



UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES
DEPARTAMENTO DE COMPUTACIÓN

Generación automática de contra-narrativas utilizando información argumentativa para combatir mensajes de odio

Tesis presentada para optar por el título de Doctor de la Universidad de Buenos
Aires en el área de Ciencias de la Computación

Damián Ariel Furman

Directora de tesis:
Co-Directora de tesis:

Dra. María Vanina Martínez
Dra. Laura Alonso Alemany

Consejero de Estudios:

Fernando Schapachnik

Lugar de trabajo:

Instituto de Ciencias de la Computación, UBA/CONICET

Fecha de defensa:

04/06/2025

Buenos Aires, 2024

Generación automática de contra-narrativas utilizando información argumentativa para combatir mensajes de odio

En el siguiente trabajo elaboramos un *dataset* de *tweets* de odio anotados con componentes argumentativos generales y específicos del dominio y con tipos distintos de contra-narrativas definidas según estrategias basadas en estos componentes, con el objetivo de utilizarlos para mejorar el rendimiento de distintos modelos de lenguaje en la tarea de generación automática de contra-narrativas para combatir la xenofobia.

Mostramos que se puede obtener un nivel aceptable de acuerdo entre anotadores, a pesar de la naturaleza subjetiva de la tarea, al utilizar un manual de anotación definido a través de un proceso iterativo que involucra a los anotadores y que los componentes argumentativos propuestos pueden ser identificados, luego, de manera automática con un rendimiento satisfactorio.

Estudiamos y desarrollamos las falencias que tienen las métricas utilizadas para realizar evaluaciones automáticas de generación de texto en la tarea de generación de contra-narrativas, tanto aquellas basadas en superposición de n-gramas como aquellas basadas en comparación de *embeddings* y proponemos categorías de evaluación que permiten definir una metodología para asignar puntajes numéricos a las contranarrativas a la vez que se explicitan las características deseables que deben tener y definen qué significa que una contra-narrativa sea aceptable o buena.

Con esta herramienta realizamos una evaluación con anotadores humanos a través de la cual concluimos que para la familia de modelos Flan-T5 el factor que aumenta en mayor medida el rendimiento de los modelos es el ajuste fino a un conjunto de datos de calidad, en comparación a aumentar el tamaño del modelo o incluso a utilizar la información argumentativa. La información argumentativa no mejora el rendimiento de los modelos de manera significativa con la excepción de aquellos que se ajustan utilizando un solo tipo de contranarrativa y la información argumentativa en la que basan su estrategia.

Finalmente, utilizamos la evaluación humana para entrenar modelos para realizar evaluaciones automáticas y evaluamos así la generación realizada por distintos Grandes Modelos de Lenguaje con múltiples configuraciones de generación.

Palabras claves: Contra-narrativas, argumentación, discurso de odio, Grandes Modelos de Lenguaje, evaluación de texto generado

Atención: por favor, quede usted advertido de que este trabajo de investigación contiene instancias de discurso de odio que pueden resultar angustiantes u ofensivas para los lectores. Estas expresiones son incluidas con el único propósito de ser analizadas y criticadas y no reflejan las creencias ni cuentan con el apoyo de los autores de este trabajo ni de su institución.

Automatic generation of counter-narratives using argumentative information to fight hate speech online

In the following work, we develop a dataset of hate tweets annotated with general and domain-specific argumentative components, as well as different types of counter-narratives defined based on strategies derived from these components. The goal is to use them to improve the performance of various language models in the task of automatic counter-narrative generation to fight online xenophobia.

We show that it is possible to achieve an acceptable level of inter-annotator agreement, despite the subjective nature of the task, by using an annotation manual defined through an iterative process involving the annotators. Furthermore, we show that the proposed argumentative components can subsequently be identified automatically with satisfactory performance.

We study and address the shortcomings of metrics used for automatic evaluations of text generation in the counter-narrative generation task, including those based on n-gram overlap and embedding comparisons. We propose evaluation categories that define a methodology for assigning numerical scores to counter-narratives, explicitly stating the desirable characteristics they should possess and defining what it means for a counter-narrative to be acceptable or good.

Using this tool, we conduct an evaluation with human annotators through which we conclude that, for the Flan-T5 family of models, the factor that most significantly improves model performance is fine-tuning on a high-quality dataset, compared to increasing model size or even using argumentative information. Argumentative information does not significantly improve model performance except in cases where models are fine-tuned using a single type of counter-narrative and the argumentative information underpinning its strategy.

Finally, we use human evaluations to train models for conducting automated evaluations and use them to assess the generations produced by various Large Language Models under multiple generation configurations.

Keywords: Counter-narratives, Argumentation, Hate Speech, Large Language Models, Evaluation of Open Text Generation

Warning: Please be advised that this research paper contains instances of hate speech that may be distressing or offensive to readers. These expressions are included for analysis and critique purposes only, and they do not reflect the beliefs or endorsements of the authors or the institution.

Agradecimientos

En el transcurso de la elaboración de esta tesis sucedieron los siguientes eventos, en el siguiente orden: una pandemia mundial nos obligó a un confinamiento casi total durante casi dos años, transformando nuestra vida e introduciendo a niveles previamente inexistentes la virtualidad; casi en simultáneo, nació mi primera hija, Tatiana, y casi tres años más tarde, en una etapa distinta de la tesis, nació también mi hijo, Santiago; fui elegido delegado gremial por ATE, asumiendo el desafío de representar a mis compañerxs becarixs siendo parte también de Jóvenes Científicxs Precarizadx; y poco después, en un contexto donde la investigación pública venía sufriendo distintos golpes, asumió un gobierno que se propuso dismantelarla por completo, convirtiéndonos a los investigadores en blanco principal de sus ataques fascistas y planteando la necesidad de multiplicar la lucha previamente existente por defenderla. Lo que quiero transmitir es que esta tesis se hizo con mucho esfuerzo. Un esfuerzo que no podría haber sido realizado si no hubiera contado con el apoyo y el cariño de todas las siguientes personas:

- De mi esposa Lucía y de mis hijos Tatiana y Santiago que me guían cuando pierdo el rumbo.
- De mi papá Ricardo y mi mamá Patricia, que siempre me brindan más apoyo del que estoy dispuesto a pedir.
- De mi hermano Ale, que tiene el corazón más grande que haya visto.
- De la bola, mi familia sin vínculo de sangre.
- De Ana, Julio, Pato, Marcuser, Thael, Tomi, Flor, Lean, Tofi, Zippo, Juancito, con quienes nunca faltan risas.
- De todos mis compañerxs de militancia con quienes comparto la esperanza en el futuro.
- De Jóvenes Científicxs Precarizaxs, porque INVESTIGAR ES TRABAJAR.
- De mis suegros, Charlo y Elena, y mi cuñado Fernando, quienes colaboraron con este trabajo más de lo que ellos se imaginan.
- De Juan Manuel Pérez, por ser tan buen amigo como compañero de trabajo.
- De José Rodríguez, Pablo Torres y Diego Letzen quienes me ayudaron a que mis trabajos y publicaciones pudieran ver la luz.
- De Ricardo Rodríguez, que me abrió los brazos sin dudarlo.

- Y por último, como corresponde a una publicación de carácter científico, a mis directoras, Laura y Vanina sin quienes nada de todo esto hubiera sido posible.

A todos ustedes, les agradezco profundamente y les dedico este trabajo.

A mis hijos, Tati y Santi, y a mi compañera de la vida, Lu.

Índice general

1..	Introducción	1
1.1.	Por qué combatir el discurso de odio	1
1.2.	Limitaciones de la censura para combatir el discurso de odio	3
1.3.	Los modelos de lenguaje como herramienta para combatir discurso de odio	4
1.4.	Nuestra aproximación al problema	4
1.5.	Desafíos y dificultades	5
1.6.	Contribuciones de este trabajo	6
1.7.	Estructura de este documento	9
2..	Trabajo Relacionado	11
2.1.	Modelos Argumentativos	11
2.1.1.	Esquemas Argumentativos	11
2.1.2.	La Tabla Periódica de los Argumentos	13
2.2.	Generación Automática de Contranarrativas	14
2.3.	Evaluación de Contranarrativas	17
2.3.1.	Categorías de evaluación	17
2.3.2.	Evaluación basada en comparaciones de pares	19
3..	Desarrollo de Corpora	21
3.1.	Manual de Anotación	21
3.1.1.	Argumentativo o No-argumentativo	22
3.1.2.	Componentes argumentativos generales: Justificación y Conclusión	23
3.1.3.	Componentes específicos del dominio: Colectivo y Propiedad	24
3.1.4.	Relación argumentativa: el Pivot	25
3.1.5.	Tipos de Proposición	27
3.1.6.	Contranarrativas	28
3.2.	Data Statement	30
3.2.1.	Curation Rationale	31
3.2.2.	Equipo de anotación	32
3.2.3.	Entorno de la extracción de los datos	33
3.2.4.	Variedad dialectal y demografía de los anotadores	33
3.3.	ASOHMO	33
3.3.1.	Acuerdo entre anotadores	33
3.3.2.	Estadísticas	37
3.4.	CONEAS	38
3.4.1.	Estadísticas	38

4..	Identificación Automática de Componentes Argumentativos	39
4.1.	Modelos	40
4.1.1.	RoBERTa	40
4.1.2.	BERTweet	41
4.1.3.	XLNet-Roberta	41
4.2.	Experimentos	42
4.2.1.	Predicción de Componentes de manera Individual	43
4.2.2.	Agregar información relevante sobre otros componentes	49
4.2.3.	Segmentando el problema	52
4.2.4.	Predicción Conjunta de Componentes	54
4.2.5.	Predicción sobre el Tipo de las Premisas	57
4.2.6.	Impacto de la cantidad de datos utilizados para el entrenamiento	59
5..	Generación Automática de Contranarrativas para Combatir Mensajes de Odio	63
5.1.	Modelo base	63
5.2.	Configuraciones de Experimentación	65
5.3.	Métodos de aprendizaje	66
5.3.1.	Ajuste fino	67
5.3.2.	Aprendizaje en contexto	68
5.4.	Evaluación de modelos	69
5.4.1.	Anotación humana de las evaluaciones	72
5.5.	Parámetros de generación de textos	74
5.6.	Resultados	76
5.6.1.	Ajuste fino vs. aprendizaje en contexto	77
5.6.2.	Tipos de contranarrativas	80
5.6.3.	Información argumentativa	81
5.6.4.	Tamaño del modelo	82
6..	Evaluación automática de contranarrativas	87
6.1.	Métricas clásicas utilizadas para evaluación automática de texto generado	87
6.2.	Evaluación automática de contranarrativas utilizando modelos de lenguaje ajustados	91
6.2.1.	Partición del conjunto de datos	91
6.2.2.	Modelos utilizados	93
6.2.3.	Resultados	96
6.3.	Evaluación automática de contranarrativas generadas por Grandes Modelos de Lenguaje	98
6.3.1.	Modelos	99
6.3.2.	<i>Entrenamiento y Prompting</i>	101
6.3.3.	Resultados	102
7..	Conclusiones y Trabajo Futuro	115
7.1.	Conclusiones	115
7.1.1.	Construcción de un <i>dataset</i> de contra-narrativas aumentado con in- formación argumentativa	115
7.1.2.	Componentes argumentativos	116
7.1.3.	Generación de Contra-narrativas	117
7.1.4.	Evaluación de Contra-narrativas	118

7.2. Trabajo Futuro	120
7.2.1. Enfoque mixto entre componentes y esquemas argumentativos e incremento del conjunto de datos anotados	120
7.2.2. Incorporar las máximas de relevancia de Grice en las categorías de evaluación	121
7.2.3. Evaluación a través de un sistema de <i>rankings</i>	123
7.2.4. Evaluar distintas condiciones de generación	124
8.. Limitaciones y consideraciones éticas	125
Bibliografía	127
Apéndice	137
A.. Nuevo protocolo de evaluación de contranarrativas	139

1 Introducción

"For every Machiavelli who urges law to prohibit speech that induces hatred of the state, there is a Walt Whitman who urges us to 'hate tyrants'" [Post, 2009]

1.1 Por qué combatir el discurso de odio

El discurso de odio es un peligro creciente para todos los habitantes del planeta. Este enseña a la gente a ver a otros seres humanos como menos que humanos y como amenazas a nuestra sociedad [Benesch, 2014a], abriendo el camino luego a que se cometan crímenes y atrocidades justificadas como un acto de defensa propia.

La *International Convention on the Elimination of all Forms of Racial Discrimination*¹ de la ONU plantea que el discurso de odio *"rechaza principios básicos de derechos humanos como la dignidad humana y la equidad y busca degradar la posición de individuos o grupos en la estima de la sociedad"*².

Varios autores plantean una relación directa entre los mensajes de odio y los actos de odio. Awan and Zempi [2016] plantean que existe una continuidad y un límite difuso entre las agresiones motivadas por el odio racial o religioso en línea y fuera de línea, y que una de las consecuencias del discurso de odio es generar el miedo en las víctimas de que las amenazas o agresiones sufridas se materialicen en la vida real. Williams et al. [2019] también estudian la relación entre mensajes de odio en redes sociales y crímenes de odio y plantean que la actividad en línea puede servir para predecir (y quizás prevenir) crímenes de odio fuera de las redes sociales. Müller and Schwarz [2017] plantean que la importancia de detectar y moderar el discurso de odio es evidente, dada su fuerte conexión con auténticos crímenes de odio.

La expansión de las redes sociales en nuestra sociedad ha proporcionado un medio para que estos discursos se expandan con mayor rapidez y con mayor alcance. A través de esta amplificación, sus efectos también se amplifican dado que pueden reforzar y profundizar prejuicios y estereotipos [Citron and Norton, 2011].

Matamoros-Fernández [2017] también estudia la relación entre racismo y redes sociales y plantea un nuevo tipo de racismo que llama *"Platformed Racism"* (Racismo de Plataformas) para el cual las plataformas actúan como amplificadoras y distribuidoras del discurso racista.

Gagliardone et al. [2015] realizan un estudio extensivo sobre el discurso de odio en línea

¹ Convención Internacional para la Eliminación de toda Forma de Discriminación Racial

² United Nations Strategy and Plan of Action on Hate Speech: Detailed Guidance on Implementation for United Nations Field Presences, 2020.

en el cual intentan analizar todas las posibles respuestas que se pueden dar desde distintos ámbitos como la ley, los estados, las organizaciones no gubernamentales, las compañías y la misma sociedad civil. Si bien destacan el rol de la legislación en combatir este fenómeno, también plantean riesgos y limitaciones, expresadas en tres puntos principales:

- “La ley refuerza las costumbres del grupo dominante que controla el contenido de la ley”. Como ejemplo, muestran que las leyes contra el discurso de odio en la Sudáfrica del *apartheid* eran utilizadas para criminalizar las críticas a la dominación por parte de la población blanca. Las leyes no siempre distinguen entre expresiones orientadas contra grupos minoritarios o marginados dentro de la sociedad y grupos que gozan de una situación de privilegios dentro de la misma. La misma definición de qué es discurso de odio constituye también una disputa de carácter político que puede ser utilizada para la persecución de ideologías opuestas a una determinada posición dominante. Actualmente, por ejemplo, se hace una utilización política del anti-semitismo para atacar cualquier cuestionamiento al Estado de Israel en el contexto de los reiterados ataques que realiza contra la población civil en la franja de Gaza.
- “Un enfoque puramente legalista puede dejar de lado cómo las sociedades evolucionan a través de la contestación y el desacuerdo”. Si bien plantean que el discurso de odio pensado como expresión de disenso es una forma ofensiva y deshumanizante, puede verse también como un síntoma de tensiones y desigualdades profundamente enraizadas en la sociedad que necesitan ser atendidas más allá únicamente de la dimensión del discurso de odio. En este sentido, es interesante observar el trabajo de Benesch [2012] que toma como guía el caso de un activista Sudafricano que cantó “Dubulu Ibhunú”, traducido como “dispara al Boer” en un mitín político interno de su organización. El canto fue traducido y publicado por medios de comunicación lo cual disparó una demanda y posterior condena del activista por “incitación a la violencia” y la prohibición de la tradicional canción anti-*apartheid* vinculada a movimientos de resistencia previos a la caída del régimen racial, a pesar de que éste aclaró que su sentido no era literal y que no era su intención incitar a acciones violentas. El artículo plantea que la interpretación de la canción está directamente vinculada al contexto político del país en el cual las tensiones entre los granjeros Afrikaans (Boers) y la población de origen étnico nativo (de piel oscura) continua, debido a las diferencias socioeconómicas que aún persisten más allá de la caída del régimen de *apartheid*.
- Existen múltiples expresiones que no dan lugar a la aplicación de una sanción legal desde el punto de vista criminal o civil pero que de todas maneras constituyen una preocupación en términos de tolerancia, civilidad y respeto por el otro. Es decir, plantean que un enfoque legalista no es una respuesta suficiente para abarcar el espectro completo del discurso de odio.

La pandemia desatada a finales del 2019 que provocó el aislamiento forzado de la totalidad de la población mundial también colaboró al crecimiento de la circulación del odio en línea. Según la iniciativa “Lo Siento Hater” de UNICEF³, la navegación web aumentó cerca de un 70 % en sus inicios. También reportan un crecimiento en la exposición

³ <https://www.unicef.org/argentina/lo-siento-hater>

a situaciones de discriminación, maltrato, cyberbullying, violencia o pornografía. En el marco de esta iniciativa, se busca difundir y posibilitar el acceso a herramientas online para combatir estos fenómenos.

1.2 Limitaciones de la censura para combatir el discurso de odio

El enfoque utilizado mayoritariamente en la actualidad para tratar con el discurso de odio es el bloqueo y la censura. Este enfoque obedece principalmente a la necesidad de las plataformas *online* de obtener cobertura frente a la posible o potencial responsabilidad en la que puedan incurrir esparciendo este tipo de mensajes. Efectivamente, la censura es efectiva para prevenir que los mensajes de odio continúen expandiéndose. Sin embargo, reconocemos dos desventajas en este enfoque:

- El bloqueo o la eliminación del mensaje no contrarresta los efectos o consecuencias generadas en aquellos que ya fueron alcanzados por el mensaje. Aún más, consideramos que en general un mensaje de odio no se genera espontáneamente sino que surge de prejuicios y estereotipos enraizados previamente en la sociedad o en una parte de la sociedad en la que está inserta quien distribuye luego el mensaje. El bloqueo, en ese sentido, no contribuye a deconstruir esos prejuicios y estereotipos sino que se limita a establecer un control sobre su proliferación.
- El bloqueo o la eliminación del mensaje no da lugar a sutilezas, grados o matices: requiere una clasificación binaria para poder tomar una decisión respecto a la supresión o no del contenido. Esto implica:
 1. En primer lugar, la necesidad de apegarse a una definición concreta sobre qué es y qué no es un mensaje de odio, lo cual a su vez implica la necesidad de apegarse a una ideología y un sistema de creencias determinado, las cuales en muchos casos no son explicitadas, perjudicando la transparencia y rigurosidad del proceso.
 2. En segundo lugar, incluso con una definición completa y extensa de qué se considera lenguaje de odio, la tarea de clasificación del discurso de odio tiene un alto grado de subjetividad [Talat and Hovy, 2016] obteniendo por lo general niveles de acuerdo entre anotadores más bajos que otras tareas con grados menores de subjetividad. Adicionalmente, Davidson et al. [2019] han demostrado que las anotaciones de 5 datasets de clasificación de mensajes de odio están sesgadas, mostrando una tendencia mayor a etiquetar comentarios como ofensivos si fueron escritos por ciudadanos Afroamericanos.
 3. En tercer lugar, los propios sistemas de clasificación automáticos pueden fallar, generando acusaciones de sobrebloqueo o censura. Es decir, son sistemas que requieren niveles altos de precisión. Adicionalmente, también se ha mostrado que estos sistemas amplifican los sesgos que puedan estar presentes en los datasets que usan como entrenamiento [Davidson et al., 2019].

No sólo las redes sociales expanden los sesgos y prejuicios presentes en la sociedad. Los grandes modelos de lenguaje (LLMs) están generando que las aplicaciones de la generación automática de lenguaje abarquen cada vez más áreas de la actividad humana. Estos

modelos también tienden a amplificar estereotipos, contribuyendo a la naturalización de prejuicios y a la deshumanización de grupos sociales. La expansión de este tipo de modelos refuerza la necesidad de mejorar los mecanismos para contrarrestar los mensajes de odio.

1.3 Los modelos de lenguaje como herramienta para combatir discurso de odio

Existe una alternativa al bloqueo y la censura que ha estado ganando lugar dentro de la literatura científica en los últimos años y consiste en “oponer contranarrativas (es decir, respuestas textuales informativas) al discurso de odio [Benesch, 2014b, Chung et al., 2019]⁴. Las contranarrativas permiten contrarrestar no solamente el mensaje concreto al que responden sino a toda una narrativa de odio de la cual forma parte ese mensaje y que se encuentra presente en una parte de la sociedad con relativa independencia de un mensaje en particular. Las contranarrativas pueden llegar de esa manera no solamente a las personas que ya fueron alcanzadas por el mensaje que se busca contrarrestar sino también a nuevas personas que puedan ser particularmente susceptibles a esos mensajes, estén o no participando de la conversación. A la vez, mediante la elaboración de contranarrativas se minimiza también el impacto de errores en la clasificación del mensaje original, evitando el sobrebloqueo y la censura ya que no se elimina ningún contenido.

Sin embargo, el inmenso volumen de mensajes de odio en línea hace de la generación manual de contranarrativas una tarea imposible. En este escenario, automatizar la generación de contranarrativas se presenta como una alternativa apeladora, pero la tarea presenta aún grandes desafíos debido a los complejos patrones lingüísticos y comunicativos involucrados en la argumentación.

1.4 Nuestra aproximación al problema

En primer lugar, para analizar esta alternativa cabe la pregunta de si cualquier mensaje de odio puede ser contestado. En redes sociales, es común la estrategia de evitar responder ciertos mensajes provenientes de usuarios identificados como *trolls* (y existe una frase, “*do not feed the troll*” para referirse a esta estrategia) debido a que contestar esos mensajes puede darles mayor entidad y relevancia de la que en verdad tendrían sin respuesta ni atención. Es decir, debemos considerar que al contestar un mensaje corremos el riesgo de potenciar su alcance en lugar de limitarlo. Entonces, ante cada mensaje de odio concreto, debemos plantearnos la pregunta: ¿merece este mensaje ser contestado?

Esta pregunta no tiene una única respuesta, sino que ésta dependerá de cuál sea el objetivo a alcanzar en la contestación. Podrían analizarse distintos ejes asociados a distintas maneras de separar los conjuntos de mensajes de odio que merecen o no ser contestados, como por ejemplo, persuasividad, masividad, objetivo del ataque, entre otros. En el contexto de este trabajo, sin embargo, consideramos que el objetivo de una respuesta o contranarrativa debe ser el de generar un debate y un intercambio entre narrativas opuestas. Para que este intercambio sea posible, entendemos que debe existir en el mensaje original una estructura argumentativa que dé lugar a la posibilidad de una refutación en los mismos términos. Entendemos que no es productivo contestar insultos, descalificaciones o llamados a la acción que no estén acompañados de algún tipo de explicación o justificación.

⁴ No Hate Speech Movement Campaign: <http://www.nohatespeechmovement.org/>

Nuestro planteo para los mensajes de odio que no se consideran argumentativos es su eliminación. Es decir, la generación automática de contranarrativas se presenta como una política complementaria que permite evitar la censura siempre que sea posible, pero de ninguna manera apunta a reemplazar por completo la eliminación de contenido de odio que no permita la mínima chance de entablar un diálogo y un intercambio.

1.5 Desafíos y dificultades

Los mensajes de odio que etiquetamos provienen del dataset HatEval [Basile et al., 2019], los cuales a su vez, fueron recolectados de la red social Twitter. Esta red social se caracteriza por tener una política flexible de moderación que se ha visto flexibilizada aún más en el último tiempo, dando lugar a un incremento de la toxicidad en sus intercambios [Vaidya et al., 2024]. Se caracteriza también por la limitación en la cantidad de caracteres que es posible utilizar en los mensajes, lo cual ha dado lugar, también, a una concepción flexible de la correctitud gramatical y sintáctica, con usos frecuentes de abreviaturas, mal uso de la puntuación, entre otras cuestiones. Estos factores acrecientan las dificultades de anotación y potencialmente dificultan el procesamiento de estos textos por parte de los modelos de lenguaje.

Otro problema asociado a la generación automática de contranarrativas tiene que ver con la capacidad de los modelos de lenguaje y de aprendizaje automático de reconocer y generar argumentos. En este punto, el surgimiento reciente de Grandes Modelos de Lenguaje ha revolucionado casi todas las tareas relativas a la generación de lenguaje. Particularmente, Thorburn and Kruger [2022] muestran que ChatGPT puede encarar seis tareas de razonamiento argumentativo con cierto nivel de éxito. También muestran que la estrategia de ajuste fino para una tarea específica supera a estrategias basadas únicamente en el aprendizaje a través de insertar contexto y/o ejemplos en la entrada del modelo (conocido como *in-context learning*).

Sin embargo, tal como Hinton and Wagemans [2023a] muestran en su análisis de las capacidades argumentativas de GPT3, el texto generado por estos LLMs tiende a mostrar ciertas debilidades. Si bien el lenguaje utilizado es argumentativo y la estructura argumental generada es correcta, la mayoría de ellos no son considerados como aceptables por jueces humanos, fallando en falacias como la de “Petición de principios”, donde la conclusión se asume como verdadera en alguna de las premisas, o dando información irrelevante. Los LLMs se comportan de manera competente desde un punto de vista puramente conversacional, pero no fueron diseñados para razonar o argumentar. A su vez, muchas veces incorporan efectos dañinos que sólo pueden combatirse de manera superficial, con un enfoque basado en reglas y que sólo recientemente han incorporado técnicas más sofisticadas como *Reinforcement Learning with Human Feedback (RLHF)* [Ouyang et al., 2022] que de todos modos requieren un grado de supervisión humana. Es por eso que parece ser necesario incorporar mejoras en estos modelos que vayan más allá del pre-entrenamiento únicamente con texto sin anotar, para obtener así un tratamiento más profundo de la dimensión argumentativa del lenguaje.

En este sentido, las herramientas de análisis automático de argumentos pueden servir también como complemento y auxilio para estrategias que apunten a respuestas automáticas, como la generación automática de contranarrativas o la recuperación de argumentos de grandes bases de datos (*argument retrieval*). Los recientes avances en LLMs facilitan estos enfoques porque logran incorporar información contextual dentro de la entrada del

modelo. En particular, para la tarea de *Question Answering*, los sistemas de Generación Aumentada con Recuperación (RAG por sus siglas en inglés) incorporan un contexto recuperado automáticamente mediante un sistema de Recuperación de Información (IR por sus siglas en inglés) que condiciona la respuesta generada y mejora la performance de los sistemas en diversas tareas de respuesta a preguntas [Lewis et al., 2021, Gao et al., 2024].

1.6 Contribuciones de este trabajo

En el marco de este trabajo presentamos “Estructura Argumentativa De Mensajes de Odio Online” (ASOHMO, por el nombre en Inglés “*Argumentative Structure Of Hate Messages Online*”) que consiste en un protocolo para anotar información argumentativa en tweets de odio contra inmigrantes y un dataset anotado con este protocolo, que puede ser utilizado para entrenar un clasificador automático. Los componentes argumentativos definidos se basan en las distintas categorías que utiliza Wagemans [2016] para construir su Tabla Periódica de los Argumentos, adaptados y agregando elementos específicos relativos al discurso de odio en redes sociales. En particular, tomamos de este trabajo la división de los argumentos en dos componentes principales, justificación y conclusión, y la subsiguiente clasificación de cada uno de estos componentes en tres tipos distintos: hecho (*fact*), política (*policy*) o valoración (*value*). También definimos un elemento nuevo, el pivot, como el elemento en común mencionado en la justificación y la conclusión, si existiese. Agregamos a su vez, dos componentes interrelacionados, específicos del dominio del discurso de odio que son el Colectivo objetivo del odio y una Propiedad (puede ser un adjetivo, una consecuencia negativa o cualquier aspecto negativo que se busque asociar a ese Colectivo).

Analizar mensajes en redes sociales representa un desafío particular porque los argumentos usualmente se refieren a elementos implícitos o ausentes y la claridad del texto es baja, encontrando frecuentemente errores de ortografía o gramaticales, frases incompletas y sintaxis incoherente. A pesar de esto, logramos obtener un acuerdo entre anotadores humanos sustancial al identificar la estructura argumental de los tweets. A su vez, mostramos para un grupo de Modelos de Lenguaje, que es posible detectar con éxito y de manera sistemática algunos de estos componentes argumentativos, incluso utilizando una cantidad baja de ejemplos anotados, lo cual indica que pueden entrenarse en este dominio para realizar esta tarea de manera automática.

También presentamos en el contexto de este trabajo “Contranarrativas Explotando la Estructura Argumental” (CONEAS, por el nombre en Inglés, “*COUNTERNarratives Exploiting Argumentative Structure*”) un protocolo de anotación y un dataset que asocia contranarrativas de cuatro tipos distintos a cada uno de los tweets de odio anotados con estructura argumental que forman parte de ASOHMO. Tres de estos tipos de contranarrativas están definidos en torno a tres estrategias distintas elaboradas en relación a distintas combinaciones de componentes argumentativos, mientras que el cuarto tipo es cualquiera que no encaje claramente en las otras tres estrategias. Las tres estrategias consisten en:

- Negar la relación de implicación entre la justificación y la conclusión.
- Negar la relación entre el colectivo objetivo del odio y la propiedad que se le quiere endilgar.
- Negar la justificación basandose en su tipo: si es de tipo “Hecho” pedir pruebas, fuentes o negar su veracidad; si es de tipo “Valoración”, destacar que es una va-

loración xenófoba o que causa daño a otros; si es de tipo “Política” proponer una política opuesta

Junto con el *dataset*, presentamos un análisis del impacto que tienen distintas combinaciones de la información argumentativa en mejorar la calidad de las respuestas generadas automáticamente por los modelos de lenguaje contra los mensajes de odio. En particular, exploramos cuatro combinaciones:

- Sin información argumentativa
- Utilizando la Justificación y la Conclusión
- Utilizando el Colectivo y la Propiedad
- Utilizando toda la información argumentativa: Justificación, Conclusión, Pivot, Colectivo y Propiedad

También exploramos el impacto de sólo utilizar en el entrenamiento, únicamente un tipo de contranarrativa, para cada uno de los tres tipos de contranarrativa asociados a una estrategia de respuesta particular. A su vez, exploramos distintas estrategias de aprendizaje de los modelos: mediante la adaptación específica a la tarea o utilizando aprendizaje en contexto, proveyendo una cantidad reducida de ejemplos como parte del *prompt* que el modelo debe completar.

Para poder evaluar las distintas configuraciones de experimentación relativas a estas tres dimensiones (información argumentativa, estrategia de aprendizaje y tipo de contranarrativa) necesitamos un mecanismo de evaluación eficiente. Sin embargo, la evaluación de contranarrativas generadas para un mensaje de odio específico no es sencilla y, hasta el momento, no se ha encontrado una herramienta automática satisfactoria para tal propósito.

Las métricas basadas en n-gramas como BLEU [Papineni et al., 2002] o Rouge [Lin, 2004] fallan en capturar muchos aspectos importantes de los textos generados porque se basan en la superposición de palabras o n-gramas con ejemplos generados manualmente para problemas donde puede haber múltiples respuestas correctas [Reiter and Belz, 2009]. Estas métricas fueron concebidas originalmente para evaluar problemas de traducción y resumen, respectivamente, donde si bien también pueden existir múltiples respuestas para un determinado ejemplo, estas son limitadas y deben tener un nivel alto de superposición de palabras entre sí.

Otras métricas basadas en el cálculo de embeddings, como *BERTScore* [Zhang et al., 2020] o *MoverScore* [Zhao et al., 2019], permiten capturar distintas formas superficiales de expresar el mismo sentido semántico y son menos susceptibles a la variación de vocabulario o del orden de las palabras que las métricas basadas en n-gramas, siempre y cuando el sentido semántico sea el mismo.

Dos contranarrativas correctas, sin embargo, no sólo pueden tener variabilidad entre sí cuando semánticamente comunican lo mismo, sino que dos contranarrativas para un mismo discurso de odio pueden tener semánticas completamente distintas (y por lo tanto, dos representaciones vectoriales totalmente distintas), buscando responder distintos ángulos o aspectos del mensaje. Y más aún, tampoco existen garantías de poder abarcar para cada mensaje, la totalidad de posibles respuestas correctas en una lista frente a la cual fuera posible contrastar, para una contranarrativa generada, si esta es similar a alguna

perteneciente a la lista de respuestas posibles anotadas o no, sin correr el riesgo de evaluar de manera incorrecta posibles nuevas respuestas no contempladas durante el proceso de anotación.

Nos enfrentamos, por lo tanto, a un problema de evaluación abierta donde los enfoques simplistas basados en una clasificación binaria de respuesta correcta o incorrecta no ofrecen buenos resultados. La evaluación manual basada en jueces humanos permite superar estas limitaciones, pero presenta otras dificultades, siendo la primera de ellas la escalabilidad.

Sin embargo, incluso desestimando ésta, definir con un criterio sistematizable qué es una buena y qué es una mala contranarrativa no es sencillo. Si consideramos el problema desde el punto de vista de nuestro objetivo, transmitir un mensaje contrario a un determinado mensaje de odio, posiblemente la mejor manera de conseguirlo dependa de varios factores, tanto comunicacionales en general como específicos al mensaje en sí. Dividir la evaluación general en categorías, en este sentido, ayuda a sistematizar el procedimiento y eliminar en el mayor grado posible la subjetividad de los anotadores, logrando mejores acuerdos entre estos.

Como una primera aproximación al problema, proponemos un protocolo para la evaluación manual de los pares discurso de odio - contranarrativa en el cual se asigna un puntaje en función de 4 categorías o ejes de evaluación: Ofensividad, Posición, Informatividad y Felicidad. Realizamos la evaluación manual de 540 pares de discurso de odio - contranarrativa utilizando jueces humanos y comparando distintas configuraciones de experimentación para poder evaluar el impacto que tiene la estrategia de entrenamiento, las distintas combinaciones de componentes argumentativos y los distintos tipos de contranarrativas.

A partir de esta evaluación, determinamos que:

- Incorporar la información argumentativa en el proceso de generación no produce en sí una mejora en la generación de contranarrativas.
- Las estrategias de aprendizaje basadas en el ajuste fino utilizando la totalidad del dataset generado para entrenar un modelo específico de generación de contranarrativas ofrecen mejores resultados que el aprendizaje mediante ejemplos provistos en tiempo de inferencia a través del aprendizaje en-contexto (*fewshot-learning*).
- La información argumentativa sí permite dirigir la respuesta generada hacia un tipo de contranarrativa particular y produce mejoras en escenarios donde la cantidad de datos de entrenamiento para ajuste fino es escasa, un resultado que muestra correlación con [Gupta et al., 2023] quienes proponen un sistema de generación de contra-discurso condicionado por la intención (informativa, denuncia, pregunta/clarificación, positiva y humor).

Las preguntas que orientan nuestro trabajo son las siguientes:

- ¿Hasta qué punto los *tweets* de odio argumentativos y los componentes argumentativos propuestos pueden ser identificados automáticamente? ¿Se pueden pensar variaciones en el proceso de detección que mejoren la identificación automática?
- ¿Qué aspectos argumentativos pueden ser útiles para mejorar el rendimiento de un LLM en tareas sutiles y riesgosas como la generación automática de contranarrativas contra el discurso de odio online?

- ¿Qué valor tiene la creación de un *dataset* con contranarrativas de calidad en la generación de contranarrativas y qué técnicas de aprendizaje se adaptan mejor a utilizarlo?
- ¿Cómo afecta al entrenamiento utilizar un tipo particular de estrategia para responder y cómo afecta la información argumentativa al aplicar una estrategia de respuesta concreta?
- ¿Cuál es la mejor forma de evaluar la generación automática de contranarrativas?
- ¿Puede sistematizarse y automatizarse nuestra propuesta de evaluación de contranarrativas, utilizando Grandes Modelos de Lenguaje? ¿Hasta qué punto es fiable esta evaluación, qué aspectos contempla y qué aspectos deja afuera?
- ¿Qué rendimiento tienen los distintos modelos de lenguaje utilizando distintas combinaciones de información argumentativa, técnicas de aprendizaje y tipos de contranarrativas?

1.7 Estructura de este documento

En el capítulo 2 analizamos el trabajo previo realizado, por un lado, en relación a las distintas teorías argumentativas revisadas para elaborar nuestra propuesta de anotación y por el otro, a los recursos desarrollados para la generación automática de contranarrativas para combatir el discurso de odio. En el capítulo 3 presentamos los conjuntos de datos anotados y describimos los protocolos y procesos de anotación, el *curation rationale* y las estadísticas de los *datasets* creados. En el capítulo 4 detallamos los experimentos relacionados a la detección automática de componentes argumentativos. En el capítulo 5 presentamos los resultados de generar contranarrativas utilizando modelos de la familia Flan-T5 junto con un protocolo de evaluación basado en cuatro categorías que utilizamos para realizar una evaluación con jueces humanos. En el capítulo 6 analizamos distintas métricas de evaluación automática para concluir que no pueden aplicarse a esta tarea y proponemos utilizar las evaluaciones humanas realizadas para entrenar un modelo que realice la evaluación de forma automática, obteniendo resultados prometedores a pesar de utilizar un conjunto acotado de datos y finalmente evaluamos con estos modelos las contranarrativas generadas por muchos otros tipos distintos de modelos de lenguaje de mayor tamaño. Finalmente, en los capítulos 7 y 8 analizamos las conclusiones de nuestro trabajo y planteamos las limitaciones y el trabajo futuro.

2 Trabajo Relacionado

2.1 Modelos Argumentativos

Existen muchas propuestas sobre cómo modelar los aspectos argumentativos de distintas clases de textos, incluso si solo consideramos aquellas dirigidas o utilizadas para aplicaciones computacionales. No es el objetivo de este trabajo presentar una visión exhaustiva de todos los enfoques, sino solo algunos ejemplos para motivar y enmarcar el modelo argumentativo que aplicamos en este trabajo.

Una de las principales distinciones entre las propuestas existentes consiste en si son de propósito general o específicas de cierto dominio. Los enfoques específicos de un dominio proponen categorías a medida, como por ejemplo, “*background*”, “*aim*” o “*comparison*”, propuestos por Teufel et al. [1999] para artículos científicos, o “*anecdote*” o “*statistics*” propuestos por Al-Khatib et al. [2016] para el análisis argumentativo de editoriales. En general, este tipo de *frameworks* tienden a lograr un buen acuerdo entre anotadores y buena precisión en la identificación automática, pero no son portables a otros dominios.

Los modelos de argumentación de propósito general tienen enfoques muy diferentes entre sí. Muchas propuestas orientadas al cómputo automático se basan en la teoría del argumento práctico de Toulmin [2003]. En esta, se distingue entre dos componentes principales: la “conclusión” (también llamada “*postulado*”) y el “hecho” (también llamado “*justificación*” o “*premisa*”). Usualmente intentan identificar relaciones entre componentes y entre argumentos, con el objetivo de crear un árbol argumentativo completo que explique la estructura argumentativa de un texto. Este tipo de modelo se ha aplicado a ensayos Stab and Gurevych [2014] o al discurso generado por usuarios Habernal and Gurevych [2017]. Es muy general, por lo que es fácilmente portable a diferentes dominios. Al mismo tiempo, no es muy estable, ya que el acuerdo entre anotadores no suele ser alto, y la información que proporciona sobre el argumento no es tan rica como en el caso de los enfoques específicos de dominio.

2.1.1 Esquemas Argumentativos

Otro enfoque para modelar argumentos en textos son los esquemas. Los esquemas argumentativos son “*patrones de razonamiento informal*” [Walton et al., 2008] que “*representan formas de argumentar que son usadas ampliamente en las conversaciones cotidianas*” [Macagno et al., 2018]. El enfoque tradicionalmente propuesto para los esquemas argumentativos por Walton et al. [2008] sigue una aproximación “*bottom-up*” que empieza con ejemplos reales y concretos de argumentos que luego pueden ser agrupados juntos en un esquema, y luego esos esquemas, a su vez, pueden ser agrupados nuevamente en clasificaciones más generales.

Cada patrón de razonamiento definido por un esquema argumentativo define un con-

junto de “preguntas críticas” orientadas a probar las condiciones de derrota de dicho patrón. Las condiciones de derrota son los “puntos débiles” del argumento, las distintas razones por las que el argumento podría no tener validez. Los esquemas, por lo tanto, se definen parcialmente en función de estrategias generales para derrotar los argumentos que contienen. Las preguntas críticas proveen información interna y accionable sobre el argumento, que puede ser utilizada luego en el contexto de otras aplicaciones como la construcción de un sistema de generación automática de contra-argumentos o contranarrativas.

Por ejemplo, un esquema de autoridad plantea que cierta afirmación debe ser válida porque lo dice o respalda una figura de autoridad. Por lo tanto, un argumento de tal tipo puede ser atacado por una pregunta como “¿cuan confiable es esta autoridad?” o “¿Es esta autoridad un experto en el campo de discusión?”. Las preguntas críticas son una parte esencial de la definición de cada esquema y, por lo tanto, los esquemas argumentativos plantean una taxonomía basada en las condiciones de derrota de los argumentos. Esta forma de agruparlos parece, *a priori*, ideal para sistematizar la respuesta automática a argumentos típicos dentro de un dominio específico como puede ser el lenguaje de odio.

Anotar esquemas argumentativos, sin embargo, es una tarea difícil. En el trabajo de Lindahl et al. [2019] se anotó un corpus de textos en sueco sobre política usando dos anotadores entrenados, obteniendo un acuerdo bajo, tanto en el número y componentes de los argumentos como en los esquemas reconocidos. Reportaron que el acuerdo respecto a los esquemas mejoró al concatenar, achicar y simplificar la cantidad de esquemas a anotar.

En Musi et al. [2016] se realizó un estudio piloto de anotación de 30 ensayos persuasivos con anotadores no entrenados y no expertos, obteniendo un acuerdo muy bajo. Posteriormente, se volvió a anotar parte del corpus con anotadores expertos entrenados y se refinó y mejoró las instrucciones, logrando una mejora significativa en el acuerdo. Este estudio destaca la importancia de construir un manual de anotación bien adecuado.

Otra dificultad tiene que ver con la falta de capacidad de generalización de los esquemas a múltiples dominios o a una aplicación general, lo que obliga a realizar un análisis de cuáles son los argumentos típicos dentro de cada dominio, una descripción detallada de cuáles son esos esquemas definiendo las preguntas críticas pertinentes, un manual para su correcta identificación, un análisis de la factibilidad de tal tarea, estudiando el acuerdo entre anotadores y un proceso iterativo donde se redefinan criterios, categorías y los mismos esquemas en sí basándose en la experiencia y dificultades reportadas por los anotadores.

Varios autores han adaptado los esquemas propuestos por Walton et al. [2008] a propósitos específicos, incluso proponiendo alternativas a las preguntas críticas [Atkinson and Bench-Capon, 2018, Kökciyan et al., 2018]. La principal desventaja de esta estrategia tiene que ver con que el inventario de esquemas puede volverse rápidamente pródigo y licencioso. Si crece mucho, pierde capacidad de generalización y al aparecer ejemplos nuevos para ser anotados no siempre logran encajar de manera unívoca en la categoría de esquemas propuestos, generando la necesidad de agregar más esquemas que contemplen los nuevos ejemplos observados o produciendo un mayor nivel de desacuerdo entre anotadores. Si por el contrario, se lo limita a una cantidad fija de casos generales, sucede que no siempre las preguntas críticas se adaptan a atacar o exponer correctamente las condiciones de derrota del argumento. En términos generales, la identificación de un esquema dado en un determinado texto es una tarea difícil, tanto para realizar de forma manual como automáticamente.

Más allá de los esquemas, consideramos que las preguntas críticas en sí son un aspecto

fundamental para poder utilizar el análisis argumentativo en la generación automática de respuestas. Creemos que es posible generalizar estas preguntas críticas, no a un tipo particular de esquema, sino a un dominio en particular, como lo es el discurso de odio en línea. Consideramos que se pueden elaborar estrategias de respuesta basadas en características particulares del dominio (como por ejemplo, la existencia necesaria de un colectivo que está siendo atacado y una propiedad negativa que se busca asociar con éste). Esta generalización, sin embargo, relaja la propiedad más importante que tienen las preguntas críticas: que siempre representan una condición de derrota para un tipo de argumento. En este caso, si bien la mayoría de los mensajes de odio tiene un Colectivo y una Propiedad asociada a este, por ejemplo, algunos podrían no tenerlos. En estos casos, la estrategia de respuesta no podría aplicarse y deberá buscarse la elaboración de otro tipo de respuesta.

2.1.2 La Tabla Periódica de los Argumentos

Con el objetivo de buscar un punto medio entre el detalle excesivo de los esquemas argumentativos y la escasa información provista por los enfoques de premisa-conclusión, Wagemans [2016] propone un enfoque analítico sobre los esquemas argumentativos que apunta a obtener los esquemas centrales con menos categorías y basado en un conjunto limitado de características generales de los argumentos.

Esta es una característica que encontramos particularmente útil para construir un sistema simple que sea fácil de anotar sin realizar un esfuerzo enorme y lograr un alto nivel de acuerdo entre los anotadores humanos, lo que nos lleva a esperar una mayor reproducibilidad en los modelos inferidos. Además, un enfoque analítico permite determinar qué aspectos de la argumentación son más factibles de detectar automáticamente e identificar qué aspectos particulares son más útiles para una aplicación dada, como los componentes que podrían usarse para elaborar una respuesta.

Todos los argumentos bajo el sistema de Wagemans tienen una premisa y una conclusión etiquetadas con un “Tipo de Declaración” cada una. Sin embargo, no se trata simplemente de una nueva propuesta de segmentación de argumentos en premisa y conclusión. La Tabla Periódica de los Argumentos es una tipología factorial de argumentos que ofrece una visión general completa de los diversos tipos que puedan existir, al describirlos como una combinación única de tres características básicas [Wagemans, 2019]:

1. **argumento de primer orden o de segundo orden.** Un término común entre la premisa y la conclusión transfiere la aceptabilidad de uno a otro. Si este término común es explícito, y no es necesario reconstruir el argumento para encontrarlo, entonces es un argumento de primer orden. Si se necesita una reconstrucción, (generalmente transformando una premisa en el sujeto de otra premisa con el predicado “es verdad”), entonces es un argumento de segundo orden.
2. **argumento de predicado o de sujeto.** Si el término común está en el sujeto de las proposiciones que forman la premisa y la conclusión, entonces es un argumento de sujeto; de lo contrario, es un argumento de predicado.
3. **política (policy), hecho (fact) o valor (value).** La conclusión y la premisa pueden etiquetarse cada una como una declaración de **Política** (el hablante manda o afirma que algo debería hacerse), una declaración de **Valor** (el hablante emite una opinión sobre algo) o una declaración de **Hecho** (el hablante transmite una proposición como un hecho verdadero).

A partir de todos los posibles valores de estas características, se construye un marco sobre el cual luego se basa una taxonomía de todos los posibles esquemas argumentativos. Las primeras dos características definen una tabla de cuatro cuadrantes llamados α , β , γ y δ y dentro de cada cuadrante, existen a su vez nueve columnas definidas de acuerdo a los posibles valores de la tercer característica (Fact, Policy o Value) para cada uno de los dos predicados (justificación y conclusión). Los esquemas argumentativos se acomodan y clasifican en cada columna de acuerdo a las características que demuestren en su forma canónica.

Visser et al. [2021] realizaron una investigación exhaustiva sobre la anotación del corpus del debate presidencial de EE.UU. de 2016 utilizando tanto los esquemas de Walton como la Tabla Periódica de Argumentos de Wagemans, y sugirieron revisiones a las directrices para anotar esquemas que mejoraron el κ de Cohen (una de las métricas clásicas usadas para medir el acuerdo entre anotadores). Aunque lograron obtener una puntuación aceptable para los esquemas de Walton, reportaron un mayor acuerdo para la tipología de Wagemans, especialmente sin considerar la clasificación entre argumentos de primer y segundo orden. Además, sostienen que para la tipología de Wagemans, “*la división en sub-tareas independientes simplifica la anotación mientras se mantiene la robustez*”.

Si bien tanto Wagemans como Walton proponen, en última instancia, una taxonomía de esquemas, Wagemans la construye con un enfoque *Top-Down*, tomando un conjunto de características generales y construyendo, a partir de estas, esquemas más concretos y específicos. Esto permite tener más control sobre el nivel de especificidad del conjunto de esquemas con el que se busca trabajar.

En el marco de este trabajo buscamos utilizar la teoría argumentativa para mejorar las respuestas a argumentos de odio. No vamos a concentrarnos en definir esquemas y clasificar argumentos según esos esquemas debido a que consideramos que agrega un nivel de complejidad a la tarea innecesario. En cambio, proponemos utilizar una adaptación al dominio (mensajes de odio en redes sociales), de las características de Wagemans directamente para definir estrategias de respuesta sin tener que pasar por una clasificación taxonómica precisa. Buscamos identificar elementos que puedan ser relevantes tanto para un humano como para una máquina en la tarea de analizar o contrarrestar el discurso de odio.

Enfocándonos en el discurso de odio en Twitter, debemos tener en cuenta que muchos tuits de odio argumentativos se basan en suposiciones justificadas por prejuicios o información contextual que es difícil de recuperar. Esto significa que, en muchos casos, es difícil refutarlos desde la perspectiva de la lógica deductiva formal. Creemos que un enfoque basado en la lógica informal, como el propuesto por Wagemans [2016], es más adecuado para capturar este tipo de argumentos que están organizados con relaciones informales.

2.2 Generación Automática de Contranarrativas

La generación automática de contranarrativas se ha abordado aprovechando los recientes avances en generación de lenguaje natural utilizando grandes modelos de lenguaje. Como en la mayoría de las tareas de generación, en los últimos años el enfoque básico de aprendizaje automático ha sido entrenar o ajustar una red neuronal generativa con ejemplos específicos de la tarea objetivo. Esto hace que los conjuntos de datos sean una parte crucial de este proceso.

El *dataset* CONAN [Chung et al., 2019] es, según nuestro conocimiento, el primer

conjunto de mensajes de odio con sus correspondientes contranarrativas anotadas. Tiene 4078 pares originales escritos manualmente por operadores de ONG, traducidos a tres idiomas: inglés, francés e italiano. Los datos se aumentaron utilizando parafraseo automático y traducciones entre idiomas para obtener 15024 pares finales de discurso de odio y contranarrativa. Desafortunadamente, este conjunto de datos no es representativo del lenguaje utilizado en las redes sociales, ya que tanto los tuits de odio como sus contranarrativas correspondientes fueron generados por los anotadores y se presentan en un estilo limpio y ordenado que rara vez se encuentra en redes sociales.

Enfoques similares fueron llevados a cabo por Qian et al. [2019] y Ziems et al. [2020]. Qian et al. [2019] han publicado un conjunto de datos con conversaciones de 5020 publicaciones de Reddit con 22324 comentarios y 11825 conversaciones de Gab con 33776 comentarios. Trabajadores de *Mechanical Turk* etiquetaron comentarios específicos en cada conversación como odiosos o no y luego escribieron respuestas a ellos. Se escribieron 21747 respuestas para el conjunto de datos de Gab y 7641 para Reddit. Aunque este conjunto de datos agrega un gran valor al contextualizar el discurso de odio dentro de una conversación, la respuesta a esos mensajes se enfoca principalmente en la moderación del debate (por ejemplo, advertencias contra el uso de lenguaje ofensivo) en lugar de generar contranarrativas adecuadas contra el mensaje de odio en sí. Ziems et al. [2020] también anotaron 3355 tuits de una base de datos de Twitter relacionada con el covid-19 y no produjeron nuevos textos, sino que etiquetaron tuits relacionados con el covid-19 como discurso de odio, contra-discurso o neutral según su nivel de odio hacia los asiáticos. Sin embargo, los tuits en este conjunto de datos solo se clasifican según su postura. No son necesariamente argumentativos y el contra-discurso no está asociado a un tuit de odio específico.

En un trabajo subsiguiente al trabajo seminal de CONAN, Tekiroğlu et al. [2020] aplicaron LLMs para ayudar a expertos a crear un corpus, con GPT-2 generando un conjunto de contranarrativas para un discurso de odio dado y expertos editándolas y filtrándolas. Fanton et al. [2021] refinaron iterativamente un LLM donde las contranarrativas generadas automáticamente fueron filtradas y post-editadas por expertos y luego alimentaron al LLM con estos ejemplos adicionales para ajustarlo, en varias iteraciones. Bonaldi et al. [2022] aplicaron este mismo enfoque para obtener un conjunto de datos generado por máquina de diálogos entre personas que producen discurso de odio y expertos en contranarrativas.

Como una mejora adicional a la metodología basada en generación mediante Grandes Modelos de Lenguaje, Chung et al. [2021a] incorporaron una arquitectura de recuperación de conocimiento al proceso de elaboración de contranarrativas. El módulo de recuperación de conocimiento recupera conocimiento relevante a nivel de oración utilizando un motor de búsqueda asociado a pares de discurso de odio y contranarrativas del conjunto de datos. Varios modelos, incluidas versiones de GPT-2 [Radford et al., 2019] y XNLG Chi et al. [2020], fueron entrenados con y sin el conocimiento relevante adicional y su rendimiento para la generación de contranarrativas se evaluó utilizando las puntuaciones BLEU-2 y ROUGE-L. Los autores también midieron la novedad y diversidad de las respuestas, a través de la similitud de *Jacqard* y la tasa de repetición, respectivamente.

Ashida and Komachi [2022] usan Grandes Modelos de Lenguaje para generación con un enfoque de *prompting*, en lugar de afinarlos con ejemplos creados o curados manualmente. También proponen una metodología para evaluar la salida generada, basada en la evaluación humana de algunas muestras. Este mismo enfoque es aplicado por Vallecillo-Rodríguez et al. [2023] para crear un *dataset* de contranarrativas para el español. Ambos

enfoques están dirigidos a texto generado por usuarios, estrechamente relacionado con las redes sociales.

Sin embargo, ninguno de los *datasets* o enfoques mencionados anteriormente para la generación de contranarrativas incluye o integra información adicional anotada aparte del mensaje de odio, posiblemente su contexto y su respuesta.

Conocemos dos trabajos que tienen en cuenta aspectos argumentativos de las respuestas a los mensajes de odio o cercanos a la argumentación. Chung et al. [2021b] exploran estrategias para clasificar automáticamente contranarrativas en distintos idiomas según siete posibles tipos definidos a partir del trabajo de Benesch et al. [2016] y tal como fueron anotados originalmente en el dataset CONAN [Chung et al., 2019]:

- Presentación de **Hechos** contrarios al mensaje de odio
- Denuncia de **Hipocresía** o contradicción en el mensaje de odio
- **Denuncia** de las consecuencias que tiene el mensaje de odio
- **Preguntas** elaboradas para provocar o exigir mayor evidencia al usuario odiante.
- Utilización del **Humor** para ridiculizar la postura de odio
- Demostrar algún tipo de **Afiliación** con el usuario odiante o su situación para luego distanciarse del argumento de odio
- La contranarrativa habla de **Otro** tema y no contesta el mensaje original

Los autores demuestran, para cuatro configuraciones de entrenamiento distintas, que es posible ajustar modelos de lenguaje pre-entrenados para predecir los tipos de contranarrativas con un buen rendimiento sobre las clases mayoritarias, pero también con un rendimiento aceptable en las clases minoritarias.

Gupta et al. [2023], por otro lado, han generado un *dataset* de contranarrativas donde cada una tiene asignada una de cinco posibles intenciones:

- **Informar:** La contranarrativa se basa en proveer hechos y datos que contradicen las asunciones o justificaciones presentes en el discurso de odio o en discutir los datos presentados en el discurso de odio.
- **Preguntar:** La contranarrativa se expresa como una pregunta que busca desafiar la cadena de razonamiento del hablante y obligarlo a responder de manera convincente o renunciar a su comentario original.
- **Denuncia:** La contranarrativa consiste en denunciar la opinión o expresión del hablante como algo inaceptable que no se puede tolerar.
- **Humor:** La contranarrativa consiste en utilizar el humor para destacar el absurdo del discurso de odio y ayudar a desviar la atención de aquellos siguiendo el diálogo
- **Positiva:** La contranarrativa consiste en utilizar una respuesta positiva y empática para reducir la animosidad y la violencia.

Si bien la información sobre posibles tipos o intenciones de las contranarrativas puede resultar útil para dirigir la generación automática de respuestas a partir de seleccionar alguna posible estrategia, no se anota información argumentativa sobre el mensaje original que se busca responder que pueda servir para mejorar la comprensión del modelo sobre éste.

Es por eso que consideramos un enfoque alternativo que apunta a alcanzar la generalización no por la mera cantidad de ejemplos, sino proporcionando un análisis más rico de tales ejemplos que guíe al modelo en la búsqueda de generalizaciones adecuadas. Creemos que la información sobre la estructura argumentativa del discurso de odio puede ser utilizada como restricción para la generación automática de contranarrativas.

2.3 Evaluación de Contranarrativas

Evaluar las contranarrativas generadas no es una tarea lineal. Hasta el momento, no existe ninguna técnica automática satisfactoria para realizar la tarea. Tal como argumenta Reiter and Belz [2009], las métricas superficiales basadas en n-gramas que normalmente se utilizan para evaluar tareas como traducción o sumarización de textos (BLEU [Papineni et al., 2002] o Rouge [Lin, 2004]) fallan en capturar importantes aspectos del texto generado porque se basan en la superposición del texto a evaluar con ejemplos generados manualmente y, por lo tanto, resultan inadecuadas en escenarios donde existen distintas posibles salidas correctas del modelo con diferencias significativas entre sí. En la opinión de Zubiaga et al. [2024], “evaluar contranarrativas es particularmente desafiante porque hay muchas posibles respuestas aceptables a un discurso de odio y suele ser difícil evaluar qué constituye una respuesta correcta”. Tampoco funcionan las técnicas de evaluación que se basan en la codificación de oraciones para medir similaridad semántica [Yang et al., 2019] debido a que existen muchas posibles respuestas correctas a un *tweet* de odio, y que existan diferencias semánticas de una contranarrativa generada con otra tomada como referencia no significa que sea peor. *A priori*, dado un *tweet* de odio, no es posible definir un conjunto limitado de contranarrativas válidas para ese *tweet* de tal modo que, si la contranarrativa generada no muestra similitudes con alguna de estas, podamos garantizar que no es adecuada.

Específicamente para textos argumentativos, Hinton and Wagemans [2023b] muestran que los LLMs producen argumentos de baja calidad que, sin embargo, son enunciados bien formados. Esto dificulta también el proceso de evaluación, ya que no alcanza con evaluar la calidad del texto generado en abstracto, sino que es necesario determinar categorías específicas que reflejen y expliciten lo que se espera de una contranarrativa “buena”.

Muchos autores reconocen estas limitaciones en los mecanismos de evaluación automáticos aplicados a contranarrativas, planteando como alternativa mecanismos de evaluación más específicos y matizados que destaquen y valoren características relevantes específicas de las contranarrativas generadas.

2.3.1 Categorías de evaluación

Ante la falta de métricas automáticas adecuadas para la tarea, muchos autores han realizado evaluaciones manuales para contranarrativas generadas automáticamente. Este método de evaluación cuenta con la desventaja de no poder escalar a una cantidad potencialmente infinita de datos que los modelos pueden generar más rápidamente que lo

que un anotador humano puede anotar, pero puede realizarse con una muestra de datos limitada para comparar el desempeño de una cantidad acotada de modelos, al menos de forma indicativa. A la vez, las evaluaciones manuales suelen distinguir diferentes aspectos de la adecuación de un texto dado como contranarrativa para otro. Encarar el problema de este modo permite profundizar en la definición y, a la vez, explicitar qué se considera una buena contranarrativa. También permite desglosar el problema de la adecuación de la anotación a un criterio definido, reduciendo, en la medida de lo posible, la subjetividad del anotador individual y aumentando la reproducibilidad de la anotación medida a través de los valores de acuerdo entre anotadores.

Conocemos tres propuestas de evaluación manual de pares de mensaje de odio/contranarrativa.

Chung et al. [2021a] evalúan tres aspectos de la adecuación de éstas al mensaje de odio que responden:

- **Adecuación:** indica si la contranarrativa es adecuada como respuesta al mensaje de odio original.
- **Informatividad:** indica cuán específica o genérica es la respuesta en relación al mensaje que responde.
- **Intra-coherencia:** evalúa la coherencia interna de la contranarrativa independientemente del mensaje al que responde, es decir, si es correcta sintáctica y semánticamente, coherente, clara y fluida en la transmisión del mensaje.

Ashida and Komachi [2022], por otro lado, evalúan otros tres aspectos:

- **Ofensividad:** indica si la respuesta es ofensiva para el colectivo objetivo del odio o cualquier otro colectivo.
- **Posicionamiento:** indica si la respuesta tiene una postura contraria al mensaje de odio o no.
- **Informatividad:** al igual que Chung et al. [2021a], indica si la respuesta es específica respecto al mensaje que responde o es una respuesta genérica.

Gupta et al. [2023] comparan distintos mecanismos de evaluación para las contranarrativas generadas por distintos modelos entrenados con sus contranarrativas anotadas con distintas intenciones, que incluyen métricas automáticas basadas en la superposición superficial de caracteres o partes de palabras, como ROUGE [Lin, 2004] o METEOR [Banerjee and Lavie, 2005], métricas automáticas basadas en mecanismos de codificación semántica, como BertScore [Zhang et al., 2020] o Similitud Coseno utilizando *Sentence Embeddings*, métricas automáticas libres de referencia (no toman en cuenta un *gold standar* sino que evalúan batches de texto generado de conjunto), como “novedad” y “diversidad”, siguiendo a Wang and Wan [2018] y, finalmente, una evaluación manual tomando una muestra aleatoria de las contranarrativas generadas según cuatro categorías:

- **Independencia:** indica si la contranarrativa generada puede ser interpretada como tal con independencia del mensaje al que responde.
- **Condicionamiento:** indica si la contranarrativa generada es una respuesta específica al mensaje de odio que responde o es genérica.

- **Adecuación:** indica si la contranarrativa generada es correcta en términos gramaticales y semánticos, coherente y fluida.
- **Toxicidad:** indica si la contranarrativa generada puede ser considerada como ofensiva o tóxica.

2.3.2 Evaluación basada en comparaciones de pares

Zubiaga et al. [2024] también llevan a cabo un estudio donde demuestran que las métricas tradicionales automáticas tienen una relación pobre o inexistente con respecto a juicios realizados por seres humanos. De este modo, proponen un nuevo mecanismo de evaluación basado en comparaciones entre pares de triplas compuestas por un discurso de odio, una contranarrativa anotada como *gold standar* por un anotador humano y una contranarrativa generada automáticamente por un modelo de lenguaje. Cada una de las contranarrativas generadas es comparada mediante esta tripla con todo el resto de las contranarrativas generadas a evaluar, al estilo de un torneo de todos contra todos.

Los autores utilizan el modelo *JudgeLM* [Zhu et al., 2023], ajustado para evaluar la calidad de textos generados, para elegir la mejor contranarrativa en cada comparación de a pares. En concreto, en cada torneo, *JudgeLM* asigna un puntaje a cada tripla y la que obtenga el puntaje más elevado se considera la ganadora, sumando un punto para el modelo que la generó. Como todos los modelos generan una sola contranarrativa para cada uno de los mensajes de odio utilizados para la evaluación, todos participan en la misma cantidad de torneos y tienen la misma posibilidad de sumar puntos. Finalmente, se suman los puntos obtenidos por cada una y se construye un *ranking* que refleja el rendimiento de cada modelo en esta tarea. Los autores, además, muestran que los *rankings* generados tienen una alta correlación con las preferencias humanas, obteniendo un ρ^1 score de 0.88.

Esta metodología de evaluación cuenta con la ventaja de utilizar un modelo *off-the-shelf* y, por lo tanto, poder prescindir del proceso de definición de categorías, posterior anotación con anotadores humanos y entrenamiento de un modelo específico para la tarea. La desventaja, en cambio, es que se utiliza para la evaluación un modelo que fue entrenado y ajustado para evaluar generación de texto abierta, pero no para la tarea específica de evaluación de contranarrativas. Al mismo tiempo, se pierde interpretabilidad, ya que es imposible reconstruir o interpretar siquiera el porqué de las elecciones realizadas por el modelo, mientras que tampoco existe un marco claro que defina lo que es una buena y una mala contranarrativa, que pueda servir para detectar anomalías o posibles problemas en el proceso de evaluación.

¹ El coeficiente de correlación de rangos de Spearman (ρ score) es una medida no paramétrica para evaluar la relación entre dos variables ordinales o, de forma indirecta, entre rankings.

3 Desarrollo de Corpora

El objetivo de nuestro modelo de argumentación es proporcionar un análisis que se pueda utilizar para generar contranarrativas para *tweets* de odio, tanto en el momento de la anotación sugiriendo al anotador cómo escribirlas basándose en los componentes etiquetados previamente, como en el momento de la predicción utilizando modelos de lenguaje. Es por esto que todas las anotaciones se realizan considerando cómo podrían ser de ayuda en esta tarea, y deben ser revisadas si en el momento de escribirlas, una alternativa ayuda a construir una mejor. Nuestro objetivo es exponer el núcleo del razonamiento que apoya un mensaje de odio para tener una comprensión más clara de cómo refutarlo.

Etiquetamos dos tipos de información: específica del dominio y general argumentativa. La información específica del dominio permite explotar características particulares de los mensajes de odio en Twitter: siempre mencionan un colectivo que está implícita o explícitamente asociado con una propiedad, acción o consecuencia negativa. La estructura argumentativa general se basa en una simplificación de la propuesta de Wagemanns que está destinada a aumentar el acuerdo entre anotadores. Alcanzar niveles aceptables de acuerdo es muy importante para nuestro trabajo, ya que indica que el proceso de anotación puede ser sistematizado y posiblemente automatizado.

3.1 Manual de Anotación

Creamos un manual de anotación¹ en el que ambos tipos de información argumentativa se definen de manera procedural. Se redactó una propuesta inicial a partir de discusiones con el equipo de anotación utilizando ejemplos generados manualmente para representar posibles casos borde. Luego, éste fue sometido a un proceso iterativo en el cual los anotadores anotaron un bache común de *tweets* de odio intentando seguir los lineamientos definidos en el manual, y luego discutieron entre sí criterios para sistematizar mejor el proceso y lograr un mejor acuerdo.

De este modo, se definieron seis pasos en el proceso de anotación:

1. Determinar si el tweet de odio es o no argumentativo. Si no lo es, el proceso concluye aquí y se procede con el siguiente ejemplo a anotar.
2. Identificar conjuntamente la justificación y la conclusión del argumento
3. Identificar el colectivo hacia quien el odio se dirige y la propiedad que se busca asociar a ellos, siempre y cuando sea explícita en el tweet

¹ La versión final del manual puede encontrarse en https://docs.google.com/document/d/1AZ1hbrxhZ-n6Lzb9fuaBZd8kAmuLc0lahG6I9dUGz_g/edit?tab=t.0

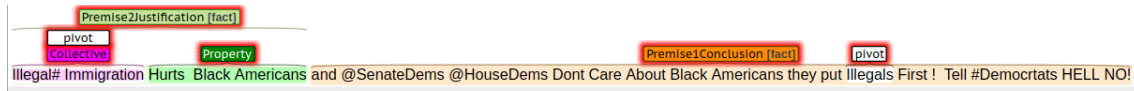


Fig. 3.1: Ejemplo de un tweet de odio argumentativo, etiquetado mediante la herramienta *brat* con Justificación, Conclusión (ambos de tipo “Hecho”/Fact), Colectivo, Propiedad y Pivot.

4. Identificar un elemento utilizado para transferir razón entre la justificación y la conclusión y anotarlo si fuera explícito.
5. Determinar tanto para la premisa como para la conclusión si son de tipo *Política*, *Hecho* o *Valoración*.
6. Escribir las contranarrativas según su tipo, intentando, siempre que fuera posible, escribir una de cada tipo.

La figura 3.1 muestra un tweet de odio anotado como argumentativo con Justificación y Conclusión, ambos de tipo “Hecho”, Colectivo, Propiedad y Pivot. El planteo principal del mensaje es que a los representantes demócratas del Congreso de los Estados Unidos no les importan los americanos negros y priorizan a los inmigrantes ilegales, y lo justifican planteando que la inmigración ilegal hace daño a los americanos negros. El colectivo objetivo del odio son los “ilegales” y la propiedad que se les endilga es “hacer daño a los americanos negros”. El elemento común entre la justificación y la conclusión que pretende transferir razón entre una y otra son los “ilegales”.

A continuación presentamos una descripción de cada uno de estos pasos de anotación y de los componentes argumentativos identificados en cada uno.

3.1.1 Argumentativo o No-argumentativo

Consideramos que un tweet es argumentativo si hay *una afirmación que se presenta con el fin de apoyar otra afirmación siempre que la aceptabilidad de esta última esté en duda* [Wagemans, 2019]. Siguiendo Wagemans [2019], *un argumento (...) consiste en dos afirmaciones, a saber, una conclusión – la afirmación que se duda – y una premisa*. Consideramos que un tweet es argumentativo si es posible dividirlo en estos dos componentes. Para hacer esto, consideraremos las relaciones entre los segmentos principales dentro de cada uno de ellos. Es por eso que esta fase de anotación y la siguiente en la cual se busca identificar efectivamente la justificación y la conclusión están interrelacionadas: durante esta etapa, debemos corroborar que el tweet está conformado por más de un componente y puede segmentarse en unidades que se relacionan entre sí como emisores/receptores de apoyo o refutación. Si bien no es necesario en esta etapa definir cuál es cada uno de estos componentes, a nivel cognitivo, la definición de un tweet como argumentativo presupone que el anotador ya evaluó algún o algunos escenarios concretos sobre cómo realizar esta segmentación.

Ejemplos de *tweets* no argumentativos pueden ser: exhortaciones a alguna acción sin justificación, insultos, descalificaciones, apoyo a una política en particular o descripción de hechos sin una conclusión explícita. Todos los *tweets* de nuestro conjunto de datos están clasificados como argumentativos o no argumentativos. Si el tweet no es argumentativo, no se realiza ninguna anotación adicional y no se escriben contranarrativas. En las figuras 3.2

y 3.3 podemos observar un ejemplo etiquetado como no argumentativo. El primer *tweet* consiste en una consigna rechazando la instalación de campos de migrantes (refugiados) en Libia, pero no presenta ninguna justificación para tal pedido ni busca ligarla a otro planteo a través de una relación argumental. El segundo *tweet* parece el titular de una noticia sobre un supuesto plan siniestro de inmigración, pero, nuevamente, esta afirmación no se encuentra ligada a ninguna justificación ni a ningún otro elemento respaldatorio.

No to #EU migrant camps in Libya, PM al-Serraj

Fig. 3.2: Example of non-argumentative tweet

Stephen Miller's Sinister New Immigration Proposal :link:

Fig. 3.3: Example of non-argumentative tweet

3.1.2 Componentes argumentativos generales: Justificación y Conclusión

Si el tweet es argumentativo, esto significa, según nuestra definición, que es posible identificar una Conclusión y una Justificación. Todos los *tweets* argumentativos deben estar etiquetados con una y solo una de cada una de estas, aunque pueden estar separadas en muchas partes no contiguas dentro del tweet. Todas las partes no contiguas, dado que forman parte de la misma premisa, deben ser del mismo Tipo. Se instruyó a los anotadores para que eligieran la Conclusión y Justificación más largas que pudieran encontrar, tratando de abarcar la totalidad del contenido significativo del tweet, dejando fuera solo hashtags que indicaran temas, enlaces, menciones de usuarios o palabras o información no relevantes.

Las justificaciones y, en menor medida, también las conclusiones, pueden ser argumentos en sí mismas, teniendo su propia estructura anidada que involucra diferentes divisiones internas en justificación y conclusión. Pero solo buscamos anotar la construcción argumentativa principal, representada por un único par de justificación y conclusión. Buscamos capturar el punto de vista central que el usuario quiere que sea aceptado con el objetivo de tener información relevante que permita construir una contranarrativa a partir de negar la relación entre estas premisas. La profundidad en la que las estructuras pueden anidarse, de todos modos, está limitada por la longitud del mensaje, que no puede superar los 280 caracteres estipulados para todos los mensajes de la red social Twitter.

Al etiquetar estos componentes, no estamos considerando la estructura sujeto-predicado propuesta por Wagemans. Visser et al. [2021] advirtió sobre cómo este modelo presupone que las premisas y conclusiones de los argumentos consisten en proposiciones categóricas completas que comprenden una estructura sujeto-predicado clara, lo cual no siempre es el caso cuando se trabaja con texto informal generado por usuarios en redes sociales.

El ejemplo 3.4 muestra una Justificación etiquetada como Hecho y una Conclusión etiquetada como Política. El punto de vista principal del mensaje debe identificarse primero como Conclusión, y luego cualquier parte del tweet que cumpla la función de proporcionar razones para ese punto de vista se identifica como Justificación. Un patrón típico frecuente en muchos *tweets* es expresar un mandato o política que debe seguirse, generalmente en forma de una frase o hashtag usando el modo imperativo, y un Hecho (o, con menos frecuencia, otro mandato) que apoya y busca explicar por qué debe seguirse ese mandato.

Victims of Illegal Alien Crime describe heartbreak, frustration
 (H) #BuildTheWall #ProtectAmerica #EndChainMigration
 #EndIllegalBirthrightCitizenship #NeverForget the American Victims of
 Illegal Alien Migration (P)

Fig. 3.4: Ejemplo de un *tweet* argumentativo etiquetado con Justificación (en azul) y Conclusión (en rojo). La justificación es de tipo **H**echo, mientras que la conclusión es de tipo **P**olítica.

3.1.3 Componentes específicos del dominio: Colectivo y Propiedad

Todos los mensajes de odio están dirigidos contra un grupo específico determinado por una característica protegida, por definición. Usualmente, el contenido del mensaje es asociar a este grupo, directa o indirectamente, con una propiedad negativa, con una acción o consecuencia indeseable o con algún problema concreto. En tanto y en cuanto esta propiedad negativa sea explícita y se mencione en el contenido del tweet, la etiquetamos. Si la propiedad negativa se asume implícitamente, no se asocia directamente con el colectivo, o, por ejemplo, se define por la negativa (la ausencia del colectivo genera una propiedad positiva), entonces no la anotamos. Esta información tiene como objeto orientar la elaboración de una contranarrativa que tome como estrategia negar la relación entre el colectivo y la propiedad, consecuencia o problema con el que se lo busca relacionar.

La información etiquetada en este paso puede ser utilizada para definir una sub-tarea independiente: el reconocimiento automático de qué se está diciendo acerca de un colectivo determinado, la cual puede aportar, a su vez, a un análisis dirigido hacia un grupo en particular y, potencialmente, a la generación orientada a defender a un grupo concreto.

@user Time to leave the uk commonwealth and Europe that would end
 immigration people do not want more refugees enough is enough

Fig. 3.5: Ejemplo de un Colectivo y Propiedad anotados en un tweet. El colectivo se muestra en rojo mientras que la propiedad se muestra en azul

El ejemplo 3.5 muestra una anotación de un par Colectivo-Propiedad. La Propiedad es cualquier concepto negativo, adjetivo, consecuencia o aspecto de la realidad que esté asociado con el objetivo del mensaje de odio. En este caso, el tweet afirma que el Colectivo inmigrantes tiene la Propiedad de “no ser deseados por la gente”.

Para realizar el etiquetado, requerimos que ambos componentes argumentativos se encuentren presentes y referenciados explícitamente. Si se menciona una Propiedad pero no se la asocia claramente a un Colectivo, o viceversa, entonces no anotamos en ambos casos, ninguno de los dos componentes. Los casos en los que sucede esto son diversos, pero podemos catalogarlos en tres grandes categorías:

- el Colectivo objetivo del odio se referencia implícitamente y/o no se menciona explícitamente. La propiedad asignada se referencia de modo complejo y con una reconstrucción necesaria.
- la Propiedad asignada al Colectivo se referencia implícitamente o se deduce de otro tipo de información contextual ausente en el texto.
- se pone el énfasis en una acción que se debe llevar a cabo y no en asociar una Propiedad al Colectivo. La Propiedad podría potencialmente ser reconstruida apelando

a encontrar la motivación de estas acciones anunciadas, pero no puede etiquetarse explícitamente en el texto del tweet.

`Good this makes it a safe country immigrants can now go home`

Fig. 3.6: Ejemplo de un *tweet* donde no debe etiquetarse ni el Colectivo ni la Propiedad. En este caso, la propiedad mencionada se asocia a la ausencia del Colectivo mucho más que al Colectivo en sí, por lo que la Propiedad se define de manera implícita o indirecta y no se anota.

El ejemplo 3.6 muestra un caso del primer tipo, en el que en lugar de definir una propiedad negativa asociada con el colectivo objetivo, el usuario define una Propiedad positiva asociada con la ausencia de ese Colectivo. El ejemplo 3.7 muestra un caso del segundo tipo, en el que una Propiedad negativa se asocia con el Colectivo objetivo, pero de una manera indirecta que debe ser reconstruida utilizando información contextual no incluida en el mensaje de odio. En este caso, la referencia se hace a través de la mención de “Operación SOAR”, una operación realizada por el organismo policial que controla la inmigración y las fronteras en los Estados Unidos, que se dirige específicamente a inmigrantes registrados como delincuentes sexuales y por el hashtag “StopTheInvasion”, que se refiere a una narrativa construida contra los inmigrantes como si hubiera un plan coordinado para invadir un país. Si bien se referencia al colectivo de “Inmigrantes”, éste no se menciona explícitamente.

`ICE officers arrest 32 sex offenders on Long Island as part of 'Operation SOAR' :link: #StopTheInvasion #SecureTheBorde`

Fig. 3.7: Ejemplo de un *tweet* donde no debe etiquetarse ni el Colectivo ni la Propiedad. En este caso, el colectivo no se menciona explícitamente sino que se referencia indirectamente a través de otros conceptos supuestamente asociados.

`Canada is an immigrant country Don't change it to refugee country please`

Fig. 3.8: Ejemplo de un *tweet* donde no debe etiquetarse ni el Colectivo ni la Propiedad. En este caso, el foco del mensaje está puesto en una acción que debe ser tomada y no en asociar una Propiedad al Colectivo objetivo del odio.

El ejemplo 3.8 muestra un caso del último tipo en el que el punto de vista principal del tweet es una acción que debe llevarse a cabo y no hay ninguna mención explícita de ninguna Propiedad.

3.1.4 Relación argumentativa: el Pivot

Wagemans [2016] clasifica los argumentos como de sujeto o predicado si sus premisas tienen un elemento común en el sujeto o en el predicado, y como de segundo orden o primer orden si necesitan una reconstrucción usando una oración auxiliar como “es verdadero” (T) o no, respectivamente. Encontramos que esta clasificación es difícil ya que el elemento común no siempre forma parte del mismo componente sintáctico e incluso cuando lo hace, puede cumplir múltiples roles, por lo que no es fácil establecer una correspondencia sintáctica entre justificación y conclusión.

Why do foreign individual dump money (and refugees) into our country? We don't need their money and their programs.

Fig. 3.9: Ejemplo de un *tweet* con un Pivot etiquetado. La Justificación se muestra en azul mientras que la conclusión se muestra en rojo. El Pivot etiquetado se muestra subrayado.

Nice tweet , Joyce, Truth is they flee Iran etc but want to bring their hate to the Eu even in refugee camps Christians not safe.

Fig. 3.10: Ejemplo de un *tweet* sin un Pivot anotado. La Justificación se muestra en azul mientras que la Conclusión se muestra en rojo. El vínculo entre las dos premisas se basa en la relación entre “traer su odio” y que los “Cristianos no están seguros”.

Sin embargo, asumimos que puede haber indicios textuales de esta transferencia, en forma de un elemento común a ambos componentes. Llamamos a este elemento el Pivot. Identificamos al Pivot como dos secuencias de palabras, una para cada premisa, que pueden estar relacionadas con el elemento que esas premisas tienen en común.

Esta relación generalmente no es única; el terreno común subyacente entre las premisas podría expresarse de diferentes formas o podría presentar múltiples aspectos señalados por diferentes palabras. Siempre que este elemento sea explícito en el texto (podría no serlo), lo anotamos.

El Pivot tiene relación con las categorías de argumentos de primer y segundo orden definidas en el Procedimiento de Identificación de Tipos de Argumentos [Wagemans, 2021]. Si un argumento se considera de primer orden, significa que el elemento común entre las premisas debe ser explícito (por definición, debe ser de la forma *a* es *X* porque *a* es *Y* o *a* es *Z* porque *b* es *Z*). Para un argumento de segundo orden, puede haber un Pivot explícito o no. Si bien desde un punto de vista taxonómico resulta importante conocer la forma y el cuadrante del argumento analizado, consideramos que desde el punto de vista de la generación de contranarrativas puede resultar aún más útil conocer el elemento común, que no siempre se presenta con el mismo rol sintáctico, ni siquiera con las mismas palabras, requiriendo, a veces, asumir conocimientos o relaciones implícitas y, por lo tanto, constituyendo una tarea difícil en sí, más allá de la clasificación.

A diferencia del resto de los componentes, el Pivot no fue concebido como un elemento asociado a una estrategia de respuesta determinada. Queremos observar su impacto en la generación más allá de la adopción de estrategias en concreto.

El ejemplo 3.9 muestra un *tweet* con un Pivot. En este caso, el usuario plantea que el “dinero” de los (programas) extranjeros fomenta la inmigración para concluir que ese “dinero” no es necesario. Todos los *tweets* presentan algún aspecto que vincula la Justificación con las Conclusiones, pero no siempre esa relación se menciona directamente. El ejemplo 3.10 muestra un *tweet* donde no se etiquetó un Pivot. El vínculo entre las premisas se basa en las suposiciones implícitas de que el odio que supuestamente traen a la UE es contra los cristianos y que, debido a ese odio, los cristianos no están seguros. Pero para reconocerlo, la relación debe ser reconstruida utilizando información implícita definida por el contexto. De lo contrario, es imposible establecerla. En estos casos, no etiquetamos el Pivot por simplicidad. Requerimos que la relación entre las dos frases que constituyen un Pivot sea directa y fácil de detectar.

3.1.5 Tipos de Proposición

Wagemans propone “una caracterización de los tipos de argumentos basada en la combinación de los tipos de proposiciones que instancian” Wagemans [2016]. Estos Tipos se toman de la teoría del debate Schut [2014], donde se hacen tres distinciones sobre las proposiciones: (1) Política (P), (2) Valor (V) y (3) Hecho (F, por la denominación en inglés *Fact*).

Etiquetamos nuestras proposiciones utilizando los mismos tipos y añadimos a nuestro manual de anotación diferentes pautas sobre cómo reconocer cada una:

- Una proposición de tipo Política es un mandato que a menudo se expresa como una orden, un imperativo o una acción o acciones que necesitan llevarse a cabo en el dominio público.
- Una proposición de tipo Hecho es una afirmación sobre la cual se presupone un valor de verdad absoluto (o bien puede decirse que es verdadera o bien puede decirse que es falsa).
- Una proposición de tipo Valor es una opinión que conlleva una participación explícita del hablante en la afirmación expresada utilizando adjetivos de opinión o verbos de pensamiento (“creo”, “pienso”, “me parece”, “siento”, etc). Se distingue de las proposiciones de tipo Hecho debido a que, dada una proposición V de tipo Valor, no es posible afirmar “V es/no es cierta” sin realizar categorizaciones sobre el estado interno del usuario que escribe el mensaje original (por ejemplo, negar el sentimiento de una persona respecto a algún tema).

Cada tipo de premisa está asociada en su concepción a una estrategia para refutarla y el etiquetado debe realizarse considerando si el tipo aporta al desarrollo de tal estrategia. A las sentencias de tipo Política se les debe oponer una política contraria o superadora. A las sentencias de tipo Hecho se les debe oponer la acusación de posible falsedad o se debe pedir pruebas o fuentes que la respalden. A las sentencias de tipo Valor se les debe oponer realizando el carácter de opinión de la afirmación y relativizándola como una opinión xenófoba.

Se reportó, durante el proceso de anotación, que los Tipos “Hecho” y “Valor” son más difíciles de diferenciar entre sí que en relación al Tipo “Política”. Encontramos múltiples ejemplos donde se expresan opiniones como si fueran hechos o donde parece existir una estrategia de confundir la frontera entre ambos de manera deliberada.

Aunque para muchos lectores puede resultar evidente que se trata de una opinión, puede no suceder lo mismo con todos. Como criterio general, proponemos definir el tipo, cuando exista una duda, en función de la adecuación a las posibles estrategias de contranarrativas. Si es coherente, en este sentido, cuestionar la veracidad o pedir pruebas o fuentes a la afirmación realizada, entonces la consideramos como Hecho. Si no es así, entonces la consideramos como opinión.

**Anyone who, ACTIVELY OR PASSIVELY, subscribes to immigration and
especially assimilation is joining the battle to destroy White**

Fig. 3.11: Conclusión de un tweet etiquetada como “Hecho”.

Los tres ejemplos de las figuras 3.11, 3.12 y 3.13 muestran tres Conclusiones, cada una etiquetada con un Tipo distinto.

Looks like no one's running Britain

Fig. 3.12: Conclusión de un tweet etiquetada como “Valor”. La Justificación que apoya esta conclusión plantea “*Our cities are out of control thanks to EU migration :link: an all to often picture of modern UK*” y fue etiquetada como “Hecho”

Go home and stay home!

Fig. 3.13: Conclusión de un tweet etiquetada como “Política”.

El ejemplo de la figura 3.11 muestra una conclusión dentro de un tweet donde el usuario exagera una opinión y la presenta como si fuera un hecho verificable. Aunque podría ser discutible que exprese una Valoración sobre un tópico como “inmigración” o “asimilación”, consideramos todas las expresiones que no hagan explícito su carácter de opinión y que puedan ser atacadas por preguntas críticas que apunten a cuestionar su veracidad o pedir fuentes como de tipo “Hecho”. En efecto, vemos que este fragmento puede ser atacado planteando que no es cierto que suscribir a la inmigración signifique querer destruir a los blancos.

El ejemplo de la figura 3.12 sí explicita el involucramiento y la posición del hablante mediante el verbo “*Looks like*” que relativiza la afirmación posterior como una percepción del observador, lo cual dificulta la generación de una respuesta que cuestione la veracidad de la afirmación, dado que el hablante puede defenderse planteando que es su parecer y su opinión subjetiva.

El ejemplo de la figura 3.13 muestra una Conclusión en modo imperativo que se exploya como orden hacia el colectivo atacado, planteando que se vayan. En este caso, el Tipo asignado es el de “Política” dado que la única respuesta posible parecería ser plantear que los inmigrantes no se vayan o que no tienen por qué irse.

3.1.6 Contranarrativas

Cada tweet argumentativo se asocia con contranarrativas de tres tipos diferentes definidos mediante la aplicación de transformaciones sistemáticas sobre los componentes argumentativos del tweet, y un cuarto tipo que consiste en cualquier contranarrativa que no encaje en ninguno de los otros tres.

Todas las contranarrativas, independientemente de su tipo, siguen las directrices del proyecto *Get The Trolls Out*²:

1. no seas agresivo ni abusivo
2. no difundas odio tú mismo
3. intenta desescalar la conversación
4. responde pensando en una audiencia más amplia que la persona que publicó el tweet original
5. intenta construir una narrativa

² <https://getthetrollsout.org/stoppinghate>

Tenemos que deportar a los inmigrantes ilegales porque están infringiendo la ley y la ley dice que hay que deportarlos .

Fig. 3.14: Tweet de odio anotado con su conclusión (subrayada) y su justificación (en negrita). En este caso, el llamado a una acción contra el colectivo atacado busca ampararse en la ley tomada en abstracto

Se sugirió a los anotadores que intentaran escribir al menos una contranarrativa de cada tipo, pero solo si al hacerlo, esta sonaba de manera natural. De lo contrario, podían dejarla en blanco.

Es posible que dos contranarrativas escritas siguiendo las estrategias de dos tipos de contranarrativa distintas terminen siendo similares. No consideramos que esto resulte problemático y, en todo caso, se sugiere al anotador tratar de hacer énfasis en distintos aspectos de la contranarrativa para que sean lo menos similares, siempre que sea posible.

Las instrucciones para generar cada tipo de contranarrativa son las siguientes:

Negar la Relación Entre Justificación y Conclusión (Tipo A) Este tipo de contranarrativa se basa en atacar el núcleo del razonamiento del usuario negando la relación implícita entre la justificación y la conclusión. Si bien es común la utilización de razonamientos atacables o incluso falaces para los cuales se puede encontrar una contranarrativa de manera directa, existen algunos casos donde es necesario dar una respuesta de manera indirecta o concentrarse en un aspecto particular del razonamiento que es más vulnerable frente a un ataque.

El ejemplo 3.14 muestra un mensaje de odio cuya conclusión llama a una acción contra un colectivo objetivo amparándose en la ley tomada en abstracto. Si se aplica de manera mecánica el procedimiento para la elaboración de una contranarrativa, se podría generar una que plantee que “solo porque la ley plantea algo no significa que sea necesario hacerlo”, pero esta narrativa no resulta muy convincente. Agregando un nivel extra de interpretación, en cambio, se puede plantear que “porque la ley estipule algo, eso no quiere decir que sea correcto desde un punto de vista moral” o “las leyes pueden cambiar si no se ajustan a la realidad de personas inocentes que sufren”.

Negar la Asociación Entre Colectivo y Propiedad (Tipo B) Este tipo de contranarrativa tiene como objetivo atacar la relación entre la propiedad, acción o consecuencia que se está asignando al grupo objetivo y el grupo objetivo en sí.

Atacar la Justificación Según Su Tipo (Tipo C) Reaccionar al discurso de odio no puede basarse únicamente en detectar fallos en el razonamiento, sino que también necesita cuestionar los valores apoyados por el discurso de odio. Este tipo de contranarrativa tiene como objetivo atacar la justificación según su tipo. Si la justificación es un “Hecho”, entonces el hecho debe ser puesto en duda o se deben pedir fuentes para probarlo. Si es de tipo “Valor”, se debe resaltar que la premisa es en realidad una opinión, posiblemente relativizándola como una opinión xenófoba. Si es una “Política”, se debe proporcionar una política alternativa.

Contranarrativa Libre (Tipo D) Si el anotador reconoce una contranarrativa que surge de manera natural pero no encaja en ninguno de los tres tipos anteriores, se le anima a escribirla bajo este cuarto tipo. Este cuarto tipo incluye todas las

TWEET:	user must deport all illegal migrants india already reeling under constant threat of muslim radicals curb population
Justificación:	india already reeling under constant threat of muslim radicals curb population (hecho)
Conclusión:	must deport all illegal migrants (política)
Colectivo:	illegal migrants
Propiedad:	muslim radicals
CONTRANARRATIVA A (Negar relación entre justificación y conclusión)	<i>Deporting illegal migrants will not mitigate the problems with muslim radicals.</i>
CONTRANARRATIVA B (Negar relación entre colectivo y propiedad)	<i>Illegal migrants are not necessarily muslim radicals.</i>
CONTRANARRATIVA C (Negar justificación basandose en su tipo)	<i>It is not true that India is reeling under threat of muslim radicals.</i>
CONTRANARRATIVA LIBRE (Libre)	<i>Deporting illegal migrants without consideration to their circumstances is an inhumane move.</i>

Fig. 3.15: Ejemplos de cada tipo de contranarrativas para un tweet de odio.

contranarrativas que el anotador propone y que no se encuadran en ninguno de los otros tres tipos.

La figura 3.15 presenta un *tweet* de odio con un ejemplo de cada una de las contranarrativas propuestas. El *tweet* presenta un hecho como justificación (que la India se tambalea ante la constante amenaza de la curva de población de radicales musulmanes) y concluye a partir de ésta una política: que hay que deportar a los inmigrantes ilegales. Ante esto, la contranarrativa de tipo A destaca la contradicción entre justificación y conclusión planteando que ésta última no es una solución para la primera. La contranarrativa de tipo B desmiente que la propiedad (radicales musulmanes) se asocie necesariamente al colectivo (inmigrantes ilegales). La contranarrativa de tipo C plantea directamente la falsedad de la justificación dado que es de tipo “Hecho”. Por último, la contranarrativa libre se centra en la conclusión del argumento agregando un elemento nuevo: que actuar así es inhumano.

3.2 Data Statement

Los *data statements* explicitan las condiciones bajo las cuales se generan los datasets públicos que luego serán utilizados por la comunidad científica y el público en general. Se han vuelto una parte fundamental en el desarrollo de recursos anotados en cualquier tipo de tarea. Permiten explicitar posibles sesgos en su construcción así como también hacer una mejor reconstrucción de cómo se obtuvieron, filtraron y procesaron los datos anotados. Para elaborar nuestro *data statement*, seguimos los criterios definidos por Bender

and Friedman [2018].

3.2.1 Curation Rationale

3.2.1.1 HatEval

La base de nuestro conjunto de datos anotados es el corpus HatEval 2019 [Basile et al., 2019], que consta de 13,000 *tweets* en inglés y 6,600 *tweets* en español conteniendo mensajes de odio contra mujeres e inmigrantes. Estos fueron recolectados entre julio y septiembre de 2018. Los autores aplicaron distintos enfoques para recolectar los *tweets*:

- Monitorear posibles víctimas de usuarios odiantes (*haters*).
- Descargar el historial de publicaciones de *haters* identificados.
- Filtrar el *stream* de Twitter con palabras clave.

Para el enfoque basado en palabras clave, los autores utilizaron claves neutrales (en línea con el trabajo de Sanguinetti et al. [2018]), palabras con carga despectiva contra el colectivo atacado y *hashtags* atravesados por una polarización fuerte. Los autores declaran como parte de su objetivo lograr reflejar la diferencia entre discurso de odio, ofensividad [Wiegand et al., 2018] y posicionamiento [Taule et al., 2017].

3.2.1.2 Filtros y Preprocesamiento

Enfocándonos en los *tweets* argumentativos, no anotamos *tweets* etiquetados como “agresivos”, que consisten principalmente en lenguaje abusivo (descalificaciones, insultos, exhortaciones a la acción y otros tipos de ataques) y *tweets* dirigidos contra individuos específicos, porque notamos que estos eran casi exclusivamente no argumentativos. También filtramos todos los *tweets* contra mujeres, ya que notamos que la mayoría también eran abusivos y no argumentativos. Después de estos filtros, quedó un corpus de 970 *tweets* en inglés y 196 *tweets* en español, con un total de 25,413 palabras en inglés y 5,436 en español y una proporción significativa de *tweets* argumentativos. Cabe destacar que los *tweets* de HatEval que fueron filtrados fueron etiquetados implícitamente como no argumentativos.

Al conjunto de datos resultante luego de aplicar este filtro se le aplicó un preprocesamiento previo a ser anotados con la información argumentativa. El preprocesamiento constituye un paso muy importante cuando se trabaja con *tweets* dado que tienden a tener una gran cantidad de caracteres no alfanuméricos, nombres de usuario, *hashtags*, emojis y todo tipo de errores de redacción, ortografía y coherencia. Para el preprocesamiento, seguimos a [Nguyen et al., 2020] y [Polignano et al., 2019]: usamos una estrategia suave de normalización que consiste en:

- Limitar la repetición de caracteres a un máximo de tres.
- Las menciones a usuarios se reemplazan por una mención genérica: @user para el inglés y @usuario para el español.
- El caracter de *hashtag* (#) se reemplaza por un token especial *hashtag* seguido del texto del *hashtag* original dividido en palabras siempre que sea posible. Por ejemplo, “#NoALaInmigración” se transforma en “*hashtag* no a la inmigración”.

- Los emojis son reemplazados por su representación textual utilizando la librería *emoji*³, y a su vez se los encierra en un token especial `emoji`.
- Los hipervínculos se reemplazan por el token especial `:link:`

Aunque los anotadores fueron instruidos para intentar en lo posible anotar palabras enteras anteceditas y precedidas por un espacio, en algunos casos esto no sucedió así. Encontramos casos, especialmente *hashtags*, donde varias palabras se escriben juntas y el anotador solo anota una o un subconjunto de estas. Para estos casos, previo a realizar el preprocesamiento anterior, realizamos una división automática de la palabra en la cual el fragmento anotado y los fragmentos anteriores y posteriores, si los hubiera, se separan cada uno como palabra nueva, de manera que cada palabra pueda tener solo una etiqueta. Considerando que luego los *hashtags* son divididos de todas maneras como parte del preprocesamiento, este procedimiento tuvo poco impacto final en el *dataset*.

3.2.2 Equipo de anotación

El proceso de anotación fue llevado a cabo por dos anotadores varones de entre 25 y 35 años en tres etapas. En las primeras dos etapas, los anotadores anotaron 10 y 20 *tweets* en inglés, respectivamente, con la información argumentativa y sus respectivas contranarrativas y luego compararon los resultados tratando de sistematizar criterios en donde hubieran diferencias, incorporando estos criterios al manual de anotación y pensando cómo éste puede generalizarse a casos distintos. La tercera etapa consistió en la anotación con los componentes argumentativos y las contranarrativas, de 170 *tweets* en inglés por parte de un anotador y 770 *tweets* en inglés y 196 *tweets* en español por parte del otro. Adicionalmente, el primer anotador anotó otros 120 *tweets* elegidos al azar de entre los 770 que anotó el segundo anotador en inglés, sólo con la información argumentativa. Esta anotación adicional, junto con los 30 *tweets* anotados en las primeras dos etapas, fue utilizada para calcular el acuerdo entre anotadores.

El primer anotador que anotó, finalmente, un 27 % del *dataset*, no recibió paga sino que realizó las anotaciones como parte integrante del equipo de investigación. El segundo anotador, que anotó un 83 % (con un 10 % de superposición), recibió una paga de 25 USD por hora, por un trabajo realizado en 80 horas en total.

Las anotaciones fueron realizadas utilizando la herramienta de anotación libre *brat* [Stenetorp et al., 2012].

Priorizamos tener la menor cantidad de anotadores haciendo la mayor cantidad posible de anotaciones. Nuestra hipótesis es que cuántos más anotadores haya, más difícil es alcanzar un criterio uniforme que pueda ser entendido de la misma manera por todos. Por lo tanto, menos anotadores haciendo más trabajo deberían llevar a anotaciones más fiables y a un mejor acuerdo entre anotadores.

El tiempo promedio que los anotadores tardan en etiquetar un tweet es de aproximadamente 3 minutos por ejemplo. El tiempo de anotación varía dependiendo de si el *tweet* es argumentativo o no. Para los *tweets* argumentativos, el tiempo promedio es de alrededor de 5 minutos, mientras que para los *tweets* no argumentativos, el tiempo promedio es de menos de 1 minuto.

³ <https://github.com/carpedm20/emoji/>

3.2.3 Entorno de la extracción de los datos

Los tweets fueron extraídos de la red social Twitter utilizando la API de la plataforma para descargarlos. Los autores no reportan haber utilizado ningún otro criterio para filtrar los mensajes más allá de las tres estrategias mencionadas en la sección 3.2.1. Observando el *dataset* presentado, podemos reconocer distintos estilos y dialectos, siendo muy comunes las frases incompletas e incoherentes sintáctica y semánticamente. Los usuarios frecuentemente hacen referencia a hechos o información ausentes en el mensaje que sólo pueden ser reconstruidos mediante el conocimiento específico particular de una región y un período de tiempo determinados. En efecto, podemos encontrar referencias a distintos países y localidades, incluso no necesariamente restringidas a aquellas de habla hispana o inglesa.

3.2.4 Variedad dialectal y demografía de los anotadores

Ambos anotadores son hablantes nativos del español en su variedad dialectal rioplatense con un alto nivel oral y escrito en inglés (nivel C2 - Proficient de acuerdo al Marco Común Europeo de Referencia - MCER o CEFR por las siglas en inglés).

El anotador que realizó la mayoría de las anotaciones (770 en inglés y 196 en español) es estudiante universitario de la Licenciatura en Filosofía. El otro anotador es el autor de este trabajo, estudiante de Doctorado en Ciencias de la Computación.

Ninguno de los anotadores ha sufrido directamente ataques xenófobos ni pertenece al colectivo objetivo del odio.

3.3 ASOHMO

Las anotaciones de los componentes argumentativos fueron publicadas en el dataset “Estructura Argumentativa de Mensajes de Odio en Línea” (ASOHMO por sus siglas en inglés, *Argumentative Structure Of Hate Messages Online*)⁴. Junto al *dataset* presentamos también los *scripts* necesarios para reproducir los experimentos realizados, calcular las estadísticas e incluso configurar un servidor de *brat* para continuar con la anotación, pudiendo incluso redefinir categorías. A continuación, presentamos el acuerdo entre anotadores para la anotación de los componentes argumentativos y las estadísticas del dataset.

3.3.1 Acuerdo entre anotadores

	Argum.	Colectivo	Propiedad	Pivot	Justif.	Concl.	Tipo de Conc.	Tipo de Just.
κ	.85	.64	.60	.52	.62	.64	.60	-.03

Tab. 3.1: Valores del acuerdo entre ambos anotadores para los distintos tipos de anotación: argumentativo o no argumentativo; colectivo y propiedad; pivot; justificación y conclusión y tipo de justificación y tipo de conclusión. Utilizamos el κ de Cohen sobre los 150 tweets etiquetados por ambos anotadores

Calculamos el acuerdo entre anotadores para evaluar la reproducibilidad de las anotaciones y la viabilidad de la identificación automática. Para esto, utilizamos los 150 *tweets*

⁴ <https://github.com/ASOHMO/ASOHMO-Dataset>

anotados por ambos anotadores, tal como se explica en la sección 3.2.2. El tamaño de la muestra para este cálculo se eligió después de determinar que es proporcionalmente mayor que varios trabajos previos analizados: Bosc et al. [2016] usaron 100 *tweets* para calcular el acuerdo sobre un conjunto de datos de 4000, mientras que Dusmanu et al. [2017] usaron 100 *tweets* para su primer conjunto de datos de 1887 *tweets*, 80 *tweets* para su segundo conjunto de datos de 1459 *tweets* y el tercer conjunto de datos entero, de 368 *tweets*. Luego, el acuerdo en cada categoría argumentativa se calculó utilizando la métrica κ de Cohen Cohen [1960]. El coeficiente κ de Cohen es una medida estadística que evalúa el grado de acuerdo entre dos evaluadores (o jueces) al clasificar elementos de la misma muestra en categorías, teniendo en cuenta la posibilidad de que parte de ese acuerdo se deba al azar. El acuerdo se calculó sobre una etiqueta binaria única por tweet para la categoría Argumentativo vs. No Argumentativo y para las categorías de Tipo de Conclusión y Justificación, usando una etiqueta con tres valores posibles que representan *hecho*, *valor* y *política*. Para todas las demás categorías, el acuerdo se calculó sobre la base de una etiqueta binaria asignada a cada palabra, marcando si pertenece a la categoría argumentativa analizada o no.

En la Tabla 3.1 podemos ver que los anotadores alcanzan un nivel sustancial de reproducibilidad, alrededor de $\kappa = ,6$ para Colectivo, Propiedad, Justificación, Conclusión y Tipo de Conclusión y ,85 para la distinción entre *tweets* Argumentativos o No Argumentativos. En contraste, el Pivote presenta un nivel moderado de acuerdo entre anotadores, y el Tipo de Justificación no presenta ningún acuerdo.

Para calcular el acuerdo de Colectivo y Propiedad, seguimos un criterio similar al de Visser et al. [2021]: al comparar dos bloques de palabras anotados, si al menos el 50 % de las palabras en el bloque más pequeño se superpone con las palabras marcadas en el otro, entonces se considera un acuerdo. Por ejemplo, si un anotador marcó “*the damage illegals do*” como una Propiedad asociada a un Colectivo y el otro anotador marcó solo “*damage*” como una Propiedad, consideramos que ambos ejemplos fueron anotados con la misma Propiedad igualando ambas anotaciones a las del ejemplo más corto, en este caso, el que estaba anotado como “*damage*”.

3.3.1.1 Análisis de desacuerdos

Los anotadores reportaron tres dificultades principales al anotar el corpus. La primera tiene que ver con identificar el Tipo de las proposiciones y, en particular, el tipo “Hecho” del tipo “Valor”. La segunda es etiquetar premisas que puedan tener dentro estructuras anidadas y, por lo tanto, puedan combinar distintos tipos de sentencias. La tercera es identificar el Pivote. Los anotadores también reportaron una dificultad menor, una vez encontradas las dos premisas principales, en clasificarlas como justificación o conclusión. Encontramos algunos casos donde los anotadores podían intercambiar las etiquetas de justificación y conclusión, y notamos que estos a veces coincidían en identificar las partes de un *tweet* pero no estaban de acuerdo en cuál era cada una.

Identificamos dos tipos de desacuerdo sobre el Tipo de Justificación. El más frecuente es entre el tipo “Valor” y el tipo “Hecho” y parece deberse a la existencia de muchos casos con cierto nivel de ambigüedad que podrían ser de uno u otro tipo según cómo se los interprete. La Figura 3.16 muestra un ejemplo que ilustra esta dificultad. Podemos observar que además del desacuerdo respecto al Tipo de la proposición, también hay cierto desacuerdo respecto a la identificación de la Justificación y la Conclusión, dado que un anotador incluyó la frase “*And increasing intolerance towards who?*” dentro de la

@user The increase of gun ownership has led to an overall decrease in homicides. And increasing intolerance towards who? (H) People smugglers? Drug cartel members? People who cheat the system essentially flip the bird to legal immigrants waiting in (H)

@user The increase of gun ownership has led to an overall decrease in homicides (H). And increasing intolerance towards who? People smugglers? Drug cartel members? People who cheat the system essentially flip the bird to legal immigrants waiting in (V)

Fig. 3.16: Desacuerdo entre anotadores respecto al tipo de justificación. El primer anotador cree que la justificación (en azul) es un *Hecho*, mientras que el anotador de abajo considera que el hablante está expresando una opinión. La conclusión se destaca, a su vez, en rojo.

Remember Kate Steinle today. It is the third anniversary of her death We must not forget. #KateSteinle#IllegalAliens #OpenBorders#BuildThatWall #MondayMorning#ImmigrationReform #ImmigrationIsAWeapon

Fig. 3.17: Ejemplo de desacuerdo respecto al tipo de la justificación. Ambos anotadores identificaron de igual manera la Justificación (en azul) y la Conclusión (en rojo). Uno de ellos etiquetó la Justificación como “Hecho” mientras que el otro la etiquetó como “Política”

Conclusión mientras que el otro anotador la incluyó dentro de la Justificación. De este modo, el anotador que considera que esta frase es parte de la Justificación le asigna el tipo “Valor” mientras que el otro anotador, sin considerar esa frase, le asigna el tipo “Hecho”. Al medir, tanto el acuerdo como el rendimiento de las predicciones realizadas con modelos automáticos, es importante considerar el impacto que tiene la tarea de identificación de la Justificación y la Conclusión y cómo los desacuerdos o errores en esta etapa pueden propagarse e impactar en esta etapa sucesiva de identificación del Tipo de proposición.

El segundo tipo de desacuerdo respecto al Tipo es menos frecuente y se da entre el Tipo “Hecho” y el Tipo “Política”. El ejemplo de la figura 3.17 muestra un ejemplo en el cual la Justificación contiene frases de tipo factual como “*It is the third anniversary of her death*” que pueden sugerir una proposición de tipo “Hecho” con frases imperativas como “*We must not forget*” y “*Remember Kate Steinle today*” que pueden sugerir una proposición de tipo “Política”.

@user @user sanctuary cities are against the law. please shut them down & arrest/prosecute all **criminal** governors & mayors

Fig. 3.18: Ejemplo de un desacuerdo entre anotadores sobre el Pivot. Un anotador está subrayado mientras el otro está en negrita. La justificación está marcada en azul y la conclusión en rojo.

Al inspeccionar ejemplos de desacuerdo entre anotadores para el Pivot, como se muestra en la Figura 3.18, encontramos que en muchos casos ambas anotaciones podrían considerarse precisas, ya que puede haber más de una posibilidad para que los anotadores etiqueten. Considerando que la relación es muy profunda dentro de las capas de significado, los anotadores pueden interpretarla como señalada por diferentes características más o menos superficiales, y como consecuencia, pueden etiquetar diferentes secuencias de palabras mientras consideran la misma relación.

@user @user you come with the usual lies an insults. **Fact is that mass immigration into Ireland has been going on for decades, most illegal and from other EU countries, still trans-formative. All the people seeking asylum**

Fig. 3.19: Desacuerdo entre anotadores sobre la Justificación y la Conclusión. Para un anotador, la Justificación está en azul y la Conclusión en rojo, mientras que para el otro anotador, la Justificación está en negrita y la Conclusión subrayada. Mientras que ambos están de acuerdo en cómo dividir el *tweet* en dos proposiciones, difieren en cuál de estas cumple el rol de Justificación y cuál el de Conclusión

@user @user The idea is to bring in the ‘‘dreamers’’ so that they vote for Democrats because Dems know they have to import their voters. That is literally the only reason the Democrats care about this issue. In the meantime, YES THEIR PARENTS

Fig. 3.20: Ejemplo de desacuerdo entre anotadores respecto al Colectivo y la Propiedad. Un anotador no etiquetó ninguno de los dos. Para el otro anotador mostramos el Colectivo en rojo y la Propiedad en azul.

En este caso en particular, podemos pensar al argumento como de la forma “Las ciudades santuario son contrarias a la ley, entonces las ciudades santuario (*them*) deben ser cerradas y sus autoridades arrestadas” de la forma “a es X” entonces “a es Y” y por lo tanto, del primer cuadrante α según la taxonomía de Wagemans, pero también podemos pensar que se toma de forma implícita, como argumento secundario, que “las ciudades santuario son contrarias a la ley, entonces sus alcaldes y gobernadores también están contra la ley” (y son criminales, lo que da lugar a que “*criminal*” sea considerado Pivot), de la forma “a es X” entonces “b es X”, del segundo cuadrante β según la taxonomía de Wagemans.

Encontrar patrones en los desacuerdos entre anotadores puede utilizarse para redefinir categorías Teruel et al. [2018]. En una segunda fase de anotación, estaremos redefiniendo la categoría de Pivot para obtener más acuerdo entre los anotadores. Entendemos que este elemento es particularmente desafiante, porque señala una relación muy profunda y su correspondencia con fenómenos textuales superficiales puede no ser directa o puede ser múltiple. Planteamos redefinir el Pivot para que los anotadores puedan anotar más de una instancia del mismo en un *tweet*, vinculando cada elemento del par con su contraparte.

El ejemplo de la figura 3.19 muestra un ejemplo donde los anotadores acuerdan en cómo realizar la división del *tweet* en dos proposiciones, pero difieren en la asignación que hacen de cuál es la Justificación y cuál es la Conclusión. En este caso, mientras que para un anotador la Conclusión es que lo que dice otro usuario son mentiras e insultos y la Justificación es que la inmigración masiva en Irlanda lleva décadas y son todos buscadores de asilo, para el otro anotador es exactamente al revés: la Conclusión es la denuncia a la inmigración y la Justificación es que los dichos del otro usuario son mentiras e insultos.

Al analizar los desacuerdos en torno al Colectivo y la Propiedad encontramos que un desacuerdo frecuente consiste en que uno de los anotadores etiqueta un par de Colectivo y Propiedad mientras que el otro no lo hace. El ejemplo de la figura 3.20 muestra un *tweet* donde un anotador considera que no hay un Colectivo referenciado y, por lo tanto, no hay ninguna Propiedad que se le asigne, mientras que el otro considera que los “soñadores”, en referencia a los inmigrantes, son el Colectivo objetivo del odio y se les asigna la Propiedad de ser votantes de los demócratas.

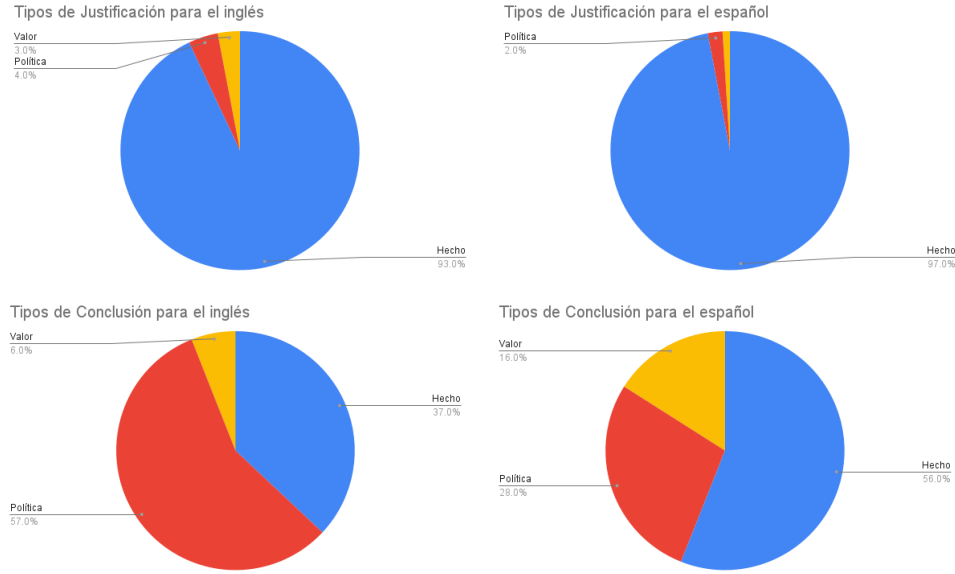


Fig. 3.21: Porcentaje de Justificaciones y Conclusiones etiquetados con cada Tipo de premisa, para el inglés y el español. Puede observarse cómo para la justificación, el tipo “Hecho” es ampliamente mayoritario, mientras que en la Conclusión las clases “Política” y “Hecho” se encuentran en proporciones más parecidas y, si bien sigue siendo minoritaria, la clase “Valor” también se encuentra en una proporción mayor que en la Justificación

3.3.2 Estadísticas

	No-Arg	Colectivo	Pivot
Inglés	25.3 %	58.2 %	45.1 %
Español	26.5 %	61.1 %	37.5 %

Tab. 3.2: Porcentaje de *tweets* etiquetados como No-Argumentativos y que fueron anotados con Colectivo-Propiedad o con Pivot

La Tabla 3.2 muestra el porcentaje de *tweets* No-Argumentativos por cada *dataset* y el porcentaje de *tweets* argumentativos que están etiquetados con Colectivo y Propiedad, y con Pivotes. Podemos observar que las proporciones de *tweets* argumentativos y no argumentativos son similares para el inglés y el español y en ambos casos son mayoritarios los *tweets* argumentativos, lo que demuestra que fue efectivo el proceso de filtrado respecto al dataset HatEval original. Observamos a su vez, que para el par Colectivo-Propiedad y para el Pivot, también son similares las proporciones entre el inglés y el español. En este caso, observamos que existe una mayoría de *tweets* que tienen una Propiedad explícita para el colectivo atacado, mientras que en el caso del Pivot, es menos de la mitad la cantidad de *tweets* que muestran este componente anotado.

Los gráficos de la figura 3.21 muestran el porcentaje de Justificaciones y Conclusiones etiquetadas como Hechos, Políticas y Valores. En este caso, es muy distinta la proporción entre estas tres clases en uno y otro tipo de premisa. Mientras que en la Justificación, la cantidad etiquetada como de tipo “Hecho” es mayoritaria (y esto se observa aún más pronunciadamente para los *tweets* en español), en la conclusión existe una distribución más

	Inglés		Español	
	#	%	#	%
A	701	97 %	140	97 %
B	350	48 %	88	61 %
C	671	93 %	127	88 %

Tab. 3.3: Cantidad de contranarrativas según su tipo (A, B o C) y porcentaje de *tweets* que tienen contranarrativa de ese tipo anotada, para *tweets* y contranarrativas en inglés y en español.

proporcional entre el tipo “Hecho” y el tipo “Política”, siendo el tipo “Política” la clase mayoritaria para el idioma inglés y el tipo “Hecho” la clase mayoritaria en español. El tipo “Valor” es minoritario en ambas premisas, aunque en la conclusión es proporcionalmente más significativo, especialmente en español, donde constituye un 16 % de los casos.

3.4 CONEAS

Las anotaciones de las contranarrativas, asociadas a sus respectivos *tweets* etiquetados con sus componentes argumentativos, se publicaron en el *dataset* “Contranarrativas Explotando la Estructura Argumentativa” (CONEAS, por las siglas de su nombre en inglés, COUNTERNARRATIVES EXPLOITING ARGUMENTATIVE STRUCTURE)⁵.

3.4.1 Estadísticas

Considerando que los anotadores fueron instruidos para no anotar una contranarrativa de un determinado tipo si sentían que la estrategia no lograba generar una que fuese satisfactoria, no todos los *tweets* tienen contranarrativas de todos los tipos asociados a ellos. A su vez, las contranarrativas de tipo B están limitadas por una restricción extra: que el tweet original debe tener anotado a su vez, un colectivo y una propiedad. Podemos observar de la tabla 3.2 que sólo el 58,2% de los ejemplos en inglés y el 61,1 % de los ejemplos en español cumple con esta condición.

La tabla 3.3 muestra la cantidad de contranarrativas de cada tipo (A, B o C) y el porcentaje de *tweets* que están anotados con ese tipo de contranarrativa para los *dataset* en inglés y en español. Las contranarrativas de tipo A se encuentran presentes en casi la totalidad de los *tweets*. Respecto a las contranarrativas de tipo B, podemos observar que, mientras que cerca del 10% de los ejemplos en inglés tienen anotado un colectivo y una propiedad pero no una contranarrativa de tipo B, en español, casi la totalidad de las contranarrativas que tienen un colectivo y una propiedad anotadas tienen también una contranarrativa de tipo B. Las contranarrativas de tipo C se encuentran anotadas también en un porcentaje alto de ejemplos pero con menor presencia que las de tipo A. A diferencia de las contranarrativas de tipo B, las de tipo A y C tienen una proporción similar entre ambos idiomas.

El conjunto de datos consta de un total de 1722 contranarrativas para 725 *tweets* argumentativos en inglés y 355 contranarrativas para 144 *tweets* en español (un promedio de 2.38 y 2.47 por tweet, respectivamente).

⁵ <https://github.com/ConeasDataset/CONEAS>

4 Identificación Automática de Componentes Argumentativos

Realizamos varios experimentos para evaluar la viabilidad de identificar automáticamente diferentes aspectos argumentativos usando Grandes Modelos de Lenguaje en entornos monolingües (inglés) y multilingües (inglés y español). No realizamos experimentos monolingües utilizando solo la porción en español del conjunto de datos debido a que su tamaño es demasiado pequeño para impactar en el entrenamiento de los modelos.

La distribución de ejemplos por partición para cada configuración lingüística puede verse en la tabla 4.2. Para todos los experimentos monolingües, utilizamos 770 *tweets* en inglés como entrenamiento (79 %), 100 *tweets* en inglés como desarrollo (10.5 %) y 100 *tweets* en inglés como *test* (10.5 %). Los experimentos multilingües fueron de dos tipos: utilizando tanto inglés como español para entrenamiento y *test*, y utilizando inglés para entrenamiento y desarrollo, y español para *test*. En el primer caso, utilizamos 770 *tweets* en inglés y 120 en español como entrenamiento (76.3 % del conjunto de datos), 100 *tweets* en inglés y 26 en español como desarrollo (10.8 %) y 100 *tweets* en inglés y 50 en español como *test* (12.9 %). En el segundo caso, utilizamos 850 *tweets* en inglés como entrenamiento (73 % del total), 120 *tweets* en inglés como desarrollo (10 %) y los 196 *tweets* en español para prueba (17 %).

Realizamos una validación de 3-*folds* utilizando tres particiones aleatorias distintas del conjunto total del dataset, generadas con tres semillas. Luego entrenamos tres modelos utilizando cada una de estas particiones, respetando siempre la proporción adecuada según el experimento tal como muestra la tabla 4.2. Reportamos el promedio del puntaje obtenido para las métricas *F1*, *Precision* y *Recall*, y la desviación estándar para el *F1*. Para todas las tareas que involucran clasificación con etiquetas múltiples, los puntajes se calculan utilizando el promedio a nivel *macro* de todas las etiquetas posibles. De lo contrario, reportamos el puntaje obtenido por la clase objetivo (la clase “positiva” en una clasificación binaria). También reportamos los valores del *F1* por clase para los tres posibles Tipos de premisas: Hecho, Valor y Política.

En todos los casos, probamos 5 valores diferentes para el parámetro *learning rate* (**1e-05**, **2e-05**, **5e-05**, **5e-04** y **5e-06**) y utilizamos el conjunto de datos de desarrollo para implementar una detención temprana (“*early stopping*”) en un máximo de 10 épocas. Reportamos el rendimiento sobre las predicciones hechas en el conjunto de *test* del modelo que haya obtenido el puntaje más alto de este parámetro en el conjunto de validación.

La tabla 4.1 muestra los valores de los hiperparámetros que mantuvimos fijos en todos los modelos entrenados con nuestros ejemplos utilizados en este trabajo.

Batch Size	16
Optimizer	AdamW
Dropout	0.1
Epochs	10
Weight Decay	0.01
Adam ϵ	1e-06
Adam β_1	0.9
Adam β_2	0.99

Tab. 4.1: Hyperparametros utilizados para el entrenamiento de todos los modelos de detección de componentes argumentativos

	Monolingüal	Multilingüal Mixto	Multilingüal Cruzado
Entrenamiento	770 (79 %)	890 (76,3 %)	850 (72,9 %)
Desarrollo	100 (10,5 %)	126 (10,8 %)	120 (10,3 %)
Test	100 (10,5 %)	150 (12,9 %)	196 (16,8 %)

Tab. 4.2: Cantidad de ejemplos y porcentajes de cada partición del dataset utilizada para nuestra experimentación según la configuración del lenguaje. Cada experimento dentro de cada una de estas configuraciones utilizó tres particiones aleatorias conformadas por distintos ejemplos para entrenar tres modelos y luego reportar la *performance* promedio entre estos, pero aunque distintas, cada partición respetó las cantidades y proporciones descriptas en esta tabla

4.1 Modelos

Seleccionamos tres modelos que forman parte del estado del arte en tareas de clasificación de *tokens*: un modelo de dominio general en inglés, un modelo específico de Twitter en inglés y un modelo de dominio general multilingüal. Los tres modelos se basan en la arquitectura original RoBERTa, propuesta por Liu et al. [2019] la cual a su vez se basa en la arquitectura de *transformers* preentrenados BERT, propuesta por Devlin et al. [2018].

4.1.1 RoBERTa

RoBERTa (Robustly Optimized BERT Approach) [Liu et al., 2019] es un modelo de lenguaje basado en *transformers* propuesto como una mejora del modelo original BERT presentado por Devlin et al. [2018]. Ambos modelos comparten una arquitectura que consiste en apilar múltiples capas de *encoders* de *transformers* diseñados para procesar texto en paralelo a una alta velocidad, utilizando mecanismos de atención. Pero RoBERTa propone varias modificaciones que le permiten hacer un mejor uso de una mayor cantidad de datos y obtener una mayor generalización, lo que genera que incremente su rendimiento en una gran variedad de tareas de Procesamiento del Lenguaje Natural.

Una de estas modificaciones consiste en quitar el objetivo de *Predicción de la Próxima Oración* (NSP por las siglas de su nombre en inglés, *Next Sentence Prediction*) que se utilizaba en BERT como “tarea pretexto” para mejorar el entendimiento del modelo respecto a las relaciones entre oraciones. En cambio, RoBERTa se focaliza únicamente en la tarea pretexto de *Modelo de Lenguaje Enmascarado* (MLM por las siglas de su nombre en inglés, *Masked Language Model*). Esta tarea consiste en enmascarar un porcentaje de

los *tokens* del *input* y luego entrenar al modelo para predecir los tokens enmascarados. Al abandonar la tarea de *Predicción de la Próxima Oración*, RoBERTa puede concentrar el uso de los recursos disponibles en la tarea de *Modelo de Lenguaje Enmascarado* lo que resulta en un entendimiento del lenguaje más robusto.

Otra modificación importante respecto a BERT consiste en aumentar la cantidad de datos y tiempo de entrenamiento. RoBERTa fue entrenado por una combinación de múltiples *datasets* de distintos tamaños pertenecientes a distintos dominios sumando un tamaño final de 160GB de datos sin comprimir, a diferencia de BERT que sólo fue entrenado utilizando dos *corpora*, **BookCorpus** [Zhu et al., 2015] y **Wikipedia** sumando un tamaño total de 16GB. El entrenamiento también involucra secuencias de texto más largas y más pasos de entrenamiento, lo que contribuye a una *performance* superior del modelo. RoBERTa también utiliza una técnica de enmascaramiento dinámico a través de la cual cambia en cada época el patrón de enmascaramiento, exponiendo al modelo a una mayor variedad de contextos y previniendo que realice un sobreajuste de determinadas posiciones.

Cuando RoBERTa fue publicado en 2019 se estableció como el nuevo estado del arte en 4 de 9 tareas del *benchmark* GLUE e igualó el estado del arte en otras 2.

4.1.2 BERTweet

BERTweet [Nguyen et al., 2020] es un modelo de lenguaje basado en la arquitectura de RoBERTa especialmente diseñado para procesar y entender texto de redes sociales, particularmente, de Twitter. Esta especialización le permite a BERTweet tener un rendimiento especial en tareas que involucren texto de redes sociales, el cual suele ser más informal, compacto y ruidoso, comparado con otros tipos de texto, y puede estar incompleto, ser incoherente, utilizar lunfardo, abreviaciones o un lenguaje específico de redes sociales, así como también apelar a referencias puntuales del contexto geotemporal que pueden ser difíciles de recuperar. BERTweet demuestra así la capacidad de entender los fenómenos lingüísticos únicos que sólo se encuentran en tweets, como hashtags, menciones de usuario, emojis, etc.

BERTweet está entrenado en un dataset masivo de 850 millones de tweets en inglés, recolectados en un período de tiempo suficiente para captar una variedad amplia de tópicos y usos del lenguaje. Se aplicó un filtro para remover *spam* y contenido que no fuese en idioma inglés y se aplicó un proceso de normalización del texto. El modelo, al igual que RoBERTa, se entrenó únicamente con la tarea pretexto de Modelo de Lenguaje Enmascarado, pero al utilizar como entrenamiento base un *corpus* de *tweets*, BERTweet logra capturar con mayor profundidad las particulares características lingüísticas de los mensajes de esta red social.

4.1.3 XLM-Roberta

XLM-RoBERTa [Conneau et al., 2019] es un modelo de lenguaje multilingual basado en la arquitectura y la metodología de entrenamiento de RoBERTa, pero adaptado al aprendizaje de múltiples idiomas. Fue entrenado con un corpus de 2.5 TB de datos web provenientes de CommonCrawl¹ en 100 idiomas. Los autores redujeron la proporción de textos en inglés, que constituyen la cantidad mayoritaria de textos del repositorio, a una treceava parte del total del corpus para favorecer a lenguajes de menores recursos.

¹ <https://commoncrawl.org/>

En el momento de su publicación, XLM-RoBERTa superó a los modelos de lenguaje multilingües que constituían el estado del arte en una variedad de *benchmarks* de lenguajes cruzados, demostrando, además, un rendimiento particularmente superior para lenguajes de bajos recursos como el urdu o el swahili, lo que sugiere que logra capturar mejor patrones lingüísticos más allá de un lenguaje en particular y, por ende, aplicar conocimientos cruzados entre los lenguajes.

Otra característica distintiva es que, a diferencia de los modelos de lenguaje multilingües anteriores, XLM-RoBERTa no utiliza vocabularios específicos para cada lenguaje, sino que utiliza un vocabulario unificado compartido entre todos los lenguajes, el cual es ampliado en proporción a la cantidad de lenguajes utilizados, agrandando a la vez el tamaño del modelo. Esto permite explicar también su capacidad de aprender representaciones más generales del lenguaje que permiten, a la vez, transferir mejor el conocimiento entre lenguajes. También permite explicar, en definitiva, el mejor rendimiento en lenguajes de menos recursos para los cuales el modelo ha visto una cantidad de ejemplos mucho menor.

Los autores discuten un *trade-off* necesario que hay que hacer durante la fase de entrenamiento para potenciar el rendimiento, entre lenguajes de bajos y altos recursos, demostrando la importancia y la diferencia que pueden ejercer los hiperparámetros en esta situación. Más allá de esto, los autores demuestran que a pesar de tener un buen rendimiento en lenguajes de bajos recursos, XLM-RoBERTa también tiene un rendimiento similar al de los modelos monolingües de los lenguajes de mayores recursos, planteando la posibilidad de que a futuro los modelos de lenguaje multilingües reemplacen a aquellos monolingües.

Utilizando este modelo pre-entrenado como base, realizamos dos tipos distintos de ajuste fino:

- Entrenamiento **Mixto**: entrenamos un modelo base XLM-RoBERTa con una partición de entrenamiento compuesta por ejemplos en inglés y en español, en la proporción detallada en la sección 4. Este modelo fue evaluado con ejemplos en inglés y español.
- Entrenamiento de Lenguaje Cruzado (**XL**): entrenamos un modelo base XLM-RoBERTa pero utilizamos una partición de entrenamiento compuesta únicamente por ejemplos en inglés. Buscamos evaluar la capacidad del modelo de hacer transferencia de aprendizaje de un lenguaje a otro. Concretamente, que aprenda la tarea de identificación automática de componentes argumentativos del inglés, un lenguaje rico en recursos lingüísticos, y luego aplique esos conocimientos para realizar la misma tarea en español, un lenguaje con menos recursos.

4.2 Experimentos

Basándonos en la información argumentativa anotada, definimos cuatro tareas diferentes. Éstas parten de la necesidad de automatizar el proceso de identificación de esta información para que pueda ser utilizada con fiabilidad para la elaboración automática de contranarrativas (o para alguna otra tarea, llegado al caso). Realizamos cuatro experimentos donde entrenamos distintos modelos de lenguaje para que realicen cada una de estas tareas. También realizamos un quinto experimento donde repetimos el entrenamiento realizado en la primera tarea, pero utilizando tamaños incrementales de datos, con el objetivo

de analizar cuánto mejora el rendimiento de los modelos en la medida que existen más datos disponibles. Buscamos evaluar la mejora marginal de ir expandiendo el *corpus* con el fin de discutir la suficiencia o insuficiencia de los recursos disponibles hasta el momento en relación, a su vez, con la dificultad de realizar esta expansión, que requiere de varias horas de trabajo manual de anotación.

En las siguientes secciones describimos en detalle los experimentos realizados.

4.2.1 Predicción de Componentes de manera Individual

Entrenamos diferentes tipos de modelos para recrear automáticamente el proceso de anotación, un componente a la vez: uno para realizar una clasificación binaria de secuencias, para predecir si un tweet es argumentativo o no; cinco modelos para la clasificación de tokens, para predecir para cada palabra si está etiquetada como parte del Colectivo, la Propiedad asociada a ese colectivo, el Pivot, la Justificación o la Conclusión, respectivamente; y dos modelos para la clasificación de secuencias, alimentados solo con el texto correspondiente de la premisa (Justificación o Conclusión), para predecir el Tipo asociado a ella (Hecho, Valor o Política). Los resultados de este experimento se muestran en la tabla 4.3.

	RoBERTa			BERT _{tweet}			XLM-R-Mix			XLM-R-XL		
	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec
Arg./No-Arg.	.89 ±.02	.84	.95	.88±.01	.84	.93	.87±.04	.84	.91	.84±.03	.84	.85
Justificación	.73±.05	.69	.76	.77 ±.05	.75	.78	.76±.05	.71	.81	.75±.01	.71	.80
Conclusión	.55±.02	.60	.51	.61 ±.03	.64	.58	.60±.02	.59	.61	.54±.03	.59	.49
Tipo Just.	.41±.09	.48	.39	.42 ±.09	.48	.41	.35±.05	.34	.37	.33±.03	.33	.35
Tipo Conc.	.58±.05	.62	.57	.65 ±.11	.67	.65	.61±.06	.65	.62	.63±.02	.66	.62
Colectivo	.59 ±.03	.56	.64	.58±.05	.55	.62	.59 ±.06	.58	.60	.27±.07	.41	.21
Propiedad	.46±.04	.52	.41	.47±.03	.50	.43	.50 ±.03	.57	.43	.42±.04	.42	.43
Pivot	.45 ±.04	.52	.41	.40±.08	.43	.39	.39±.08	.42	.38	.33±.08	.41	.27

Tab. 4.3: F1, *Precision* y *Recall* de la clase objetivo en la tarea de detección automática de componentes argumentativos en *tweets*. Cada experimento fue realizado con tres particiones aleatorias a través de una validación de 3 *folds*, por lo que los valores de la tabla representan el promedio de estos tres experimentos. Para el F1, presentamos a su vez, la desviación estandar entre estas corridas disintas. Los mejores resultados del F1 para cada categoría se destacan con negrita.

Podemos observar que la tarea de identificar cuáles tweets son argumentativos obtiene un puntaje satisfactorio en todos los modelos, siendo el mejor el modelo RoBERTa con un puntaje de .89 de F1. Este resultado, por sí solo, demuestra que la tarea de identificar cuáles *tweets* merecen ser contestados y cuáles no, puede ser resuelta de manera automática por un modelo de lenguaje ajustado con una cantidad pequeña de datos. Nos referimos, con “merecer”, a aquellos casos en los que una contranarrativa ajustada (en oposición a una genérica) puede tener un efecto positivo en el entorno en el que pueda aparecer el mensaje de odio.

En general, los componentes con un mayor valor de acuerdo entre anotadores obtuvieron un mejor rendimiento. Se observa una diferencia entre la Justificación y la Conclusión en este aspecto. Mientras que la Conclusión obtuvo un valor incluso mayor que la Justificación en el acuerdo entre anotadores ($\kappa = .64$ vs $\kappa = .62$), todos los modelos mostraron una capacidad predictiva mucho mayor a la hora de identificar la Justificación (con un valor

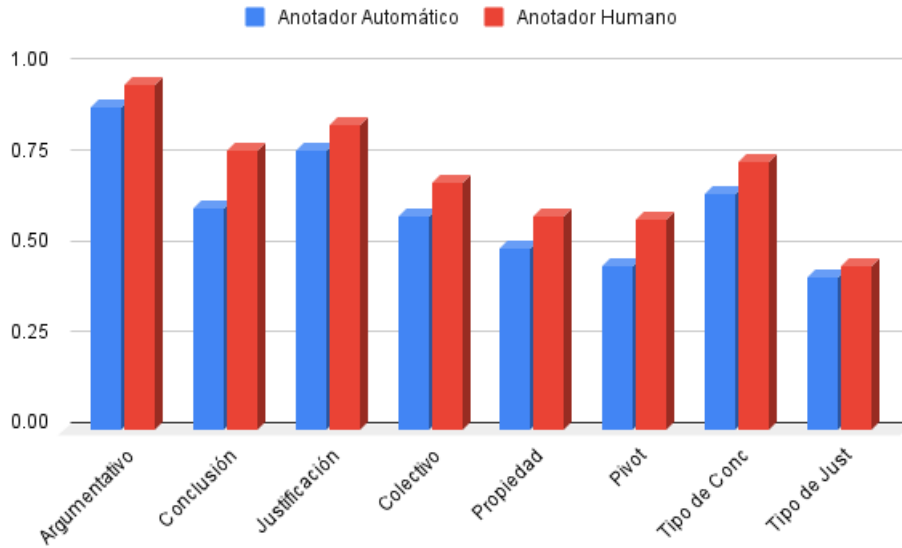


Fig. 4.1: Valores de F1 para “predicciones” hechas por el anotador humano, comparadas con las predicciones realizadas por el modelo de mejor rendimiento (BERTweet para la Justificación, Conclusión y sus Tipos, RoBERTa para detectar *tweets* argumentativos y al Pivot y XLM-Roberta para el Colectivo y la Propiedad).

máximo de F1 de .77 obtenido por el modelo BERTweet) que al identificar la Conclusión (con un valor máximo de F1 de .61 también obtenido por BERTweet).

El resultado obtenido por los modelos para la identificación de los tipos de premisas sí refleja, al igual que en el acuerdo entre anotadores, una mayor capacidad de realizar esta tarea en el caso de la Conclusión (con un acuerdo de $\kappa = .60$) que en el caso de la Justificación (con un desacuerdo de $\kappa = -0.03$). En efecto, la identificación del Tipo de Conclusión obtuvo un valor de F1 de .65 con el modelo BERTweet mientras que la identificación del Tipo de Justificación obtuvo un valor de F1 de .42, siendo el componente con peor rendimiento, a pesar de ser una tarea de clasificación de toda la oración en lugar de ser una tarea de clasificación de tokens que da más lugar al ruido y a los errores.

La tabla 4.4 muestra la matriz de confusión de las predicciones sobre el Tipo de la Justificación hechas por un modelo XLM-RoBERTa-Mix de uno de los tres *folds*. Las columnas se corresponden con las predicciones realizadas, mientras que las filas representan los valores reales de los ejemplos predichos. Por ejemplo, la primera fila muestra que de los 107 ejemplos etiquetados como “Hecho” que fueron evaluados, 101 fueron predichos por el modelo como “Hechos”, ninguno como “Valor” y 6 como “Política”. En la última fila de la tabla observamos el F1 de cada clase. Podemos observar que el mal desempeño del modelo se debe principalmente al bajo rendimiento de este para predecir las clases minoritarias. La métrica *accuracy*, que no considera el desempeño según las clases, como sí hace el F1, para este conjunto de predicciones es de .89, pero el modelo predice casi invariablemente sólo la clase “Hecho”.

El Pivot, que obtuvo el segundo peor valor de acuerdo entre anotadores (con $\kappa = .52$), obtuvo también el segundo puntaje más bajo en términos de su identificación automática (con un F1 de .45 para el modelo RoBERTa). Revisando los errores cometidos por este modelo (ver sección 4.2.1.2), observamos que en muchos casos, éste logra identificar

conceptos relacionados que podrían ser utilizados para transferir conocimiento entre la Justificación y la Conclusión, pero que se encuentran dentro de la misma premisa y, por lo tanto, no cumplen verdaderamente el rol de Pivot. Esto sugiere que los modelos logran captar cierta noción de qué es (dos palabras relacionadas que pueden usarse como nexo común en dos construcciones lingüísticas) pero fallan en conceptualizarlo en el marco del esquema “Justificación-Conclusión” que utilizamos para desglosar los argumentos, posiblemente, debido a que no cuentan con información sobre cuál es la Justificación y cuál la Conclusión y por ende, deben aprender dos tareas a la vez: separar el *tweet* en sus premisas y luego encontrar el nexo entre estas.

Con respecto al Colectivo y la Propiedad, el rendimiento de los modelos de lenguaje ajustados a la tarea de su identificación es proporcionalmente peor en comparación a la Justificación y la Conclusión. Considerando que tienen valores de acuerdo entre anotadores similares ($\kappa = .64$ para el Colectivo y $\kappa = .60$ para la Propiedad), observamos que el *score* de F1 obtenido por los mejores modelos en la tarea de predicción de estos componentes es menor en relación al mismo resultado para la Justificación y la Conclusión (.59 para el Colectivo con los modelos RoBERTa y XLM-RoBERTa y .50 para la Propiedad con XLM-RoBERTa).

Esta diferencia puede deberse, quizás, a la proporción de la clase positiva en uno y otro caso: mientras que la clase positiva en las justificaciones y las conclusiones es de aproximadamente la mitad de las etiquetas totales, en el caso del Colectivo y la Propiedad la clase positiva puede ser una o unas pocas palabras dentro de todo el ejemplo, constituyendo de manera mucho más marcada la clase minoritaria entre las dos clases posibles (1 o 0 para cada palabra indicando si pertenece o no al componente que se quiere identificar). Otra explicación posible para esta diferencia en el rendimiento de los modelos puede ser que, mientras que la Justificación y la Conclusión son componentes generales presentes en muchos modelos de argumentación, el Colectivo y la Propiedad son componentes específicos del dominio del discurso de odio, con menos o incluso nula presencia en otro tipo de textos. Es posible, por tanto, que los modelos de lenguaje preentrenados tengan incorporado internamente algún tipo de conocimiento sobre qué es la Justificación y qué es la Conclusión, que hayan adquirido como parte de su preentrenamiento y que esto les sea de utilidad para realizar la tarea, mientras que es menos probable que ocurra lo mismo para el Colectivo y la Propiedad.

Respecto a los modelos, observamos que aquellos monolingües tienen un rendimiento ligeramente superior respecto a los modelos multilingües, con excepción de la Propiedad, donde el modelo que mejores resultados obtiene es la versión **Mixta** de XLM-RoBERTa y el Colectivo, donde este mismo modelo iguala a RoBERTa. Mientras que BERTweet obtiene el mejor rendimiento en la identificación automática de la Justificación, la Conclusión y sus respectivos Tipos, RoBERTa tiene una mejor performance en la identificación del Colectivo, el Pivot y los tweets Argumentativos. El rendimiento del modelo multilingual mixto, en términos generales, de todos modos, es similar al de los modelos monolingües, con la única excepción del tipo de justificación.

El rendimiento del modelo multilingual pero aplicado a lenguajes cruzados (**XL**), donde un modelo multilingual de base se ajusta con *tweets* en inglés y luego se evalúa sobre *tweets* en español, también es peor para el Tipo de Justificación, al igual que con el modelo entrenado de manera **Mixta**. Pero sobre todo, tiene un rendimiento notoriamente inferior para el Colectivo, con un F1 de .27. Esto demuestra la importancia de observar durante el entrenamiento las denominaciones de los Colectivos objetivo del odio en el lenguaje

Predicciones			
	H	V	P
H	101	0	6
V	3	0	1
P	3	0	0
	.94	.00	.00

Tab. 4.4: Matriz de confusión de las predicciones realizadas por uno de los tres *folds* del modelo XLM-RoBERTa-Mix para el Tipo de la Justificación. En la última línea también se muestra el F1 de cada categoría. Podemos observar que el mal desempeño del modelo se debe al bajo rendimiento que tiene en las clases minoritarias

original, y las dificultades de aprender esta tarea de forma cruzada entre lenguajes debido al léxico específico usado para nombrarlos, con una alta cantidad de jerga y palabras fuera de vocabulario.

Los modelos multilingües también demuestran resultados más bajos en la tarea de reconocer de manera automática al Pivot, especialmente en la configuración experimental de lenguajes cruzados (XL).

4.2.1.1 Comparación con respecto a un predictor humano

La comparación del rendimiento de los modelos entrenados con respecto a los puntajes obtenidos al calcular el acuerdo entre anotadores para cada componente argumentativo sugiere que este último podría funcionar como indicador, de manera indirecta, de cuál es el techo respecto a lo que puede aprender un modelo, sin importar cuán bueno sea. Intuitivamente, considerando que la tarea de reconocer componentes argumentativos es altamente subjetiva, el límite respecto a lo que un modelo de lenguaje debería poder aprender debería estar fijado por la capacidad de un ser humano de reconocer estos componentes. Para poder realizar una comparación en este sentido y contextualizar así, de mejor modo, los resultados obtenidos por nuestros modelos ajustados a la tarea, utilizamos los 150 ejemplos que fueron anotados por ambos anotadores y los interpretamos como si el anotador que anotó menor cantidad de ejemplos sólo en inglés intentara predecir las anotaciones del otro anotador, evaluándolo entonces como un “predictor humano”. Esta interpretación tiene fuertes bases en la realidad debido a que ambos anotadores discutieron juntos los criterios de anotación y luego fueron instruidos para intentar anotar los ejemplos intentando obtener el mayor acuerdo entre ellos.

La figura 4.1 muestra la comparación entre el F1 obtenido por el mejor modelo para cada categoría y el F1 del “predictor humano” calculado sobre los 150 *tweets* que fueron anotados en común. Podemos observar que existe una fuerte correlación entre los resultados obtenidos por uno y otro para cada componente argumentativo, con resultados similares en la mayoría de los casos, donde el predictor automático se ubica ligeramente por debajo del predictor humano. El gráfico aporta un indicio de qué componentes tienen, aún, lugar para mejorar su rendimiento. En particular, la mayor diferencia entre el predictor humano y el predictor automático se encuentra en la Propiedad (.65 vs .50 respectivamente), en la Conclusión (.76 vs .61 respectivamente) y en menor medida en el Pivot (.54 vs .45 respectivamente). El tipo de justificación muestra un resultado paradójico en el que el predictor automático obtiene un mejor resultado que el predictor humano. Esto puede explicarse por el hecho de que es el único componente argumentativo que muestra un

Video: (part 1) London #BNP a frame trailer with patriotic sound system on the road in and around our capital city ‘‘say no to **immigration**’’ #Brexit #Immigration #ImmigrationBan #London #England #BrexitBorder #Brexiteer #Brexiteers #BrexitGoodNews #BrexitChaos

Fig. 4.2: Ejemplo de una predicción sobre el Colectivo, marcada en color rojo. Aunque el modelo encuentra una mención a un Colectivo que efectivamente está siendo atacado, no se encuentra ninguna Propiedad explícita sino que el mensaje sólo incluye un mandato o una acción a realizar

acuerdo entre anotadores negativo, es decir, un desacuerdo.

Es importante aclarar que los conjuntos de ejemplos utilizados para evaluar las predicciones hechas por los modelos automáticos y por el evaluador humano son distintos y que la comparación realizada es meramente indicativa.

4.2.1.2 Análisis de predicciones

El ejemplo 4.2 muestra una predicción donde el modelo identifica correctamente un Colectivo mencionado en un tweet xenófobo, pero no se le asigna explícitamente una Propiedad, por lo que se encuentra indebidamente etiquetado. Si bien el modelo no siempre comete este tipo de errores cuando no se menciona la Propiedad, observamos que, de todos modos, es un error frecuente en proporción al total de errores cometidos. Este descubrimiento nos llevó a proponer el experimento descrito en 4.2.3 separando el problema en dos: primero identificando si es necesario etiquetar un Colectivo y una Propiedad y luego encontrándolos en el tweet.

El ejemplo 4.3 muestra una predicción incorrecta sobre la Propiedad. Aunque la trata de personas podría considerarse una consecuencia negativa, el tweet no la asocia explícitamente a un Colectivo, sino que la asociación se hace de manera implícita poniéndola como contra-punto de un ‘‘rescate’’. Estos modelos tienden a identificar frases con connotaciones negativas como Propiedades, sin considerar si están asociadas al grupo objetivo. Este problema surge independientemente de la presencia de una Propiedad real y generalmente todas las palabras o frases que podrían considerarse ‘‘negativas’’ son etiquetadas por los predictores automáticos.

Please dont call it rescue - it’s human trafficking #PortsClosed
#SendThemBack #BenefitSeekers

Fig. 4.3: Ejemplo de una predicción de una Propiedad. La Propiedad predicha está subrayada. El ejemplo original no tenía etiquetado ninguna Propiedad o Colectivo debido a que este último no se mencionaba explícitamente.

Respecto al Pivot, encontramos ejemplos donde el modelo predice más de un par de palabras como miembros de este componente argumentativo. Si bien las predicciones realizadas parecen ser correctas, el ejemplo original sólo tiene anotado un solo par, dado que así se instruyó a los anotadores. La figura 4.4 muestra un ejemplo en el cual el anotador reconoce a ‘‘Salvini’’ como el nexo entre la justificación y la conclusión, mientras que el modelo anota no solamente a ‘‘Salvini’’ en ambas premisas, sino también a un segundo posible Pivot, compuesto por ‘‘invade’’ e ‘‘invasion’’, que potencialmente también podrían funcionar como nexo entre estas. Nuevamente, estos ejemplos abonan a la necesidad de

redefinir la categoría de Pivot, permitiendo identificar múltiples nexos entre una premisa y la otra.

Salvini prosecuted for defending italian sovereignty and finally preventing hundreds of migrants to invade Italy grande Salvini, help us preserve the european culture against the invasion #StopIslamization #ComplicediSalvini #StopInvasion #RefugeesNotWelcome

Fig. 4.4: Ejemplo de predicción de un Pivot. La Justificación verdadera se muestra en azul mientras que la Conclusión verdadera se muestra en rojo. El Pivot real se muestra subrayado mientras que el Pivot predicho está en negrita. Como puede observarse, el modelo predice más de un Pivot en el mismo ejemplo.

La figura 4.5 muestra otro error frecuente también concerniente al Pivot. Aquí, el modelo, que no cuenta con información sobre otros componentes argumentativos como la Justificación y la Conclusión, etiqueta un par de palabras con una fuerte relación semántica y conceptual entre sí. Sin embargo, estas dos palabras se encuentran ambas dentro de la Justificación, y por lo tanto no están siendo utilizadas para establecer una correlación argumental entre Justificación y Conclusión, por lo que no se encuentran anotadas en el ejemplo original.

Pressure on Spain's maritime border: Boatloads of #Illegal #Migrants Storm Spanish Tourist Beaches & Scatter #StopTheInvasion #Unregistered #UnVetted

Fig. 4.5: Ejemplo de predicción sobre el Pivot. La Justificación se etiqueta en azul mientras que la Conclusión se etiqueta en rojo. El Pivot predicho se encuentra subrayado. El ejemplo no tiene un Pivot etiquetado por una persona humana. Aunque las palabras etiquetadas tienen una clara relación conceptual entre sí, ambas son parte de la Justificación por lo que fueron etiquetadas de manera incorrecta.

Respecto a las premisas, los ejemplos 4.8 y 4.9 muestran dos errores cometidos al predecir la Justificación y la Conclusión, respectivamente. En ambos casos observamos el mismo fenómeno: cada uno de los dos modelos distintos que predicen cada componente divide el *tweet* correctamente en dos componentes, pero asigna la premisa a la mitad contraria a la que la asignan los anotadores humanos. Como ambos modelos hacen lo mismo, la combinación de ambos modelos genera un ejemplo bien anotado (en el sentido de que ambas premisas son disjuntas) pero intercambiado respecto a la anotación original realizada por los anotadores humanos. Este tipo de errores es análogo a los desacuerdos encontrados durante el proceso de anotación, que se muestran en la figura 3.19, en la sección 3.3.1.1.

Finalmente, los ejemplos 4.10 y 4.11 muestran dos errores al predecir el tipo de la Justificación y la Conclusión, respectivamente. En ambos casos, el modelo predice equivocadamente el tipo “Hecho”. En el caso de la Justificación, la premisa es una frase en modo imperativo que instiga al lector a realizar una acción, por lo tanto concluimos que el error del modelo se debe a la tendencia a predecir la clase mayoritaria. Para la conclusión, consideramos que el ejemplo muestra un mayor grado de ambigüedad: mientras que el anotador consideró el uso del verbo “*want*” como un verbo que demuestra el involucramiento del usuario en el juicio que omite, podría considerarse, también, que al utilizarse este verbo en plural, el usuario demuestra la pretensión de mostrar su opinión particular como una

opinión compartida por una mayoría invisible e indeterminada a la cual apela, por lo cual consideramos que no es del todo errada la predicción de la premisa como “Hecho”.

@user utter lunacy: these migrant people have zero entitlement or use in Europe

Fig. 4.6: Ejemplo de un *tweet* no argumentativo que fue etiquetado como argumentativo por el predictor automático

citizens call @user :phone-number: tell him to bring the mandatory e-verify bill to a vote \$hireamericans not illegals :hashtag: trump :hashtag: maga :hashtag: red nation rising :hashtag: walk away :link: via @user

Fig. 4.7: Ejemplo de un *tweet* argumentativo que fue etiquetado como no argumentativo por el predictor automático

americans agree with @user on :hashtag: immigration we can not afford to give welfare to illegals while us citizens are homeless :hashtag: votedemsout :hashtag: families belong togheter march :link:

Fig. 4.8: Ejemplo de una predicción sobre la Justificación. La Justificación real se destaca en azul mientras que la predicción realizada por el modelo se muestra subrayada. Podemos observar que hay cierto nivel de acuerdo respecto a cómo dividir el *tweet* pero hay desacuerdo respecto a qué parte es la Justificación

americans agree with @user on :hashtag: immigration we can not afford to give welfare to illegals while us citizens are homeless :hashtag: votedemsout :hashtag: families belong togheter march :link:

Fig. 4.9: Ejemplo de una predicción sobre la Conclusión. La Conclusión real se destaca en rojo mientras que la predicción realizada por el modelo se muestra subrayada. Podemos observar que, igual que con la Justificación, hay cierto nivel de acuerdo respecto a cómo dividir el *tweet* pero hay desacuerdo respecto a qué parte es la Conclusión

see canada in decay by ricardo duchesne for the similar reality of canada

Fig. 4.10: Ejemplo de una justificación etiquetada por los anotadores como “Política” y predicha por el modelo automático como “Hecho”

We don't want an :hashtag: african basket of deplorables in :hashtg: europe

Fig. 4.11: Ejemplo de una conclusión etiquetada por los anotadores como “Valor” y predicha por el modelo automático como “Hecho”

4.2.2 Agregar información relevante sobre otros componentes

El Colectivo y la Propiedad son dos componentes argumentativos que se anotan en el mismo paso de anotación, señalando la intención del usuario de asignar la segunda al

primero. Por lo tanto, se encuentran profundamente conectados en su significado. A la vez, comparten una restricción que surge del proceso de anotación: no puede existir uno sin el otro. El Pivot, si bien se anota en un paso de anotación subsiguiente, también tiene una conexión con otros dos componentes argumentativos como la Justificación y la Conclusión. Y de esta conexión también surge una restricción sobre su anotación: el Pivot debe estar constituido por dos segmentos y cada uno de estos debe estar contenido en una premisa distinta: uno en la Conclusión y otro en la Justificación.

Los anotadores humanos consideran más de un componente argumentativo en simultáneo cuando realizan el proceso de anotación. Los modelos entrenados en el experimento de la sección 4.2.1, sin embargo, son entrenados para predecir y encontrar un único componente argumentativo a la vez y carecen de cualquier tipo de conocimiento o información sobre otros componentes. Buscamos evaluar, por lo tanto, en qué grado esta falta de conocimiento afecta la capacidad predictiva del modelo y cuál es el nivel de dependencia de cada componente respecto a otros componentes relevantes.

Por este motivo, el siguiente experimento consiste en entrenar los modelos y realizar las predicciones agregando información sobre otros componentes argumentativos relevantes, para poder evaluar el rendimiento sobre las predicciones realizadas asumiendo condiciones ideales. En el caso del Pivot, los anotadores humanos al momento de realizar esta anotación ya conocen cuál es la Justificación y cuál la Conclusión, con lo cual, el procedimiento de este experimento emula de mejor modo el proceso que siguieron. En el caso del Colectivo y la Propiedad, usualmente un anotador encuentra primero uno de estos componentes y luego busca si existe el complementario, con lo cual también podemos asumir que de este modo representamos esta última tarea de buscar el componente “complementario”.

Re-procesamos los datos utilizados para el experimento de la sección 4.2.1 y elaboramos versiones distintas dependiendo del componente para el cuál queremos entrenar los modelos para su identificación. Para el modelo que busca identificar al Colectivo, agregamos al texto del tweet original un sufijo “Propiedad:” que indica cuál es el texto de la Propiedad que el anotador humano encontró en ese ejemplo; para el modelo que busca identificar a la Propiedad, agregamos al texto del tweet un sufijo “Colectivo:” indicando cuál es el colectivo anotado por el humano; y para el Pivot, dos sufijos “Justificación:” y “Conclusión:” indicando cuál es la Justificación y la Conclusión respectivamente. Para el caso del Colectivo y la Propiedad, si uno de los dos componentes no está anotado, de igual modo agregamos el sufijo correspondiente, pero sin ninguna información ya que no hay una anotación para agregar, señalando, por tanto, mediante este mecanismo de sufijos, no sólo la información sobre otros componentes relevantes sino también si es necesaria la anotación o no, algo que no sucede en el caso del Pivot. El ejemplo de la figura 4.12 muestra el mismo *tweet* de la figura 3.1 en el formato de entrada propuesto para este experimento.

Debido a que el modelo devuelve una etiqueta por cada *token*, tomamos para el cálculo de las métricas reportadas únicamente las etiquetas predichas para los *tokens* originales del tweet, excluyendo las etiquetas de los *tokens* que agregamos en la entrada del modelo a modo de “sufijos”. Los resultados de este experimento pueden observarse en la tabla 4.5.

Podemos observar que el rendimiento de los modelos en la identificación del Colectivo sube 16 puntos para RoBERTa y XLM-RoBERTa-Mix, 15 puntos para BERTweet y 18 puntos para XLM-RoBERTa-XL respecto al experimento de la sección 4.2.1. Respecto a la versión XLM-RoBERTa-XL ajustada con ejemplos en inglés y evaluada sobre ejemplos en español, observamos que sigue existiendo una diferencia importante respecto al rendimiento de los demás modelos, alrededor de 30 puntos por debajo respecto a su versión

Ejemplo para predecir Propiedad:
 Illegal Immigration Hurts Black Americans and user user Dont Care About Black Americans they put Illegals First ! Tell #Democrats HELL NO!.
 Collective: Illegal Immigration

Ejemplo para predecir Colectivo:
 Illegal Immigration Hurts Black Americans and user user Dont Care About Black Americans they put Illegals First ! Tell #Democrats HELL NO!.
 Property: Hurts Black Americans

Ejemplo para predecir Pivot:
 Illegal Immigration Hurts Black Americans and user user Dont Care About Black Americans they put Illegals First ! Tell #Democrats HELL NO!.
 Justification: Illegal Immigration Hurts Black Americans. Conclusion:
 and user user Dont Care About Black Americans they put Illegals First !
 Tell #Democrats HELL NO!

Fig. 4.12: Ejemplo de la figura 3.1 con la información sobre los componentes argumentativos relevantes para predecir Propiedad, Colectivo y Pivot. Estos ejemplos muestran el formato con el cual los modelos de lenguaje realizan el aprendizaje y las predicciones sobre los componentes respectivos, en el experimento detallado en la sección 4.2.3

con datos mixtos, lo que evidencia que en este experimento persiste el problema de la dificultad de transferir conocimiento entre lenguajes para la identificación del colectivo.

Los modelos entrenados para identificar el Pivot, sin embargo, no demuestran una mejora en su rendimiento respecto a los modelos entrenados en la sección 4.2.1 e incluso muestran un leve empeoramiento, con la única excepción de BERTweet, cuyo valor de F1 sube en .03.

Dado que para el caso del Colectivo y la Propiedad la información adicional no sólo identifica el componente relevante sino que también señala su presencia e, implícitamente, también la del componente a identificar, y que esto no sucede con el Pivot, donde los componentes relevantes, la Justificación y la Conclusión, siempre se encuentran presentes en los *tweets* argumentativos y no tienen una incidencia tan directa en identificar o no la presencia del Pivot, cabe preguntarse si las mejoras en la *performance* se deben verdaderamente a la identificación de los componentes relevantes o simplemente a la identificación de si el ejemplo debe ser anotado con el componente o no. Nos proponemos dar respuesta a esta pregunta en la sección 4.2.3.

	RoBERTa			BERTweet			XLM-R-Mix			XLM-R-XL		
	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec
Colective	.75 ±.03	.78	.73	.73±.03	.70	.77	.75 ±.04	.75	.74	.45±.04	.69	.34
Property	.52±.08	.59	.48	.53±.01	.52	.56	.55 ±.01	.55	.55	.52±.02	.46	.61
Pivot	.41±.12	.51	.37	.43 ±.04	.39	.48	.35±.05	.38	.33	.32±.01	.35	.30

Tab. 4.5: Resultados de la identificación de componentes argumentativos proveyendo información relevante previamente anotada.

4.2.3 Segmentando el problema

Algunos componentes como el Colectivo, la Propiedad y el Pivot no siempre se anotan aunque el *tweet* sea argumentativo. En efecto, existe un número significativo de *tweets* argumentativos que no tienen estos componentes anotados, tal como puede observarse en la tabla 3.2.

Al analizar los resultados del experimento de la sección 4.2.1 encontramos que un error frecuente de los predictores es etiquetar el Colectivo o la Propiedad en ejemplos que no los tenían etiquetados originalmente debido a que no se encuentra presente el componente complementario (la Propiedad en el caso del Colectivo y el Colectivo en el caso de la Propiedad). Como el predictor sólo toma en cuenta uno de los componentes, no tiene forma de aprender por sí solo el criterio necesario para saber si, dado un Colectivo, debe etiquetarlo o no. En el caso del Pivot, también encontramos casos donde el modelo intenta encontrar relaciones entre conceptos donde originalmente no había ninguna anotada (lo que sucede cuando esta relación no es explícita).

Esto nos llevó a pensar que los modelos que intentan identificar estos componentes corren con una desventaja respecto al resto: deben realizar dos tareas en una. Mientras que los modelos que identifican a la Justificación y la Conclusión y sus tipos solo deben ocuparse de identificar el componente o el tipo dentro del *tweet* excluyendo la posibilidad de que este no esté contenido en el texto, los modelos que identifican al Colectivo, la Propiedad y el Pivot deben, de manera conjunta, adivinar si el *tweet* debe ser etiquetado con este componente y luego, al igual que en el caso anterior, encontrarlo.

Para este experimento decidimos hacer explícita la separación entre las dos tareas realizadas en simultáneo y separarlas en dos tareas independientes. Por un lado, entrenamos dos conjuntos de modelos para predecir una única etiqueta por *tweet*: el primer conjunto indicando si el *tweet* debe ser anotado con Colectivo y Propiedad y el segundo indicando si el modelo debe ser anotado con un Pivot. Por el otro, entrenamos otros tres conjuntos de modelos para identificar estos tres componentes (Colectivo, Propiedad y Pivot), exactamente del mismo modo que en el experimento de la sección 4.2.1 pero excluyendo tanto del entrenamiento como de la evaluación a todos los *tweets* que no tienen anotado el componente para el cual se entrena el modelo. El porcentaje de ejemplos que son utilizados en este entrenamiento respecto del entrenamiento descrito en la sección 4.2.1 puede observarse en la tabla 3.2.

Buscamos evaluar una alternativa a la detección de los componentes que se construya en dos etapas sucesivas: una primera en la que se identifican los *tweets* que merecen ser anotados y una segunda en la que aquellos *tweets* identificados en la primera etapa se anotan con un modelo que presupone su presencia. Los resultados de este experimento pueden observarse en la tabla 4.5.

Observamos que la tarea de detección de la presencia de Colectivo y Propiedad obtiene buenos resultados, con un puntaje para la métrica F1 de .80 en el mejor caso (XLM-RoBERTa-Mix) y .78 para todos los demás modelos, incluyendo la versión de lenguaje cruzado de XLM-RoBERTa. Destacamos, a su vez, que los resultados tienen baja varianza, mostrando, además, que los modelos son robustos. La tarea de detección de la presencia de Pivot, por su parte, obtiene un valor de F1 máximo de .68 para XLM-RoBERTa-Mix, y luego valores superiores de .66 para BERTweet, .62 para XLM-RoBERTa-XL y .60 para RoBERTa.

Podemos observar que la identificación del Colectivo en ejemplos donde se garantiza su presencia aumenta su rendimiento significativamente respecto al experimento de la

sección 4.2.1 a pesar de sólo utilizar un 58 % de los ejemplos. Comparando el mejor modelo, RoBERTa en ambos casos, observamos un incremento de .15 puntos, .14 para BERTweet y .13 para XLM-RoBERTa-Mix. En el caso de la versión de lenguajes cruzados de XLM-RoBERTa, la mejora en el rendimiento es de .22, la mayor entre los cuatro modelos. La Propiedad, por su parte, mejora su rendimiento en .03 teniendo en cuenta el mejor modelo, XLM-RoBERTa-Mix en ambos casos, y obtiene mejoras de .06 para RoBERTa, .04 para BERTweet y .10 para XLM-RoBERTa-XL.

Para estos dos componentes, los resultados son, en mayor o menor medida, análogos a aquellos obtenidos en el experimento de la sección 4.2.2, donde las mejoras respecto al experimento de la sección 4.2.1 para el Colectivo son prácticamente idénticas y para la Propiedad son también similares, siendo ligeramente superior el resultado obtenido en la sección 4.2.2. Este resultado sugiere que la razón principal de la mejora en el rendimiento en el experimento de la sección 4.2.2 se debe al conocimiento de que había que anotar el *tweet*, mucho más que a cuál era el contenido del componente “complementario” en sí.

Si comparamos en cambio los resultados obtenidos para el Pivot, podemos observar que la tarea de identificación automática de este componente Argumentativo mejora significativamente, a diferencia de lo previamente observado en la sección 4.2.2 donde en este caso, la información agregada no servía para identificar si el ejemplo debía ser anotado o no. Observamos una mejora de .12 para el mejor modelo, RoBERTa, .11 para BERTweet, .10 para XLM-RoBERTa-Mix y .13 para XLM-RoBERTa-XL. Estos resultados, en combinación con los obtenidos en la sección 4.2.2, sugieren que los modelos interpretan la tarea como una búsqueda de segmentos de palabras con cargas semánticas similares pero no logran comprender la razón subyacente para anotar o no el Pivot: que cada uno de estos segmentos esté en una premisa distinta. Esta dificultad persiste incluso cuando las premisas se agregan como información complementaria a la entrada del modelo.

Los resultados muestran que el componente argumentativo que se ve afectado en mayor modo por el hecho de sólo lidiar con ejemplos donde el componente se encuentra efectivamente anotado y la tarea sólo consiste en encontrar dónde es el Colectivo. Los modelos que intentan identificar el Colectivo aprenden a identificarlos sin considerar a la Propiedad y si esta se encuentra presente. La Propiedad, por su parte, si bien mejora al acotar el experimento a los ejemplos donde ésta está anotada, no muestra una mejora tan significativa, lo que indicaría que posiblemente sea más frecuente encontrar ejemplos que mencionen al Colectivo pero no a la Propiedad que se busca asociar a éste que el caso contrario (es decir, que cuando la Propiedad es explícita, en general el Colectivo también lo es). De este modo, una posibilidad de encadenar estos experimentos para obtener un mejor resultado podría consistir en utilizar el predictor de la sección 4.2.1 para encontrar la Propiedad y definir si es necesaria su anotación y, si lo es, utilizar luego el predictor que encuentra el Colectivo asumiendo que éste se encuentra presente.

4.2.3.1 Análisis de predicciones

Al delimitar el problema a encontrar un Colectivo en un *tweet* en el que sabemos que se le está asignando una propiedad y por lo tanto, siempre es correcto anotarlo, la mayoría de los errores producidos por los clasificadores automáticos pasan a ser discrepancias en la cantidad de palabras utilizadas para referirse al colectivo (como en el ejemplo 4.13) o cuando el *tweet* menciona múltiples colectivos además del objetivo del mensaje de odio (como en el ejemplo 4.14). Creemos que el primer caso revela una oportunidad de mejora en el proceso de anotación, donde a veces, un colectivo podría haber sido etiquetado usando

	RoBERTa			BERTweet			XLM-R-Mix			XLM-R-XL		
	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec
Predecir la presencia o no de los componentes												
Col-Prop	.78±.03	.74	.84	.78±.02	.69	.90	.80±.00	.72	.89	.78±.03	.70	.88
Pivot	.60±.01	.63	.58	.66±.08	.60	.73	.68±.06	.63	.73	.62±.03	.60	.66
Identificación del componente en los ejemplos donde está anotado												
Colectivo	.74±.02	.74	.76	.72±.02	.71	.74	.72±.06	.70	.74	.49±.03	.63	.40
Propiedad	.52±.02	.58	.48	.51±.01	.54	.50	.53±.02	.55	.50	.52±.02	.46	.59
Pivot	.57±.08	.60	.54	.51±.09	.55	.49	.49±.09	.50	.49	.46±.01	.57	.38

Tab. 4.6: Resultados de predecir si un componente debe ser anotado en un *tweet* o no (en la parte superior) y de reconocer automáticamente ese componente utilizando sólo ejemplos filtrados donde los componentes están anotados siempre (en la parte inferior).

Chain migration imported 120K foreign nationals from **terrorist-funding countries** since 2005 - breitbart @user @user #EndChainMigration #EndDACA #NoAmesty #EndBirth-rightCitizenshipForIllegalAliens #BuildTheWall #KeepAmericaSafe

Fig. 4.13: Ejemplo de predicción del Colectivo cuando se garantiza su presencia. En rojo se indica el Colectivo identificado por el anotador humano y en azul la propiedad. Subrayado y en negrita se indica la predicción sobre el Colectivo realizada por el modelo. Podemos observar que ésta es parcialmente correcta pero incluye la palabra “chain” asociada con “migration”. Estas diferencias surgen frecuentemente cuando aparecen frases construidas comunmente, como “Chain Migration” o “illegal immigrants”

una palabra y otras veces usando varias.

At this time, **w-organized crime/returning jihadists** it’s a matter of national security. #Italy #Salvini must ignore international social engineers/cultural **marxists** #V4 Itali Kurz others must challenge empty threats from un-eu **migration** pimps. What can they really do about it?

Fig. 4.14: Ejemplo de predicción del Colectivo cuando su presencia está garantizada. En este caso, el modelo identifica correctamente el Colectivo “migration” pero predice también como parte de este la palabra “jihadists” que forma parte de la propiedad que se le endilga al verdadero Colectivo (que son parte del crimen organizado o jihadistas retornando). También predice la palabra “marxists” como parte del Colectivo cuando en verdad este grupo no es el objetivo principal del mensaje de odio sino receptores secundarios acusados de defender o asociarse con el verdadero Colectivo.

Otro error al que los modelos automáticos son propensos es etiquetar porciones de texto más grandes o más pequeñas que las que deberían. El ejemplo 4.15 muestra una predicción realizada por un modelo entrenado y evaluado con ejemplos donde siempre hay una Propiedad explícita. El modelo identificó correctamente “invadir ilegalmente los EE.UU.” como parte de la Propiedad, pero omitió el resto.

4.2.4 Predicción Conjunta de Componentes

El objetivo de este experimento es medir el rendimiento de los modelos al ser entrenados para predecir componentes que se anotan en el mismo paso de anotación en simultáneo: por un lado, el Colectivo y la Propiedad y por el otro, la Justificación y la Conclusión. Queremos evaluar hasta qué punto realizar la predicción conjunta ayuda o no a mejorar un sistema de anotación automático de componentes argumentativos. Existen dos potenciales

@user the disgrace is the **illegal** parent who **brought their kids on their crime spree to illegally invade the U.S.** so taxpayers pay for their kids education wic and medicaid. We don't owe **illegals** our tax dollars #SendThemBack #WalkAway #Trump #MAGA #RedNationRising

Fig. 4.15: Ejemplo de predicción de Propiedad en ejemplos donde se garantiza su presencia. La propiedad real se destaca en color azul mientras que el colectivo se destaca en rojo. La propiedad predicha por el modelo se destaca en negrita y subrayado. En este caso, el modelo predice parte de la propiedad pero deja una parte sin marcar

razones para realizar la predicción conjunta de componentes. En primer lugar, porque los modelos se entrenan con información de todos los componentes relacionados entre sí, contando en los hechos, con mayor información. Este escenario, sin embargo, fue previamente contemplado al plantear nuestro experimento de la sección 4.2.2. La segunda razón es que entrenar los modelos de esta manera considera una particularidad del proceso de anotación: que en casi la totalidad de los casos, los componentes que se anotan en el mismo paso de anotación son disjuntos. Al anotar todos los componentes disjuntos utilizando un solo modelo en una sola pasada hacia adelante se garantiza la disjunción, ya que el modelo predice una sola etiqueta por palabra, asignando, por lo tanto, cada palabra a un único componente.

En el caso del Colectivo y la Propiedad, la disjunción es una restricción discutida e incorporada al proceso de anotación que fue respetada por todos los anotadores. En el caso de la Justificación y la Conclusión, se discutió un caso particular en el cual la Justificación podía estar contenida dentro de la Conclusión. Es el caso de los Argumentos de Autoridad, donde se justifica un hecho o declaración porque fue dicha por una supuesta autoridad en el tema. Suelen ser de la forma “X es verdadero porque A dijo X” pero muchas veces se omite la primera parte y el argumento se presenta en una forma comprimida como “A dijo X”. Para este caso particular, se discutió anotar “X” como la conclusión y “A dijo X” (es decir, todo el mensaje) como la justificación, dando a entender que se concluye X justificándose en que A dijo X. Este es el único caso en que se permite que la Justificación y la Conclusión se superpongan. El ejemplo 4.16 muestra un caso como el descrito.

[South Koreans stand against islamic immigration](#) :link: via @user

Fig. 4.16: Ejemplo anotado con Justificación (en azul) y Conclusión (subrayada) superpuestas. Mientras que la Conclusión es de tipo “Política” y pide tomar postura contra la inmigración, la Justificación es de tipo “Hecho” y plantea que esa postura es la que tienen los Surcoreanos.

Para realizar este experimento, por lo tanto, utilizamos un clasificador de *tokens* con tres posibles clases por *token* como salida: una indicando que el token no pertenece a ningún componente argumentativo y las otras dos indicando que pertenecen a alguno de los dos componentes que buscamos predecir de manera conjunta: en un caso el Colectivo y la Propiedad y en el otro, la Justificación y la Conclusión.

La tabla 4.7 muestra los resultados de este experimento. La tabla muestra el F1, *Precision* y *Recall* para las dos clases positivas (Colectivo y Propiedad, en un caso, y Justificación y Conclusión en el otro) y el promedio Macro entre las tres categorías predichas por el modelo para cada uno de los dos modelos entrenados. Podemos observar que el rendimiento respecto a la identificación del Colectivo y de la Conclusión es similar al

@user Immigration in a picture :link: **Some basic truths: Access to White people is not a human right.**

Fig. 4.17: Ejemplo de una predicción conjunta sobre la Justificación y la Conclusión. La Conclusión predicha se muestra en azul mientras que la real está en negrita. La Justificación predicha está en rojo mientras que la real está subrayada. Como puede observarse, el modelo divide correctamente el texto en dos proposiciones pero falla en reconocer correctamente cuál es la Justificación y cuál es la Conclusión.

del experimento 4.2.1. Para la Justificación y la Propiedad, observamos un rendimiento menor. Sin embargo, la disminución no parece ser significativa, al menos en el caso de la Justificación (.03 puntos de diferencia). La predicción conjunta nos permite, a la vez, tener una evaluación más general de las posibilidades de identificación de componentes interrelacionados que conforman un único paso de anotación y que suelen realizarse necesariamente de conjunto a nivel cognitivo por los anotadores humanos. En este aspecto, la tarea de Identificación de Premisas (Justificación y Conclusión) obtiene un F1 de .68 en su mejor modelo, BERTweet, y luego obtiene valores entre .65 y .66 para el resto de los modelos, lo cual nos permite concluir que esta tarea puede ser automatizada con un rendimiento aceptable. La tarea de identificar los componentes argumentativos que son específicos del dominio del discurso de odio, Colectivo y Propiedad, obtiene un F1 de conjunto de .51, lo cual nos permite concluir que esta tarea puede ser automatizada, hasta cierto punto, con un rendimiento aceptable, dependiendo del contexto, pero con lugar aún para mejorar.

Otra ventaja de realizar la predicción de manera conjunta es que consume una cantidad considerablemente menor de recursos y tiempo, ya que con solo agregar una etiqueta se disminuye a la mitad el tiempo y los recursos necesarios para el entrenamiento y la inferencia.

En vista de los resultados obtenidos, concluimos que la predicción conjunta de componentes muestra amplias ventajas respecto a la predicción individual, particularmente en el caso de la Justificación y la Conclusión.

4.2.4.1 Análisis de predicciones

La predicción conjunta del Colectivo y la Propiedad y de la Justificación y Conclusión mejora el rendimiento de los modelos, especialmente en este segundo caso. Sin embargo, observamos que continúa ocurriendo un error que podemos relacionar con un fenómeno observado al analizar el acuerdo entre anotadores, tal como muestra la figura 3.19: que la división del tweet en dos proposiciones se hace con mayor facilidad, pero luego existen mayores dificultades para decidir cuál de éstas es la Justificación y cuál la Conclusión. El ejemplo de la figura 4.17 muestra una predicción conjunta de la Justificación y la Conclusión realizada por el modelo de mejor rendimiento. Podemos observar que la división del tweet en dos partes se realiza exactamente igual que fue realizada por el anotador humano. Sin embargo, el modelo asigna la categoría de Justificación y de Conclusión de manera inversa. Los ejemplos de las figuras 4.18 y 4.19 muestran las predicciones hechas sobre el mismo ejemplo por los modelos descritos en la sección 4.2.1. Como puede observarse, estos modelos realizan, por separado, la misma predicción que el modelo conjunto, mostrando la persistencia de la misma dificultad entre ambos casos.

@user Immigration in a picture :link: **Some basic truths: Access to White people is not a human right.**

Fig. 4.18: Ejemplo de predicción únicamente sobre la Conclusión realizado con el modelo de la sección 4.2.1. La Conclusión predicha se muestra en azul mientras que la real se muestra en negrita.

@user Immigration in a picture :link: Some basic truths: Access to White people is not a human right.

Fig. 4.19: Ejemplo de predicción únicamente sobre la Justificación realizado con el modelo de la sección 4.2.1. La Justificación predicha se muestra en rojo mientras que la real se muestra subrayada.

	RoBERTa			BERTweet			XLM-R-Mix			XLM-R-XL		
	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec	F1	Pr	Rec
Collective	.57±.04	.58	.58	.59±.03	.59	.59	.57±.01	.59	.54	.36±.02	.34	.49
Property	.41±.03	.41	.48	.43±.02	.43	.46	.45±.01	.45	.54	.40±.04	.42	.31
Macro Avg	.49±.04	.50	.53	.51±.02	.51	.53	.51±.01	.49	.52	.38±.03	.38	.40
Justification	.71±.06	.67	.74	.74±.01	.70	.78	.73±.02	.72	.75	.74±.00	.70	.78
Conclusion	.59±.05	.62	.57	.61±.03	.64	.58	.58±.03	.58	.58	.57±.03	.64	.51
Macro Avg	.65±.05	.65	.66	.68±.02	.67	.68	.66±.02	.65	.67	.66±.02	.67	.65

Tab. 4.7: Resultado de predecir de manera conjunta componentes argumentativos relacionados anotados en el mismo paso de anotación. La tabla muestra el F1, *Precision* y *Recall* para las dos clases positivas (Colectivo y Propiedad, en un caso, y Justificación y Conclusión en el otro) y el promedio Macro entre las tres categorías predichas por el modelo para cada uno de los dos modelos entrenados.

4.2.5 Predicción sobre el Tipo de las Premisas

Si bien cada una de las dos premisas en nuestro modelo (Justificación y Conclusión) tiene asociada una distribución sobre sus posibles Tipos, la cual puede verse en la figura 3.21, la clasificación de una premisa particular en “Hecho”, “Valor” o “Política” es independiente del rol argumentativo que cumple esa premisa. En el experimento de la sección 4.2.1, entrenamos un modelo individual para predecir el Tipo únicamente con las mismas premisas que buscamos predecir. Es decir, entrenamos un modelo para predecir el Tipo de la Justificación únicamente con las Justificaciones etiquetadas y entrenamos un modelo para predecir el Tipo de las Conclusiones únicamente con las Conclusiones etiquetadas. Si bien cada modelo se entrena con un conjunto distinto de datos, entendemos que realizan, en definitiva, la misma tarea. Por lo tanto, entrenar un modelo con ambos tipos de premisas puede resultar en un mejor rendimiento en identificar el Tipo de ambas premisas.

Entrenar un modelo utilizando toda la información disponible sobre el Tipo de una premisa tiene la ventaja directa de duplicar el tamaño del conjunto de entrenamiento, a la vez que permite superar la situación de desbalanceo de los datos que presenta, sobre todo, la Justificación, donde el Tipo Hecho es mayoritario (aunque no exclusivamente, ya que la clase Valor también es minoritaria en la Conclusión). La tabla 4.8 muestra los resultados obtenidos por este único modelo entrenado con todas las premisas y sus tipos. La tabla muestra el resultado obtenido en tres conjuntos distintos de evaluación y los compara con los resultados de los dos clasificadores entrenados de acuerdo a lo desarrollado en la sección 4.2.1. Los conjuntos referenciados con la leyenda “Tipo Just” y “Tipo Concl” son

los mismos conjuntos de evaluación utilizados para medir el rendimiento de los modelos entrenados sólo con un componente argumentativo, compuestos únicamente por el texto de la Justificación y únicamente por el texto de la Conclusión, respectivamente. Agregamos un tercer conjunto de evaluación que surge de combinar los dos anteriores y que podemos utilizar para medir el rendimiento del predictor del Tipo de sentencia, independientemente del componente argumentativo. Copiamos en las últimas dos filas de la tabla los valores obtenidos en el experimento de la sección 4.2.1.

Podemos observar un incremento significativo del rendimiento en la tarea de predecir el tipo de ambas premisas. En el caso de la Justificación, que tenía un valor de F1 mucho más bajo, observamos una mejora de .13 puntos comparando a los modelos con mejor rendimiento (BERTweet con .42 entrenado con una sola premisa vs XLM-RoBERTa-Mix con .55 entrenado con ambas). Para los modelos multilinguales, la mejora es aún mayor. XLM-RoBERTa-Mix aumenta en .20 puntos su valor de F1, comparando su puntaje al ajustarse con solo una premisa (.35 de F1) con el obtenido al ajustarse con ambas (.55 de F1), mientras que su versión de lenguajes cruzados, XLM-RoBERTa-XL, muestra una mejora de .19 puntos, de .33 a .52. Los modelos monolinguales también muestran una mejora, con un incremento de .11 puntos para BERTweet y de .08 para RoBERTa.

El rendimiento para la identificación del Tipo de la Conclusión, si bien en menor grado que para el Tipo de Justificación, también mejora, con un incremento de .05 comparando el mejor modelo (BERTweet en ambos casos, de .65 a .70). Para el resto de los modelos, RoBERTa muestra una mejora de .05 y XLM-RoBERTa-Mix una mejora de .06.

Curiosamente, la versión de lenguajes cruzados, XLM-RoBERTa-XL muestra un rendimiento peor, pasando de .63 en el caso del modelo de la sección 4.2.1 a .57 en este caso.

Respecto a la evaluación del modelo sobre el conjunto que combina los textos de las Justificaciones y las Conclusiones de manera indistinta, observamos que el mejor modelo, BERTweet, alcanza un puntaje de .69, un valor superior al obtenido por cualquiera de los dos clasificadores del experimento de la sección 4.2.1 individualmente.

Podemos observar también que el modelo entrenado con ambas premisas obtiene sobre todo, mejores resultados en las clases minoritarias, especialmente en la clase “Política” donde, al ser evaluado en la Justificación, pasa de .00 a .41 con RoBERTa; de .17 a .47 con BERTweet, de .08 a .34 en XLM-RoBERTa-Mix y de .05 a .21 en XLM-RoBERTa-XL. Para la evaluación realizada sobre la Conclusión, el valor de F1 para esta clase se mantiene igual, con una variación que va de .00 a .02, con la excepción de XLM-RoBERTa-XL en donde el rendimiento empeora, pasando de .65 a .60. Para la clase “Valor”, en la Justificación, el F1 empeora para RoBERTa, pasando de .28 a .13 pero mejora para BERTweet, de .13 a .19, para XLM-RoBERTa-Mix, de .00 a .37 y para XLM-RoBERTa-XL, de .00 a .41. En la Conclusión, para esta clase, observamos una mejora en tres modelos: para RoBERTa de .07 a .22, para BERTweet de .20 a .37 y para XLM-RoBERTa-Mix de .28 a .37. En XLM-RoBERTa-XL, sin embargo, al igual que con la clase “Política”, el F1 disminuye, de .45 a .34.

En la clase mayoritaria, “Hecho”, se mantiene el buen rendimiento con valores similares de F1 a los obtenidos en el experimento de la sección 4.2.1 e incluso un poco superiores en el caso de la Conclusión. RoBERTa pasa de .95 a .92 en la Justificación y de .81 a .82 en la Conclusión; BERTweet pasa de .95 a .93 en la Justificación y de .84 a .85 en la Conclusión. XLM-RoBERTa-Mix pasa de .97 a .94 en la Justificación y de .70 a .78 en la Conclusión; finalmente, XLM-RoBERTa-XL pasa de .95 a .93 en la Justificación y mantiene un valor de .78 en la Conclusión.

Estos resultados validan nuestra hipótesis de que entrenar un modelo con ámbos tipos de proposiciones, con distribuciones distintas de sus clases, ayuda a superar el desbalanceo y mejorar sus resultados en las clases minoritarias.

	RoBERTa				BERTweet				XLM-R-Mix				XLM-R-XL			
	Macro	H	V	P	Macro	H	V	P	Macro	H	V	P	Macro	H	V	P
Modelos entrenados con textos de Justificaciones y Conclusiones																
Type of Just	.49±.07	.92	.13	.41	.53±.08	.93	.19	.47	.55±.01	.94	.37	.34	.52±.17	.93	.41	.21
Type of Conc	.63±.14	.82	.22	.85	.70±.14	.85	.37	.87	.67±.16	.78	.37	.86	.57±.04	.78	.34	.60
Type of both	.66±.05	.90	.28	.79	.69±.12	.91	.34	.82	.67±.04	.88	.35	.79	.60±.03	.89	.39	.53
Modelos Entrenados con sólo una proposición (resultados de la tabla 4.3)																
Type of Just	.41±.09	.95	.28	.00	.42±.09	.95	.13	.17	.35±.05	.97	.00	.08	.33±.03	.95	.0	.05
Type of Conc	.58±.05	.81	.07	.85	.65±.11	.84	.20	.89	.61±.06	.70	.28	.85	.63±.02	.78	.45	.65

Tab. 4.8: Resultados para la identificación del Tipo de proposición realizada con un modelo entrenado con textos de Justificaciones y Conclusiones en conjunto. Se muestran los resultados evaluados en tres conjuntos de evaluación distintos: uno compuesto de sólo Justificaciones, otro sólo de Conclusiones y otro compuesto por los dos anteriores combinados. Los resultados se comparan contra los valores de F1 totales y por categoría obtenidos en el experimento de la sección 4.2.1, que fueron evaluados utilizando los dos primeros conjuntos de evaluación respectivamente.

4.2.5.1 Analisis de predicciones

El ejemplo 4.20 muestra una justificación que fue etiquetada como “Valor” por un anotador humano y luego fue predicha como “Política” por un modelo entrenado tal como se describe en la sección 4.2.1, el cual, como pudimos observar, tiene dificultades para etiquetar las clases minoritarias y particularmente la clase “Valor”. El mismo ejemplo fue luego etiquetado por el modelo desarrollado en esta sección, entrenado con ejemplos de los dos tipos de premisas, el cual tiene un rendimiento mejor para las clases minoritarias y particularmente para la clase “Valor” y este logró predecir correctamente el tipo de la premisa. En el caso de este ejemplo en particular, la pertenencia a la clase “Valor” es clara debido a la presencia de verbos de pensamiento tal como “*I do not want*” que señalan el involucramiento del hablante en la afirmación que realiza, de tal modo que proponer una contranarrativa, por ejemplo, cuyo eje sea negar que el usuario “no quiera” a los inmigrantes no constituiría una respuesta coherente o eficiente.

I do not want those vile thugs in our country

Fig. 4.20: La justificación fue etiquetada como “Valor” por un anotador humano y fue predicha como “Política” por un modelo entrenado siguiendo las configuraciones descritas en 4.2.1. Luego correctamente identificada como “**V**alor” por un modelo entrenado como se describe en 4.2.5.

4.2.6 Impacto de la cantidad de datos utilizados para el entrenamiento

Los modelos de lenguaje pre-entrenados han permitido, a través de la transferencia de aprendizaje con tareas pretexto, lograr un buen desempeño al ser ajustados a tareas específicas con pocos datos de entrenamiento. La tarea pretexto le permite al modelo aprender características generales del lenguaje que ayudan luego a generar una representación mejor de la tarea con un ajuste menor en sus pesos. En paralelo, el rendimiento

	Monolingual	Multilingual Mixto	Multilingual Cruzado
25 %	193	223	213
50 %	385	445	425
75 %	578	668	638

Tab. 4.9: Cantidad de datos de entrenamiento utilizados de forma incremental para entrenar los modelos monolinguales, el modelo multilingual mixto y el modelo multilingual cruzado

	25 %	50 %	75 %	100 %
Colectivo (RoBERTa)	.56	.58	.59	.59
Propiedad (XLM-RoBERTa-Mix)	.32	.39	.41	.50
Pivot (RoBERTa)	.26	.33	.36	.45
Justificación (BERTweet)	.75	.78	.79	.77
Conclusion (BERTweet)	.57	.60	.65	.61
Argumentative (RoBERTa)	.87	.87	.87	.89
Type of Just (BERTweet)	.32	.35	.40	.42
Type of Conc (BERTweet)	.60	.65	.65	.65

Tab. 4.10: Valores de F1 obtenidos para los modelos entrenados con el 25 %, 50 %, 75 % y 100 % de los datos utilizando el modelo con mejor rendimiento en el experimento de la sección 4.2.1. El modelo utilizado se indica entre paréntesis al lado del componente correspondiente. El gráfico de la figura 4.21 modela estos mismos datos en una serie temporal

marginal de expandir el conjunto de datos se vuelve cada vez menor en la medida que éstos aumentan. Nos proponemos, por lo tanto, estudiar este rendimiento marginal de nuestros datos, puntualmente para el experimento descrito en la sección 4.2.1.

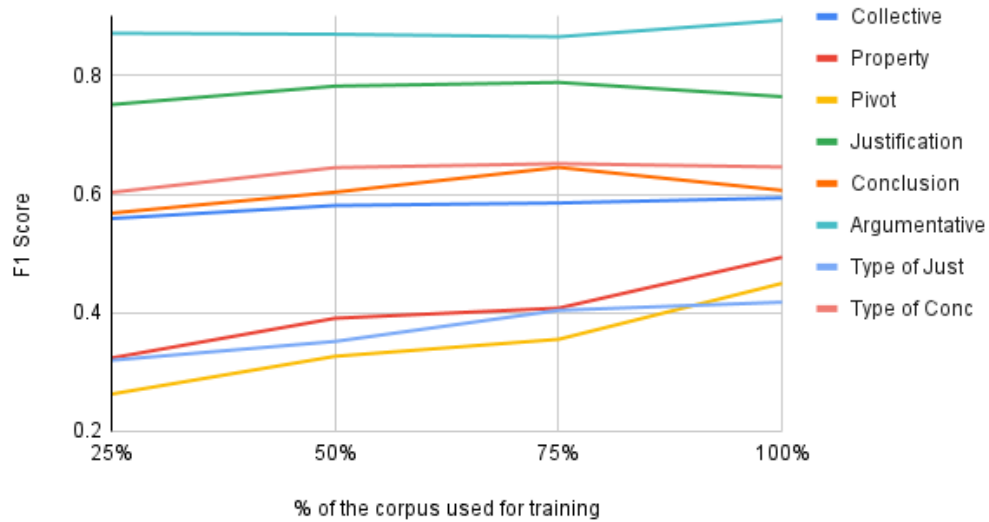
Con este fin, creamos tres porciones más pequeñas del *dataset* utilizado para el entrenamiento, seleccionadas aleatoriamente y respetando una proporción de 25 %, 50 % y 75 % respecto del original. Entrenamos, para cada componente argumentativo, tres nuevos modelos con estos datos. La tabla 4.9 muestra la cantidad de ejemplos utilizados en cada etapa incremental sucesiva para cada una de las configuraciones de entrenamiento. Estos modelos fueron evaluados utilizando el mismo conjunto de datos que en el experimento 4.2.1. Nuestro objetivo es analizar la variación de su desempeño en la medida que aumentamos el tamaño del *dataset* de entrenamiento y la ganancia relativa obtenida en el desempeño, a partir de comparar la variación de los puntajes de F1 obtenidos por cada modelo.

La tabla 4.10 muestra los resultados obtenidos para el entrenamiento progresivo, incorporando particiones cada vez mayores del *dataset* en los modelos que obtuvieron mejores resultados utilizando la totalidad de los datos. Informamos el modelo utilizado entre paréntesis al lado del componente argumentativo correspondiente. La figura 4.21 muestra estos mismos valores como una función temporal donde cada momento representa un porcentaje cada vez mayor de datos utilizado para el entrenamiento. En el primer gráfico podemos observar para cada componente, la evolución del valor del F1. En el segundo, podemos observar, también para cada componente, el rendimiento marginal obtenido (qué sucede al pasar de un modelo que usa el 25 % de los datos a otro que utiliza el 50 % y así sucesivamente). La conclusión principal que muestran es que los componentes con un mayor incremento marginal (y que potencialmente podrían mejorar aún más su rendimiento si aumentamos el tamaño del *dataset* etiquetando más datos) son los componentes que

muestran un peor rendimiento en general: el Pivot, la Propiedad y el Tipo de Justificación. Para estos tres componentes se observa una tendencia creciente de sus resultados a medida que aumenta la cantidad de datos utilizada para entrenar: la Propiedad aumenta de .32 a .50 y da su mayor salto al pasar del modelo entrenado con 75 % de los datos al modelo que utiliza el 100 %; el Pivot pasa de .26 a .45 y también da su mayor salto en el ultimo tramo; el Tipo de Justificación pasa de .32 a .42, pero en este caso, el mayor aumento en el rendimiento se da al pasar del 50 % de los datos al 75 %. Estos resultados sugieren una posibilidad alta de mejorar el rendimiento obtenido en el experimento de la sección 4.2.1 para estos componentes si se etiquetan y se utilizan más datos en el entrenamiento.

Para los demás componentes, el resultado obtenido por los modelos entrenados con menor cantidad de datos tiene escasa variación. Incluso con el 25 % de los datos, la identificación de *tweets* argumentativos y de la Justificación obtienen un valor de F1 similar al obtenido con el 100 % de los datos. El Tipo de Conclusión evidencia una mejora al pasar del 25 % de los datos al 50 % pero luego mantiene un valor de F1 idéntico al seguir incrementando el tamaño de los datos utilizado para entrenamiento. La Conclusión también evidencia una mejora al pasar de 25 % a 50 % y luego nuevamente al pasar de 50 % a 75 %, pero luego el rendimiento cae cuando usa el 100 % de los datos y apenas supera en un punto al rendimiento obtenido con el 50 %. El Colectivo, finalmente, muestra una pequeña mejora a medida que se aumenta el tamaño del dataset, pero es escasa: sólo .03 entre utilizar el 25 % y el 100 % y solo .01 entre el 50 % y el 100 %. Para estos casos, los resultados sugieren que aumentar el tamaño del dataset anotado no debería redundar en un mejor rendimiento de los modelos y que en todo caso, las limitaciones a la hora de mejorar la detección de estos componentes pasan por mejorar los criterios de anotación con el fin de superar las subjetividades propias de la tarea.

Evolución del F1 al aumentar los datos de entrenamiento



Porcentaje de mejora al aumentar los datos de entrenamiento

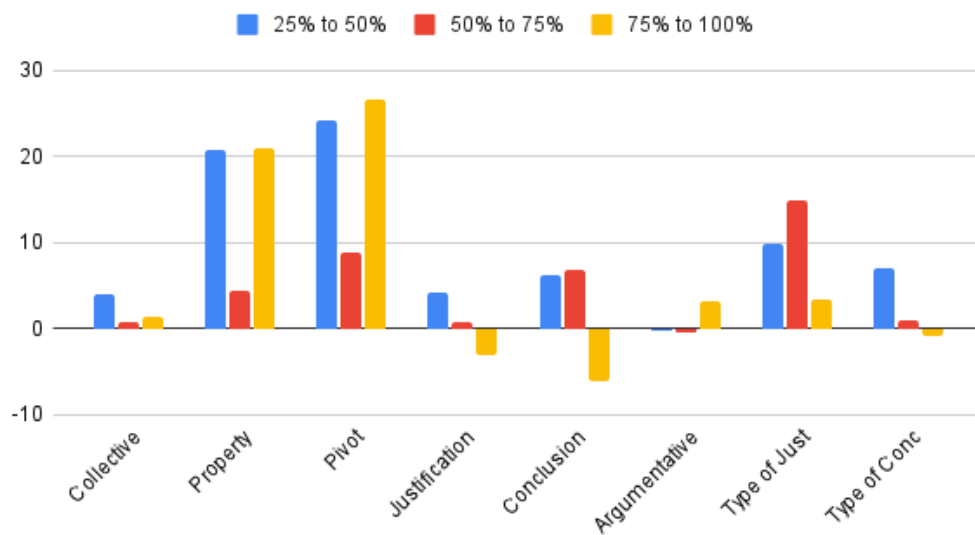


Fig. 4.21: Evolución del valor de F1 por componente argumentativo cuando se incrementa el tamaño del conjunto de datos utilizado para el entrenamiento. La primera figura muestra la evolución de los valores absolutos de F1 mientras que la segunda muestra el porcentaje de variación marginal entre cada etapa de aumento de los datos.

5 Generación Automática de Contranarrativas para Combatir Mensajes de Odio

En esta sección describimos el proceso de entrenamiento de distintos modelos de lenguaje y posterior generación automática de contranarrativas contra mensajes de odio utilizando los mismos. Definimos un protocolo de evaluación manual de las contranarrativas generadas según cuatro categorías y un puntaje agregado que puede utilizarse para comparación. Realizamos, con este protocolo, una evaluación manual con tres anotadores, buscando medir el rendimiento de distintos tamaños de modelos, estrategias de generación, uso de la información argumentativa y tipos distintos de contranarrativas. Finalmente, presentamos un análisis de los resultados destacando tanto las adecuaciones como las problemáticas del trabajo realizado.

5.1 Modelo base

Para elegir el modelo base utilizado para la generación automática de contranarrativas, consideramos los siguientes aspectos:

- Que el modelo sea abierto y por lo tanto, se pueda asegurar la reproducibilidad de los resultados.
- Que se ofrezcan versiones del mismo modelo con distintos tamaños de arquitectura base y distinta cantidad de parámetros.
- Que una de esas versiones sea de tamaño pequeño para poder ser deployada en contextos de bajo presupuesto o bajos recursos computacionales.

Para elegir el modelo a utilizar, evaluamos manualmente 10 ejemplos generados por cuatro modelos que cumplen con estas características, pasándoles como entrada una instrucción de la tarea y dos ejemplos anotados, tal como se describe en la figura 5.2:

Bloom [et. al., 2023] es un modelo de lenguaje de gran tamaño (LLM, por sus siglas en inglés) desarrollado en 2022 por BigScience en un esfuerzo colaborativo internacional que involucró a más de mil investigadores. Su versión más grande cuenta con 176 mil millones de parámetros, sin embargo, también cuenta con versiones de 560 millones, 1, 3 y 7 mil millones de parámetros que pueden ser deployados en entornos con menos requerimientos de *hardware*. Bloom se distingue por su enfoque en la transparencia, responsabilidad y replicabilidad de los resultados dado que proporciona un acceso abierto al modelo y a los datos de entrenamiento utilizados. El modelo fue entrenado en un corpus multilingüe de 46 idiomas y 13 lenguajes de programación. Además, su desarrollo incorporó prácticas responsables de IA, como la evaluación de sesgos y el impacto ético.

GPT-J [Wang and Komatsuzaki, 2021] GPT-J es un modelo de lenguaje de gran tamaño basado en la arquitectura de *transformers*, desarrollado por EleutherAI en 2021. Tiene 6 mil millones de parámetros y fue diseñado para ser una alternativa de código abierto a modelos propietarios como GPT-3, ofreciendo a la comunidad investigadora acceso a un modelo potente sin restricciones comerciales. GPT-J se entrenó en el corpus de datos *Pile*, un conjunto masivo y diverso de textos que incluye fuentes desde artículos académicos hasta códigos de programación, lo que le permite realizar tareas complejas de procesamiento de lenguaje natural con un alto nivel de competencia. Su diseño busca equilibrar la capacidad y la eficiencia computacional, y ha sido empleado en una amplia variedad de aplicaciones, desde generación de texto hasta traducción automática y análisis semántico

GPT2 [Radford et al., 2018] GPT-2 es un modelo de lenguaje autoregresivo desarrollado por OpenAI en 2019. Tiene 1.5 mil millones de parámetros y fue entrenado con un corpus masivo y diverso compuesto de contenido descargado de internet. A pesar de su tamaño relativamente menor en comparación con modelos más recientes, GPT-2 destacó por su habilidad para realizar múltiples tareas de procesamiento de lenguaje natural sin ajustes específicos, como la generación de texto, traducción y respuesta a preguntas.

Flan-T5 [Chung et al., 2022] Flan-T5 es una versión ajustada del modelo T5 (Text-To-Text Transfer Transformer), desarrollada por investigadores de Google en 2022, con el objetivo de mejorar el rendimiento del modelo original mediante el ajuste con instrucciones (instruction fine-tuning). Mientras que T5 destacaba por su enfoque basado en tratar todos los problemas de procesamiento de texto como si fueran un problema de “texto a texto”, Flan-T5 forma parte de una iniciativa cuyo enfoque se basa en convertir una amplia variedad de tareas de procesamiento de lenguaje natural en una tarea más general que consiste en seguir instrucciones. Para cada tarea se genera un prefijo describiendo la actividad a realizar que luego es demostrada con un ejemplo. A partir de entrenar el modelo base con una gran cantidad de datos, utilizando esta técnica, se obtiene una mayor capacidad de generalización y mejor desempeño en tareas nuevas no vistas previamente, sin realizar adaptaciones del modelo adicionales. Flan-T5 fue entrenado y evaluado en múltiples idiomas, consolidándose como un modelo eficiente para tareas lingüísticas complejas.

En base a la observación y evaluación humana de las respuestas generadas en un conjunto de evaluación de 10 *tweets* seleccionados al azar, y considerando las limitaciones en la disponibilidad de recursos, tanto computacionales como de anotación para evaluar con evaluadores humanos las respuestas de todos los modelos, decidimos utilizar para nuestros experimentos únicamente Flan-T5, por ser el modelo que mejores resultados obtuvo. Particularmente, utilizamos la versión Base de 250 millones de parámetros y la versión XL de 3 mil millones de parámetros (3B). Consideramos que su buen desempeño puede deberse parcialmente a su entrenamiento orientado a seguir instrucciones, lo que le brinda mayor capacidad de generalización a otras tareas, tal como fue desarrollado en Chung et al. [2022]. Los otros tres modelos generaron texto incoherente y sin sentido, con la excepción de GPT-J, que también generó una alta proporción de texto sin sentido pero con un poco más de coherencia interna, llegando incluso a producir unas pocas respuestas aceptables.

5.2 Configuraciones de Experimentación

Diseñamos una serie de experimentos para evaluar el impacto de ejemplos de alta calidad y de la información argumentativa en la generación automática de contranarrativas.

Proponemos distintos parámetros de configuración para definir el modo de aprendizaje, el tamaño de los modelos y la información argumentativa y los tipos de contranarrativas utilizadas. Buscamos evaluar el impacto que distintos factores puedan tener en la tarea de generación automática de contranarrativas.

A continuación describimos los distintos ejes utilizados para definir valores de configuración de los experimentos, que apuntan a evaluar distintos aspectos de los modelos de lenguaje y del uso que pueden o no darle a nuestro *dataset*:

Versión Base o versión XL del modelo Queremos medir el impacto del tamaño y la cantidad de parámetros del modelo a utilizar. Mientras que los modelos más grandes requieren más recursos computacionales, también logran generalizar su aprendizaje para adaptarlo más fácilmente a tareas para las que no fueron ajustados. Estos modelos demuestran mejor rendimiento que sus contrapartes de menor tamaño cuando se les provee de ejemplos dentro de la entrada misma en tiempo de inferencia (*in-context learning*). Los modelos más pequeños, por su parte, requieren menos recursos computacionales, son más portátiles, generan una huella de carbono inferior, y pueden ajustarse con mayor facilidad a una tarea determinada utilizando un conjunto de datos específicamente anotado. Sin embargo, la anotación de estos conjuntos de datos suele requerir de muchas horas de trabajo humano y, en ocasiones, de la formación de los/as anotadores/as en el campo y en el contexto relevante a la tarea. En este trabajo nos proponemos intentar responder a la siguiente pregunta: en la tarea de generación automática de contranarrativas, ¿es mejor utilizar un modelo más grande y con más parámetros de propósito general, utilizando una cantidad mínima de ejemplos anotados como prefijo de las inferencias (*fewshot learning*), o es mejor utilizar un modelo más pequeño pero ajustado a la tarea con un dataset específico de datos de calidad?

Ajuste fino vs Fewshot Con el fin de evaluar distintas estrategias de aprendizaje, comparamos un enfoque basado en ajuste fino, donde un modelo pre-entrenado mediante una tarea pretexto, se vuelve a entrenar para ser ajustado a una tarea específica, y otro basado en aprendizaje en contexto (*in-context learning*), donde los ejemplos utilizados para mostrarle al modelo como completar la tarea son provistos en la misma entrada del modelo en tiempo de inferencia. En particular, *fewshot learning* se considera un caso específico de aprendizaje en contexto en el cual el propio modelo fue pre-entrenado utilizando ejemplos anotados como prefijos del ejemplo a predecir, aprendiendo a realizar tareas a partir de ver ejemplos dentro de su entrada [Brown et al., 2020, ?].

Con o sin información argumentativa Para evaluar el posible impacto de la información argumentativa como información complementaria provista al modelo de lenguaje de forma complementaria al texto del *tweet* de odio proponemos distintas combinaciones de estos componentes utilizados para enriquecer tanto el proceso de pre-entrenamiento como el *prompt* utilizado para el aprendizaje en contexto. Las combinaciones evaluadas son las siguientes:

- Sin información argumentativa;
- Con el Colectivo y la Propiedad;
- Con la Justificación y la Conclusión;
- Utilizando toda la información argumentativa: el Colectivo y la Propiedad, la Justificación y la Conclusión con su correspondiente tipo y el Pivot

Utilizando sólo una estrategia particular de contranarrativa Para evaluar el impacto de los ejemplos de alta calidad, únicamente en los enfoques de ajuste fino (excluimos de este experimento los modelos que utilizan aprendizaje en contexto), ajustamos el modelo base con diferentes subconjuntos de ejemplos definidos por criterios argumentativos, es decir, según el tipo de contranarrativa tal como fueron descritas en la sección 3.1.6. Por lo tanto, distintos modelos base fueron ajustados utilizando:

- Todas las contranarrativas
- Solo contranarrativas de tipo A (496 ejemplos en Inglés, 101 ejemplos en Español)
- Solo contranarrativas de tipo B (238 ejemplos en inglés, 59 ejemplos en español)
- Solo contranarrativas de tipo C (467 en inglés, 97 en español)

Para cada una de las versiones que utilizan para el ajuste fino un único tipo de contranarrativa (A, B o C), entrenamos a su vez dos modelos: uno enriquecido con el tipo de información argumentativa relevante y otro sin información argumentativa adicional. La información argumentativa relevante es definida según el tipo de contranarrativa:

- para las contranarrativas de tipo A, agregamos la Justificación y la Conclusión;
- para el tipo de contranarrativa de tipo B agregamos el Colectivo y la Propiedad;
- para el tipo de contranarrativa de tipo C agregamos únicamente la Justificación.

Respecto a los modelos entrenados con ajuste fino, entrenamos solo la versión Base de Flan-T5 usando todos los tipos de contranarrativas con cada una de las variantes de información argumentativa (4 modelos) y luego un modelo utilizando únicamente uno de los tres tipos de contranarrativa y a su vez, cada uno de estos con y sin utilizar la información argumentativa relevante (2 modelos por cada uno de los 3 tipos de contranarrativas, en total 6 modelos). Luego generamos contranarrativas utilizando cada uno de estos 10 modelos ajustados a la tarea. Respecto a los modelos que utilizaron técnicas de aprendizaje en contexto (*in-context learning*), en particular *fewshot-learning*, generamos contranarrativas utilizando los modelos Base y XL de Flan-T5, cada uno de estos con cada una de las variantes de información argumentativa (2 modelos con 4 variantes, 8 configuraciones de generación). Esto nos da un total de 18 configuraciones de generación para cada idioma, es decir, un total de 36 configuraciones de generación en total.

5.3 Métodos de aprendizaje

Como mencionamos brevemente en la sección previa, existen dos estrategias para adaptar modelos pre-entrenados a una tarea específica. Una es el ajuste fino, mediante el

cual un modelo pre-entrenado en una tarea pretexto es sometido a una segunda ronda de entrenamiento con un conjunto de datos específico de la tarea a aprender y actualiza sus pesos a la nueva tarea. La otra estrategia, surgida más recientemente, es el aprendizaje en contexto (*in-context learning*), mediante el cual el modelo pre-entrenado incorpora la información necesaria para aprender la tarea en tiempo de inferencia a partir de ejemplos provistos en la entrada, manteniendo intactos los pesos del modelo pre-entrenado. La estrategia de ajuste fino es por lo general más eficaz pero, a la vez, más costosa, tanto en términos de recursos computacionales, considerando que la actualización de los pesos es mucho más costosa en términos de procesamiento que la inferencia, como en términos de recursos, dado que se necesita un conjunto de datos anotados de calidad, el cual no siempre está disponible y generarlo conlleva un costo considerable. La estrategia de aprendizaje en contexto, en oposición, requiere pocos datos para enseñar al modelo a realizar la tarea requerida pero depende de la capacidad subyacente del modelo de generalizar su conocimiento previo a tareas nuevas a partir de unos pocos ejemplos.

Los modelos ajustados a instrucciones, como el *Flan-t5* que utilizamos en este trabajo, han demostrado una mayor capacidad de adaptación a tareas no vistas previamente [Wei et al., 2022], por lo que son una buena alternativa para el aprendizaje en contexto. En función de explotar la capacidad del modelo de seguir instrucciones, ambas estrategias de aprendizaje utilizarán un prefijo en todas sus entradas instruyendo la elaboración de respuestas contrarias al mensaje de odio. El prefijo fue elaborado siguiendo el estilo sencillo y directo del conjunto de instrucciones FLAN. A la vez, utilizamos *tokens* especiales agregados previamente al vocabulario del *tokenizador* para marcar el inicio de un discurso de odio (<SHS>) y su fin (<EHS>), y el comienzo de una contranarrativa (<SCN>) y su fin (<ECN>). Debido a limitaciones computacionales al momento de realizar estos experimentos, el ajuste fino sólo se llevó a cabo con la versión base del modelo Flan-t5. Esto nos permite, de todos modos, desarrollar nuestro objetivo de comparar la diferencia marginal del desempeño respecto al mismo modelo Flan-t5 en su versión XL pero utilizando aprendizaje en contexto, pudiendo analizar si impacta más aumentar el tamaño del modelo pre-entrenado o realizar un ajuste fino sobre la versión más pequeña.

5.3.1 Ajuste fino

Para realizar el ajuste fino, separamos de manera aleatoria nuestro conjunto de datos en conjuntos de entrenamiento, desarrollo y evaluación, asegurándonos de que todas las contranarrativas a un mismo tweet queden dentro del mismo conjunto. La tabla 5.1 muestra el detalle de los tamaños y proporciones de cada partición. El entrenamiento se realizó con un *learning-rate* de $2e-05$ durante 8 epochs. La plantilla utilizada para la entrada de los modelos entrenados con ajuste fino puede verse en la figura 5.1.

Entrada: <pre>‘‘Generate a non-hateful response for the following hate messages: <SHS>’’ + Tweet + ‘‘<EHS> => ’’</pre> Etiqueta: <pre>‘‘<SCN>’’ + Counter-narrative + ‘‘<ECN>’’</pre>
--

Fig. 5.1: Formato de la entrada y la etiqueta objetivo del entrenamiento en modelos Seq2Seq como el Flan-T5 utilizado.

	Inglés						Español						
	Tweets	CNs	% corpus	A	B	C	Tweets	CNs	% corpus	A	B	C	
Train	509	1201	69.8 %	496	238	467	105	257	72.4 %	101	59	97	
Dev	71	173	10.0 %	67	38	68	12	27	7.6 %	12	8	7	
Test	145	348	20.2 %	138	74	136	27	71	20 %	27	21	23	
Proporción de <i>tweets</i> con contranarrativa													
				96 %	47 %	90 %					97 %	61 %	89 %

Tab. 5.1: Tamaño de las particiones de los conjuntos de datos en inglés y español. Las columnas A, B y C muestran la cantidad de contranarrativas utilizadas para cada partición al entrenar únicamente con contranarrativas de un tipo determinado.

5.3.2 Aprendizaje en contexto

Para el aprendizaje en contexto, además de la instrucción describiendo la tarea, incorporamos discursos de odio anotados con sus correspondientes contranarrativas, tomadas del conjunto de datos de entrenamiento, a ser incorporadas como prefijo de cualquier entrada del modelo. Los ejemplos fueron seleccionados al azar. Para determinar la cantidad adecuada de ejemplos a incorporar, probamos generar entradas de los modelos con 2, 5 y 10 ejemplos, en distintas combinaciones. El modelo Flan-T5 tiene una limitación en el tamaño de su entrada de 512 *tokens*, por lo que buscamos que las entradas generadas no superen esa longitud. En el caso de las entradas de los modelos generadas con 2 ejemplos, ninguna superó ese tamaño, mientras que para las entradas de 5 y 10 ejemplos, todas superaron ese tamaño. Por lo tanto, para todos los experimentos realizados con aprendizaje en contexto utilizaremos una configuración *2-shot*. La plantilla utilizada para generar la entrada de los modelos entrenados utilizando aprendizaje en contexto, junto con los ejemplos seleccionados y agregados también a la entrada, puede verse en la figura 5.2.

```
"Generate a non-hateful response for the following hate messages:

<SHS>unbelievable census bureau: 44 6% in ca dont hashtag speakenglish
at home 35 6% 35 6% in tx 34 5% in nm 31 7% in nj 31 0% in ny our
schools r getting 150% funding for every student who doesnt speak
hashtag english well parents job hashtag stop immigration :link: <EHS>
=> <SCN>And would ending immigration improve English? I don't think
so... <ECN>

<SHS>100 bangladeshi nationals apprehended near texas border in 3
weeks via @user hashtag noamnesty hashtag deport illegals hashtag
buildthatwall hashtag trump hashtag maga hashtag masa<EHS> => <SCN>Do
you know why where they apprehended? Why do you assume they were
illegals? <ECN>

<SHS>' ' + Tweet + ' ' <EHS> => <SCN>"
```

Fig. 5.2: Entrada del modelo para la generación utilizando *fewshot-learning* con dos ejemplos

5.4 Evaluación de modelos

La correcta evaluación de las contranarrativas generadas automáticamente por nuestros modelos es esencial para poder medir el impacto de la información argumentativa, de los tipos de contranarrativas, de las estrategias de aprendizaje, del tamaño de los modelos o de cualquier otra variable.

Las limitaciones de las métricas automáticas de evaluación fueron desarrolladas en la sección 2.3. En la sección 6.1 presentamos también un análisis propio sobre la inadecuación de algunas métricas de evaluación automática, mostrando las falencias en su aplicación a esta tarea y a nuestro conjunto de datos en particular.

La alternativa que se presenta ante la evaluación automática tradicional es la evaluación manual mediante anotadores humanos. Distintos autores han optado por este mecanismo, tal como se desarrolla en la sección 2.3.1. Debido a las limitaciones de utilizar anotadores humanos, realizamos la evaluación sobre un subconjunto limitado de ejemplos de la partición de evaluación.

A continuación detallamos el protocolo y las categorías de evaluación adoptadas en este trabajo.

5.4.0.1 Protocolo de evaluación

Sobre la base de los trabajos previos descritos en la sección 2.3, elaboramos un protocolo de evaluación de pares de discurso de odio y contranarrativas que permite identificar los aspectos clave, necesarios para distinguir su adecuación como respuestas, a la vez que mantiene la robustez y reproducibilidad del proceso de anotación.

Con este objetivo, elaboramos una propuesta con cuatro categorías de análisis. A cada categoría se le puede asignar uno de tres valores posibles (del 1 al 3) dentro de una escala Likert [Likert, 1932] donde los valores más bajos representan las opciones menos deseables y los más altos, las más deseables, con la única excepción de la categoría Informatividad, que puede anotarse con el valor 0 si no le corresponde ser anotada y, por lo tanto, tiene 4 valores posibles. Las categorías de anotación son las siguientes:

- **Ofensividad:** Tomada de Ashida and Komachi [2022]. Indica si la contranarrativa generada es ofensiva para el colectivo objetivo, el autor del *tweet* o cualquier otra persona o grupo. Sus tres valores posibles son: (1) Ofensivo; (2) Es ambiguo o no puede determinarse; (3) No es ofensivo.
- **Posición:** Tomada de Ashida and Komachi [2022]. Indica si la contranarrativa generada tiene una posición contraria al *tweet* de odio que responde o no. Sus tres posibles valores son:
 1. Tiene la misma posición que el *tweet* de odio. Este valor tiene una relación cercana con la categoría anterior dado que si comparte la misma posición del *tweet* de odio es altamente probable que también sea ofensivo.
 2. Es ambiguo o no puede determinarse. En general encontramos que las contranarrativas a las que les corresponde este valor son aquellas que cambian de tema o no responden ningún aspecto central del mensaje de odio.
 3. Tiene una posición contraria a la del *tweet* de odio y esta puede inferirse sin ambigüedad.

- **Informatividad:** Esta categoría es utilizada en Ashida and Komachi [2022] y en Chung et al. [2021a]. Indica si la contranarrativa constituye una respuesta generica y simple o específica y elaborada respecto al mensaje de odio original. Intuitivamente, queremos observar si la respuesta toma elementos del mensaje original y elabora sobre estos para contestar o si, por el contrario, es sencillamente una narrativa contraria al odio que podría aplicarse de manera genérica a un conjunto grande de mensajes de odio sin importar su contenido. Esta categoría sólo debe ser completada si el valor de la Posición es (3), es decir, si la contranarrativa tiene una posición contraria al mensaje de odio original. Hay, por lo tanto, cuatro valores posibles:

0. No asignamos un valor porque la contranarrativa no tiene una posición contraria al mensaje de odio. Este valor sólo se usa cuando el valor de Posición es distinto de 3.
1. La contranarrativa es genérica y/o poco elaborada. No se incorpora en la respuesta ningún contenido específico del mensaje que se responde por lo que la contranarrativa podría ser utilizada potencialmente en un amplio conjunto de mensajes de odio (ejemplos: “Lo que dices no es cierto”, “Eso es xenófobo”, “¿Tenés pruebas de lo que decís?”)
2. La contranarrativa considera contenido específico del mensaje de odio pero no elabora al respecto ni lo utiliza para argumentar (por ejemplo, dado el mensaje X se responde “No es cierto que [contenido parcial de X]”)
3. La contranarrativa es específica y tiene algún grado de elaboración sobre el contenido de la respuesta original. Identificamos tres patrones típicos que asociamos con este valor:
 - A - Respuestas que toman más de un elemento del mensaje de odio original y plantean alguna relación entre éstas. Por ejemplo, “No entiendo qué relación tiene [** elemento del mensaje de odio **] con [** otro elemento del mensaje de odio **]”
 - B - Una declaración sobre un elemento particular del mensaje original pero acompañada de una explicación o apreciación personal. Por ejemplo “No creo que tengamos que [** política mencionada en el mensaje de odio **]. Es una mala idea, no va a funcionar”.
 - C - Una respuesta argumentativa basada en información no necesariamente mencionada en el mensaje de odio pero inferida a partir de éste, mostrando una comprensión integral del significado del mensaje de odio. Por ejemplo, un mensaje cuya conclusión es “#BuildTheWall” (#ConstruyamosElMuro) se responde aludiendo a que el costo de construcción del muro sería más elevado que el supuesto costo generado por los inmigrantes.

- **Felicidad:** Esta categoría está relacionada con la Intra-coherencia propuesta por Chung et al. [2021a] pero considerando a su vez, la correctitud sintáctica y semántica y la fluidez del texto. Evalúa de la contranarrativa, con independencia del mensaje que responde, si el texto generado suena, en sí mismo, fluido y correcto. Hay tres valores posibles:

1. El texto es incoherente o es sintáctica o semánticamente incorrecto.

2. El texto es coherente y puede entenderse su significado sin demasiado esfuerzo, pero tiene errores menores como no coordinar número o género de las palabras, manejar mal la puntuación, repetir palabras o repetir partes literales del mensaje de odio sin adaptarlas (por ejemplo, ante un mensaje que plantee “hagamos esto y pasara esto otro”, una respuesta como “No es cierto que si hagamos esto y pasara esto otro” en lugar de “No es cierto que si hacemos esto pasará esto otro”)
3. El texto es fluido y suena correcto.

Utilizando estas cuatro categorías de manera agregada, definimos dos conceptos de clasificación de las contranarrativas con el objetivo de refinar nuestro concepto de adecuación de las mismas:

- **Contranarrativas Aceptables:** son aquellas que tienen valores óptimos (3) en las clases de Ofensividad, Posición y Felicidad. Conceptualmente, representan a aquellas contranarrativas que son seguras para un sistema automático de generación de contranarrativas. Consideramos a una contranarrativa como “segura” si no quita credibilidad al sistema de generación de contranarrativas. Si una contranarrativa reproduce odio en sí misma y/o no tiene una posición claramente contraria al mensaje de odio que responde y/o es incoherente o no parece escrita por alguien con dominio del lenguaje, además de ser una mala respuesta, quita credibilidad al sistema en sí. Una condición necesaria para considerar el rendimiento de un modelo, por lo tanto, debería ser el tener un porcentaje mínimo de generación de contranarrativas que no sean aceptables.
- **Contranarrativas Buenas:** son aquellas que tienen valores óptimos en las cuatro categorías de evaluación propuestas. Es decir, son contranarrativas que son aceptables y a su vez, tienen un valor óptimo en la categoría “Informatividad”. Conceptualmente, representan a las contranarrativas que aportan credibilidad al sistema.

5.4.0.2 Puntaje final agregado

Con el objetivo de brindar una herramienta más que colabore a simplificar la comparación del rendimiento entre distintos modelos, proponemos una fórmula para calcular un puntaje final agregado utilizando las cuatro categorías propuestas. Este puntaje es indicativo y debe ser interpretado y analizado sin dejar de considerar los valores desglosados de evaluación y, sobre todo, el porcentaje de contranarrativas Aceptables y Buenas. Para calcularlo, sumamos el puntaje de cada categoría y lo normalizamos a un rango entre 1 y 10. Como el valor más pequeño que puede obtener una contranarrativa es 1 en todas las categorías excepto en Informatividad, donde puede tener 0 (sumando por lo tanto 3) y el valor máximo que puede obtener es 3 en cada una de las cuatro categorías (sumando por lo tanto 12), proponemos entonces como valor final restarle 2 a la suma de las cuatro categorías de modo tal que el valor mínimo de esta cuenta sea 1 y el máximo 10. La fórmula para calcular este puntaje se muestra a continuación:

$$S = \frac{1}{n} \sum_i^n Of_i + Pos_i + Felic_i + Inf_i - 2 \quad (5.1)$$

Si bien la fórmula propuesta no distingue en modo alguno ni asigna pesos o prioridades distintas a ninguna categoría, la naturaleza del protocolo de evaluación propuesto provoca

que una categoría tenga mayor influencia en el valor final que el resto: la Posición. Esto se debe a que si la Posición tiene un valor menor al máximo, la Informatividad necesariamente será 0, mientras que si tiene el valor máximo, la Informatividad será al menos 1.

5.4.0.3 Limitaciones de nuestras categorías de análisis

Hay dos aspectos que consideramos importantes a la hora de evaluar la adecuación de un sistema de contranarrativas que quedan fuera de nuestra propuesta de evaluación. La creatividad o variabilidad entre las respuestas es un aspecto importante debido a que si un sistema repite constantemente o usa con excesiva frecuencia conceptos o incluso frases enteras, pierde naturalidad y corre el riesgo de revelarse como sistema automático. Sin embargo, tiene la dificultad de no poder evaluarse individualmente con un mensaje de odio y su respuesta, sino que requiere la comparación entre todas las respuestas brindadas por el sistema. La persuasión es otro aspecto importante de considerar, porque permite dejar de discutir en términos de adecuación y comenzar a hacerlo en términos de efectividad, hilando, a su vez, más fino en la evaluación de los sistemas. Sin embargo, ésta constituye un fenómeno complejo para el cual es difícil llegar a un acuerdo reproducible sobre lo que es persuasivo y lo que no, por lo que decidimos no incorporarla en nuestra propuesta.

5.4.1 Anotación humana de las evaluaciones

A continuación describimos el proceso de anotación de las contranarrativas generadas utilizando las categorías de evaluación desarrolladas en la sección 2.3.1.

5.4.1.1 Curation Rationale

Para realizar la evaluación manual seleccionamos 20 *tweets* aleatorios del conjunto de datos de test en inglés y 10 *tweets* aleatorios del conjunto de datos de test en español. Debido a que ciertas configuraciones sólo fueron entrenadas con un único tipo particular de contranarrativa, decidimos para esos casos, utilizar conjuntos de evaluación donde todos los ejemplos tengan ese tipo de contranarrativas anotadas. Por lo tanto, definimos otros dos conjuntos no disjuntos: para el primero, tomamos todos los *tweets* del primer conjunto que tuvieran anotados un Colectivo y Propiedad, y por lo tanto que tuvieran una contranarrativa de tipo B y lo completamos con otro conjunto de *tweets* extra, anotados también con contranarrativas de tipo B seleccionados de manera aleatoria del resto de los datos del conjunto de evaluación que no fueron seleccionados para el primer conjunto, de modo tal que este segundo conjunto también tenga 20 *tweets* en inglés y 10 en español. Para el tercer conjunto de evaluación realizamos el mismo procedimiento pero esta vez garantizando que todos los ejemplos tengan una contranarrativa anotada de tipo C. En este último conjunto, encontramos que todos los ejemplos ya tienen una contranarrativa de tipo A anotada, con lo cual lo usamos para evaluar los modelos entrenados con uno u otro tipo. En total, para estos tres subconjuntos utilizamos 47 *tweets* de odio en inglés de un total de 145 pertenecientes a la partición de evaluación y 12 en español de un total de 27.

Para cada uno de los modelos detallados en la sección 5.2 se generaron contranarrativas para los 20 *tweets* en inglés y 10 *tweets* en español. Habiendo 18 configuraciones distintas de ajuste de los modelos a la tarea, tenemos en total 360 pares de *tweet* de odio y contranarrativa en inglés y 180 pares en español para ser evaluados utilizando las

categorías descritas en 2.3.1. Si bien las contranarrativas generadas no sufrieron ningún tipo de post-procesamiento, identificamos un problema en los textos generados en español: la aparición de caracteres *unknown* (<unk>) en lugares donde debiera haber un caracter con tilde u otros caracteres no utilizados en inglés, como la letra ñ. Para estos casos en particular, los anotadores fueron instruidos a asumir que el texto generado utilizaba los caracteres correctos.

5.4.1.2 Equipo de anotación

La anotación de los 540 pares de discurso de odio y contranarrativa (360 en inglés y 180 en español) fue realizada por un equipo de tres anotadores de entre 25 y 35 años, dos varones y una mujer. Los dos varones son los mismos que realizaron la anotación de las contranarrativas detallada en la sección 3.2.2, es decir, un estudiante universitario de la Licenciatura en Filosofía y un estudiante del Doctorado en Ciencias de la Computación. La tercera anotadora es graduada de la carrera de Biología, sin conocimientos específicos en modelos de lenguaje, contranarrativas, argumentación ni ningún otro tema relacionado con este trabajo. Los tres anotadores son hablantes nativos del español en su variedad dialectal rioplatense y tienen un alto nivel oral y escrito en inglés. Nuevamente, ninguno de los anotadores pertenece a un colectivo víctima de odio xenofóbico o racial. Exceptuando al estudiante de Doctorado que realizó la anotación como parte de su tesis doctoral y no recibió compensación, los otros dos anotadores recibieron una compensación de 25 USD por hora, por un trabajo de 5 horas de duración en total cada uno. Los tres anotadores realizaron una capacitación de una hora donde discutieron sobre los criterios para distinguir los posibles valores de cada categoría, sistematizando los acuerdos en un documento escrito e intentando abarcar la generalidad de los casos para mejorar el acuerdo entre anotadores.

5.4.1.3 Acuerdo entre anotadores

La totalidad de los pares de discurso de odio y contranarrativas seleccionados fueron evaluados por los tres anotadores. Sobre la base de esta anotación por triplicado, calculamos el acuerdo entre ellos utilizando el κ de Cohen [Cohen, 1960]. La tabla 5.2 muestra el acuerdo entre todas las combinaciones de pares de anotadores. Podemos observar que en la mayoría de los casos el acuerdo oscila entre “moderado” (entre .4 y .6) y “sustancial” (entre .6 y .8), con la única excepción del acuerdo del anotador 3 con los otros dos anotadores para la categoría de Felicidad, donde se evidencia una diferencia importante en el entendimiento del concepto, obteniendo un acuerdo “justo” (entre .2 y .4). Los otros anotadores, sin embargo, para la misma categoría obtuvieron el acuerdo más alto de todos, de .67. La categoría Posición muestra el mejor acuerdo en promedio (“sustancial”), mientras que la Ofensividad y la Informatividad muestran valores en el rango de un acuerdo moderado. Considerando la naturaleza interpretativa y subjetiva de la tarea, consideramos que el acuerdo obtenido es razonable.

5.4.1.4 Puntaje final de la evaluación

Considerando que existen tres juicios para cada par a evaluar de discurso de odio/-contranarrativa, aplicamos un esquema de voto mayoritario para calcular el valor final de cada categoría y del valor final agregado y normalizado de 1 a 10. Dado que existen

	1 vs 2	2 vs 3	1 vs 3
Ofensividad	.47	.40	.41
Posición	.63	.58	.63
Informatividad	.49	.42	.54
Felicidad	.67	.37	.36

Tab. 5.2: Acuerdo entre los anotadores 1, 2 y 3 utilizando el κ de Cohen.

tres valores posibles para cada categoría, el voto mayoritario entre los tres anotadores no garantiza una decisión unánime. Los casos en donde no hubo una decisión mayoritaria por un valor determinado, sin embargo, fueron escasos, contabilizando en total 10. Para estas situaciones, tomamos un criterio conservador y asignamos el valor más bajo entre los posibles valores evaluados por los anotadores.

5.5 Parámetros de generación de textos

Los modelos de generación de texto permiten aplicar distintos parámetros de generación que pueden redundar en distintos textos generados aunque se utilice el mismo modelo con la misma entrada. En particular, previo a realizar la generación, nos proponemos explorar cuatro configuraciones distintas:

- **Generación Golosa (*Greedy*):** consiste en generar texto mediante una estrategia de decodificación en la que, en cada paso, el modelo selecciona la palabra (o *token*) con la mayor probabilidad de acuerdo con su distribución de salida. Esta técnica no considera alternativas menos probables, lo que puede llevar a resultados que, aunque localmente óptimos, pueden ser globalmente subóptimos, produciendo secuencias coherentes pero menos variadas o creativas en comparación a otras técnicas
- ***Beam Search*:** consiste en una técnica de decodificación orientada a maximizar la coherencia del texto generado en relación a la generación golosa. En lugar de elegir solo el *token* más probable en cada paso, *Beam Search* mantiene un conjunto de las mejores secuencias candidatas, o “*beams*”, a lo largo del proceso de generación, lo que le permite elegir la secuencia candidata en función de la probabilidad conjunta acumulada de toda la secuencia y no de un sólo *token* generado paso a paso. La longitud de las secuencias candidatas se puede configurar por parámetro: una secuencia más larga garantiza mayor coherencia en la generación pero a la vez aumenta significativamente los costos computacionales. En este trabajo, tomamos un tamaño de *beam* de 4 *tokens*.
- **Muestreo con Temperatura:** la técnica de decodificación utilizando muestreo (o *sampling* en inglés) consiste en seleccionar el siguiente *token* en función de la distribución de probabilidades generada por el modelo en cada paso de decodificación. Es decir que la elección del siguiente *token* generado es aleatoria pero condicionada a esta distribución: los *tokens* más probables tendrán mayor probabilidad de ser seleccionados mientras que los *tokens* con menor probabilidad tendrán menores probabilidades. La temperatura se introduce para modificar y controlar la dispersión de la distribución de probabilidad: una temperatura menor a 1 acrecienta la diferencia entre los *tokens* menos probables y más probables, acrecentando la posibilidad

de elegir *tokens* más probables y por ende, haciendo que la generación priorice la coherencia y se vuelva más conservadora. Una temperatura mayor a 1 achata la distribución de probabilidades de los *tokens* acrecentando la posibilidad de elegir *tokens* menos probables y por ende, provocando que la generación sea más creativa, a riesgo de perder coherencia. En este trabajo probamos y evaluamos una generación con muestreo y una temperatura de 0.7, priorizando la coherencia por sobre la creatividad y diversidad.

- **Muestreo con Top-P:** consiste en una técnica de generación con muestreo, tal como fuera descrita en el ítem anterior pero en lugar de utilizar una temperatura para modificar la distribución de probabilidades, se selecciona de forma dinámica un subconjunto de *tokens* empezando por aquellos con mayor probabilidad hasta que la suma de probabilidades acumulada alcance un determinado valor p predefinido. Luego se realiza un muestreo aleatorio y equiprobable entre los *tokens* de ese subconjunto. En este trabajo utilizamos un valor de $p = 0.92$.

En las cuatro configuraciones de generación descritas utilizamos una penalización por repetición de 4 n-gramas para evitar texto repetido y otorgar fluidez.

Por una cuestión de limitación de recursos, sobre todo de evaluación de contranarrativas generadas, no podemos repetir todos los experimentos con cada una de estas configuraciones. Por lo tanto, decidimos realizar un único experimento previo a encarar la evaluación de los modelos en sí, con un único modelo para el cual utilizamos un subconjunto de evaluación de 10 tweets en inglés seleccionados aleatoriamente. Para esta selección, generamos una contranarrativa por cada *tweet* utilizando un modelo Flan-T5-Base ajustado con el conjunto de datos de entrenamiento sin información argumentativa y luego realizamos una evaluación manual basándonos en las categorías descritas en la sección 2.3.1. Los resultados para cada una de las cuatro técnicas de generación de texto pueden observarse en las tablas 5.3 y 5.4. La tabla 5.3 muestra para cada categoría (“Ofensivo”, “Posición”, “Informatividad” y “Felicidad”) el porcentaje de ejemplos que fue etiquetado con cada posible valor (1, 2 o 3) y el porcentaje de ejemplos clasificados como “Aceptables” y “Buenos”. Notar que, mientras que para las categorías “Ofensivo”, “Posición” y “Felicidad” la suma de los porcentajes para cada posible valor debe sumar 100 %, en el caso de “Informatividad” la suma debe dar el mismo porcentaje obtenido en la columna “3” de la categoría “Posición”, dado que si el ejemplo no tiene este valor en esta categoría, la “Informatividad” se considera como 0. La tabla 5.4, por su parte, muestra el puntaje agregado, calculado tal como se describe en la sección 5.4.0.2. Los resultados demuestran que la generación mediante *Beam Search* produce los mejores resultados, seguida de cerca por la generación Golosa. Los métodos que involucran muestreo, si bien tienen una mayor variabilidad, muestran un desempeño menor en términos de calidad de la respuesta generada. Particularmente, el muestreo utilizando *Top-P* (el de mayor variabilidad, ya que la temperatura en el otro muestreo fue utilizada para aumentar las posibilidades de elegir respuestas más probables) genera una cantidad muy elevada de contenido ofensivo y con posiciones favorables al mensaje de odio original. Basándonos en estos resultados, durante todo el resto de los experimentos de generación de contranarrativas de este trabajo, utilizaremos *Beam Search* con 4 *beams*.

	Ofensivo			Posición			Informatividad			Felicidad			Accept	Buenas
	1	2	3	1	2	3	1	2	3	1	2	3		
Goloso	10 %	20 %	70 %	0 %	40 %	60 %	10 %	20 %	30 %	10 %	40 %	50 %	30 %	10 %
Muestreo c/Temp	20 %	30 %	50 %	20 %	60 %	20 %	0 %	10 %	10 %	40 %	10 %	50 %	10 %	10 %
Muestreo c/Top P	80 %	0 %	20 %	40 %	20 %	40 %	10 %	0 %	30 %	40 %	0 %	60 %	20 %	10 %
Beam Search	10 %	0 %	90 %	10 %	0 %	90 %	0 %	30 %	60 %	0 %	30 %	70 %	70 %	50 %

Tab. 5.3: Resultado de la evaluación manual de diferentes estrategias de decodificación. La tabla muestra, para cada categoría de análisis, el porcentaje de ejemplos clasificados con la respectiva etiqueta (1, 2 o 3) y luego el porcentaje de ejemplos considerados como “aceptables” y “buenos”

	Puntaje	% Aceptables	% Buenas
Goloso	7.00	30 %	10 %
Muestreo c/Temp	4.60	20 %	10 %
Muestreo c/Top p	4.90	20 %	10 %
Beam Search	8.70	70 %	50 %

Tab. 5.4: Puntajes agregados para diferentes estrategias de generación usando un modelo Flan-T5 con ajuste fino. El mejor resultado se encuentra destacado en negrita

5.6 Resultados

La tabla 5.6.4 muestra los resultados de la evaluación manual de las contranarrativas para *tweets* de odio generadas automáticamente por las distintas versiones de los modelos basados en Flan-T5, en inglés y español. Se muestran los puntajes calculados según se describe en la sección 5.4.0.2 junto al porcentaje de contranarrativas consideradas como “Aceptables” y “Buenas” para las distintas configuraciones experimentales tal como se describen en la sección 5.2. Podemos observar que 8 configuraciones tienen que ver con el aprendizaje en contexto, mediante el cual se proveen ejemplos sobre cómo se debe realizar la tarea en la entrada del modelo, sin realizar ningún ajuste de los pesos del mismo. De estas 8 configuraciones, 4 utilizan la versión Base de Flan-T5 (0.25B) mientras que otras 4 utilizan la versión XL (3B), donde a su vez, estas 4 se distinguen entre sí según la información argumentativa utilizada: ninguna, el Colectivo y la Propiedad, la Justificación y la Conclusión o toda la información argumentativa junta. Las otras 10 se corresponden con modelos ajustados a través del conjunto de datos de entrenamiento de CONEAS, consistente en pares de discurso de odio y contranarrativas. Mientras que 4 configuraciones utilizan la totalidad de las contranarrativas de este conjunto de datos, cada una con las mismas combinaciones de información argumentativa enumerada en el párrafo anterior, otras 6 utilizan sólo uno de los tres tipos de contranarrativas: A, B o C. Y a la vez, para cada tipo de contranarrativa exploramos la generación con y sin la información argumentativa relevante para ese tipo: la Justificación y Conclusión para los tipos A y C y el Colectivo para las de tipo B.

La tabla 5.8 muestra los valores desglosados por categoría utilizados para calcular el puntaje final de la tabla 5.6.4 para el inglés y el español, mientras que las tablas 5.9 y 5.10 muestran, para el inglés y el español respectivamente, el porcentaje de contranarrativas evaluadas con el menor y el mayor valor de cada categoría. A su vez, las figuras 5.4 y 5.5 muestran el porcentaje de contranarrativas generadas que fueron clasificadas como “Aceptables” y “Buenas” para el idioma inglés y español respectivamente. Por último, las tablas 5.5 y 5.6 comparan las respuestas generadas por cada modelo para un mensaje

de odio en inglés y uno en español respectivamente, seleccionados del subconjunto de evaluación que fue evaluado manualmente. Junto con la respuesta generada se presenta la evaluación final asignada por los anotadores a esa contranarrativa.

En términos generales, ninguna técnica de aprendizaje logró que ninguno de los modelos Flan-T5 tuviera un desempeño adecuado como para ser puesto en producción. El mejor modelo tiene sólo un 60 % de contranarrativas aceptables, lo cual implica que necesariamente produjo un 40 % de contranarrativas que no deberían ser generadas por un sistema en producción en un escenario de la vida real.

A continuación presentamos un análisis de los resultados de acuerdo a cada uno de los ejes planteados para el desarrollo de la experimentación.

5.6.1 Ajuste fino vs. aprendizaje en contexto

Los resultados muestran que el enfoque de ajuste fino entrenado con un conjunto de datos anotados de calidad, incluso sin ser masivo, produce un desempeño mucho mejor que el aprendizaje en contexto, al menos para los modelos Flan-T5. Comparando las técnicas de aprendizaje del mismo modelo base para el idioma inglés, podemos observar en la tabla 5.6.4 que el aprendizaje en contexto utilizando *fewshot learning* obtiene entre un 20 % y un 40 % de contranarrativas aceptables y entre 0 % y 5 % de contranarrativas buenas, mientras que el modelo ajustado con nuestro conjunto de datos obtiene entre un 25 % y un 60 % de contranarrativas aceptables y entre un 10 % y un 30 % de contranarrativas buenas. El puntaje obtenido por los modelos entrenados con ajuste fino es invariablemente entre 1.5 y 2 puntos superior al de su contrapartida entrenada mediante aprendizaje en contexto. En la tabla 5.9 observamos que la mayor diferencia se da en la categoría de “Posición”, donde los cuatro experimentos realizados con aprendizaje en contexto promedian un 25 % de contranarrativas con posición contraria al *tweet* de odio original, mientras que los cuatro experimentos realizados con ajuste fino promedian el 50 %. La categoría de informatividad también presenta diferencias significativas, por un lado, como consecuencia del aumento de contranarrativas con posición contraria al mensaje de odio, lo cual incrementa el número de etiquetas de informatividad mayores a 0, pero por el otro, la misma proporción de contranarrativas con valor máximo en la categoría (consideradas como “elaboradas” o “argumentales”) es mayoritaria en los experimentos de ajuste fino, mientras que en los experimentos de aprendizaje en contexto la categoría mayoritaria es la de menor valor (aplicada a respuestas “genéricas” no dirigidas claramente hacia un mensaje en particular). Nuevamente esta diferencia se aprecia en la figura 5.4, donde puede verse que para los modelos ajustados a los datos de entrenamiento, no sólo la cantidad de contranarrativas aceptables (gris claro + gris oscuro) es superior, sino que dentro de estas, la proporción de contranarrativas buenas es mucho mayor.

Tal como se puede ver en el ejemplo de la figura 5.5, las contranarrativas generadas por los modelos Base+Fewshot no demuestran una postura contraria al mensaje de odio clara y en varias ocasiones simplemente repiten todo o parte del mensaje original. Encontramos un número de contranarrativas generadas por modelos con la información de las premisas o de todos los componentes donde se devuelve un fragmento que coincide enteramente con la conclusión del mensaje de odio, sugiriendo que quizás el modelo no termina de comprender la tarea que debe realizar, confundiéndola con la identificación automática de los componentes argumentativos. Respecto al idioma español, notamos que el rendimiento del modelo Flan-T5 es inferior en todas las categorías experimentales, algo que puede observarse claramente al comparar las figuras 5.4 y 5.5, pero que esta diferencia es mayor,

TWEET: @user must deport all illegal migrants india already reeling under constant threat of muslim radicals curb population

		Of	Pos	Inf	Fel
Base-Fewshot-SinInfo:	"I'm sorry to hear that. I'm not sure if it's a joke or a joke, but I'd like to know."	1	1	0	2
Base-Fewshot-Colectivo:	"I don't know why they were apprehended."	1	1	0	1
Base-Fewshot-Premisas:	"@user must deport all illegal migrants"	1	1	0	2
Base-Fewshot-All:	"@user must deport all illegal migrants"	1	1	0	2
XL-Fewshot-SinInfo:	"yep, muslim radicals are a threat to india."	1	1	0	3
XL-Fewshot-Colectivo:	"yep, muslim radicals are a threat to India."	1	1	0	3
XL-Fewshot-Premisas:	"Nevertheless, I think it's a good policy."	1	1	0	3
XL-Fewshot-Todo:	"Sadly, India is already reeling under constant threat of muslim radicals."	1	1	0	2
Base-Ajustado-SinInfo:	"I don't understand what you are trying to say. Besides, it's not true that Muslims are radicals."	3	3	3	3
Base-Ajustado-Colectivo:	"Why should we deport all of them? I don't understand why you should do that."	3	3	3	3
Base-Ajustado-Premisas:	"I don't understand why you should deport all of them. They are not radicals."	3	3	3	3
Base-Ajustado-Todo:	"I don't think India is reeling under a constant threat of Muslims radicals"	3	3	2	3
Base-CNsTipoA-SinInfo:	"I don't think it's a good idea to allow immigrants to stream across our undefended border. It's just a bad idea."	1	1	0	3
Base-CNsTipoA-Premisas:	"I don't understand why you would want to allow millions of illegals to stream across our undefended border. It's a joke."	1	1	0	3
Base-CNsTipoC-SinInfo:	"I don't think it's a problem. It's not a threat to the population, it's just a threat."	3	2	0	1
Base-CNsTipoC-Premisas:	"I don't think it's true that Muslims are radicals. It's not true that India is reeling under a constant threat of radicals."	3	3	3	3

TWEET: 100 bangladeshi nationals apprehended near texas border in 3 weeks via @user hashtag noamnesty hashtag deport illegals hashtag buildthatwall hashtag trump hashtag maga hashtag masa

		Of	Pos	Inf	Fel
Base-CNsTipoB-SinInfo:	"I don't think they are deporting people. They are just trying to make a living. I think they are trying to make some money. I think it's not fair to call them illegals. I mean, they aren't. They are not criminals, they are just criminals"	2	2	0	1
Base-CNsTipoB-Colectivo:	"I don't think they are illegals. I think they are. They are a bunch of people who don't want to be a part of a society. I think that's why they don't care about immigration. I think it's just a matter of time. I'm not sure if that's true or not, but it's not true."	2	2	0	1

Tab. 5.5: Ejemplos generados por las distintas versiones de modelos Flan-T5 para tweets en inglés, junto a la evaluación manual realizada de cada una de las categorías propuestas: **O**fensividad, **P**osición, **I**nformatividad y **F**elicidad.

TWEET: la inmigración de africa y países árabes es una estrategia para entrar a europa y atacarla desde dentro como hace mil años hicieron los barbaros para destruir el imperio romano el islan ataca una europa lacsa en leyes y desunida marina lepen tiene razón nada d inmigrantes :link:

		Of	Pos	Inf	Fel
Base-Fewshot-SinInfo:	“the immigration of africa and árabes countries is a strategy to enter Europe and atacar it from within as many millennia hicieron barbaros to destroy the roman empire el islan ataca a europa lacsa in laws and desunida marina lepen tiene ninguna razón nada de inmigrante”	1	1	0	1
Base-Fewshot-Colectivo	“the immigration of africa and árabes countries is a strategy to enter Europe and atacar it from within as many millennia hicieron barbaros to destroy the roman empire el islan ataca a europa lacsa in laws and desunida marina lepen tiene ninguna razón nada de inmigrante”	1	1	0	1
Base-Fewshot-Premisas:	“the immigration of africa and árabes countries is a strategy to enter Europe and atacar it from within as many millennia hicieron barbaros to destroy the roman empire el islan ataca a europa lacsa in laws and desunida marina lepen tiene ninguna razón nada de inmigrante”	1	1	0	1
Base-Fewshot-All:	“ the immigration of africa and árabes countries is a strategy to enter Europe and atacar it from within as many millennia hicieron barbaros to destroy the roman empire el islan ataca a europa lacsa in laws and desunida marina lepen tiene ninguna razón nada de inmigrante”	1	1	0	1
XL-Fewshot-SinInfo:	“es una estrategia para entrar a europa y atacarla desde dentro como hace mil años hicieron los barbaros para destruir el imperio romano el islan ataca una europa lacsa en leyes y desunida marina lepen tiene razón nada”	1	1	0	2
XL-Fewshot-Colectivo:	“es una estrategia para entrar a europa y atacarla desde dentro como hace mil años hicieron los barbaros para destruir el imperio romano el islan ataca una europa lacsa en leyes y desunida marina lepen tiene razón nada”	1	1	0	2
XL-Fewshot-Premisas:	“es una estrategia para entrar a europa y atacarla desde dentro como hace mil años hicieron los barbaros para destruir el imperio romano el islan ataca una europa lacsa en leyes y desunida marina lepen tiene razón nada”	1	1	0	2
XL-Fewshot-Todo:	“es una estrategia para entrar a europa y atacarla desde dentro como hace mil años hicieron los barbaros para destruir el imperio romano el islan ataca una europa lacsa en leyes y desunida marina lepen tiene razón nada”	1	1	0	2
Base-Ajustado-SinInfo:	“No es cierto que la inmigración de los países árabes es una estrategia para entrar a Europa y atacarla desde hace mil años.”	3	3	2	2
Base-Ajustado-Colectivo:	“No es cierto que la inmigración de los países árabes es una estrategia para entrar a Europa y atacarla desde dentro como hace mil años hicieron barbaros para destruir el imperio romano el islan.”	3	3	2	2
Base-Ajustado-Premisas:	“No es cierto que la inmigración de los países árabes es una estrategia para entrar a Europa y atacarla, y eso no tiene razón a los inmigrantes.”	3	3	2	2
Base-Ajustado-Todo:	“No es cierto que la inmigración de los países árabes es una estrategia para entrar a Europa y atacarla. No tiene razón a los inmigrantes.”	3	3	2	2
Base-CNsTipoA-SinInfo:	“No creo que es una estrategia para entrar a Europa y atacarla, y eso no tiene razón nada a los inmigrantes.”	2	3	2	1
Base-CNsTipoA-Premisas:	“No tiene razón nada a los inmigrantes, y eso no es una estrategia para entrar a Europa.”	2	3	2	1
Base-CNsTipoB-SinInfo:	“El imperio romano es el imperio europeo, y eso no es una razón para destruir la imperio romana, y no tiene nada en el mundo. No hay nada de inmigrantes, ni nada de niños. No tiene que encontrar esas razones. No puede destruir la imperio Romano, no puede.”	2	2	0	1
Base-CNsTipoB-Colectivo:	“No es cierto que el inmigrante de EE.UU. es una estrategia para entrar a Europa y atacarla desde dentro de los mil años, sino que los barbaros tienen razones para destruir el imperio romano, eso es lo que tiene que encontrar.”	2	2	0	1
Base-CNsTipoC-SinInfo:	“No es cierto que el imperio romano tiene ninguna razón en el mundo, pero eso no es lo que tiene.”	3	2	0	1
Base-CNsTipoC-Premisas:	“No es cierto que el imperio romano el islámico está en el país, y eso no es lo que es necesario.”	2	2	0	1

Tab. 5.6: Ejemplos generados por las distintas versiones de modelos Flan-T5 para tweets en inglés, junto a la evaluación manual realizada de cada una de las categorías propuestas: **Ofensividad**, **Posición**, **Informatividad** y **Felicidad**.

sobre todo, al utilizar un enfoque de aprendizaje en contexto. En este caso, la mejora experimentada al ajustar el modelo utilizando nuestro dataset es proporcionalmente mucho mayor. Mientras que de todos los modelos que utilizaron un enfoque de aprendizaje en contexto, sólo el XL sin información argumentativa logró un 10 % de contranarrativas aceptables y el resto tuvieron un 0 %, para los modelos con ajuste fino, el porcentaje de contranarrativas aceptables osciló entre un 10 % y un 30 %. La diferencia en el puntaje también es mucho mayor, no pudiendo superar los 3 puntos ninguno de los modelos que utilizaron aprendizaje en contexto, mientras que los modelos ajustados oscilan en valores de entre 6.40 y 7.40, una diferencia de 4.60 puntos entre el mejor de cada categoría, muy superior a lo observado para el inglés.

La razón de esta diferencia radica en que la gran mayoría de las contranarrativas generadas con aprendizaje en contexto son ofensivas y tienen una postura que apoya el mensaje de odio original. Como puede verse en la tabla 5.10 entre el 70 % y el 100 % de las contranarrativas generadas con esta configuración obtienen el valor más bajo en las categorías de Ofensividad y Posición, un porcentaje muy superior al de aquellas generadas en inglés. Los bajos valores en la categoría de Posición arrastran bajos valores también en la categoría de Informatividad, ya que imponen muchas evaluaciones en esta categoría con valor 0. En la tabla 5.8 podemos observar el impacto de cada categoría en el puntaje final. Si comparamos los resultados en inglés y en español obtenidos por cada configuración experimental, observamos que el modelo Base con *fewshot learning* muestra una mayor diferencia de puntaje entre ambos idiomas en la categoría “Ofensividad”, donde las variaciones en puntaje promedian 1.05 puntos de diferencia, mientras que en las categorías de “Posición” y “Felicidad” el promedio de la variación es de todos modos también similar, con 0.86 y 0.88. La “Informatividad” aporta la menor variación al puntaje total con una diferencia promedio de 0.40. Si realizamos la misma comparación entre los mismos idiomas pero para los modelos entrenados mediante el ajuste fino, observamos que la “Ofensividad” y la “Posición” tienen una variación en promedio de 0.28 y 0.29 respectivamente, demostrando que el ajuste fino reduce también la diferencia de rendimiento de un modelo entre distintos lenguajes (igualándolos para arriba).

Al analizar los ejemplos de la figura 5.6 observamos que todas las contranarrativas generadas por los modelos que utilizaron aprendizaje en contexto son iguales y en todas parece ocurrir una falla en la comprensión de la tarea requerida: en lugar de generar una respuesta contraria, realizan una traducción de algunos fragmentos al idioma inglés. Las contranarrativas generadas por modelos que utilizaron ajuste fino, si bien también son muy similares entre sí, tienen un grado mayor de éxito en cumplir con el requerimiento de generar un mensaje contrario al *tweet* original.

5.6.2 Tipos de contranarrativas

Si comparamos el rendimiento de las dos configuraciones utilizadas para ajustar los modelos (utilizando todas las contranarrativas vs. utilizando un tipo en particular), observamos que tanto en puntaje total como en porcentaje de respuestas Buenas y Aceptables, los modelos que utilizan todas las contranarrativas tienen un desempeño mejor que aquellos que utilizan un solo tipo específico. Este comportamiento es esperable, considerando que los primeros se entrenan con casi el triple de ejemplos y puede observarse, a su vez, de forma análoga, con el desempeño de los modelos entrenados con contranarrativas de tipo B en comparación con las de tipo A y C, las cuales tienen aproximadamente el doble de ejemplos que la primera (ver tabla 3.3). Lo que observamos a partir de revisar ejemplos

Tweet de odio:

nosotros no son mas los sudacas y centroamirdicanos y lo peor es que pasan por mi país violentando gente porque ‘‘‘no tienen dinero’’’ ‘‘‘porque van en busca de una vida mejor’’’ y ademas ilegal :link:

Respuesta humana:

No es cierto que los “sudacas” sean violentos. Estás siendo xenofobico.

Respuesta automática:

No es cierto que los sudacas y centroamirdicanos son violentos. No tienen dinero, y no tienen nada.

Fig. 5.3: Ejemplo automático generado por un modelo entrenado con contranarrativas de tipo B, comparado con la contranarrativa de tipo B original escrita por un anotador humano. En el mensaje de odio original se destaca el Colectivo anotado con color púrpura y la propiedad con color verde

es que en los casos en que el modelo cumplía la tarea que le fue pedida y otorgaba una contranarrativa Aceptable, esta en general tenía una estructura similar o análoga a la de la contranarrativa con la que el modelo fue ajustado. En el ejemplo de la figura 5.3 podemos observar una respuesta automática a un *tweet* de odio generada por un modelo entrenado con contranarrativas de tipo B e información argumentativa, comparada con la contranarrativa original escrita por un anotador humano. Si bien la contranarrativa generada por el modelo tiene errores de coherencia interna, como repetir el error ortográfico del *tweet* original, se puede apreciar que el mensaje se orienta a negar la relación entre el Colectivo mencionado y la Propiedad que se le busca atribuir, tal como estipula la estrategia de las contranarrativas de Tipo B.

5.6.3 Información argumentativa

Si consideramos tanto los modelos entrenados con aprendizaje en contexto como aquellos ajustados utilizando la totalidad del conjunto de datos de entrenamiento, incorporar la información argumentativa no representa una mejora cualitativa en el rendimiento de los modelos. Al comparar las respuestas generadas por estos, observamos que muchas veces ante el mismo mensaje de odio, los modelos que utilizaban para la generación distintas combinaciones de componentes argumentativos generaban la misma respuesta o respuestas similares, como puede observarse en los ejemplos de las figuras 5.5 y 5.6. Si bien en algunos casos encontramos cambios en la generación a partir de introducir la información argumentativa, comprobamos a través de un proceso de validación posterior a través de una muestra aleatoria, que en la mayoría de los casos los cambios no parecen responder de manera directa a la información incorporada. Tanto para el modelo base con *fewshot* o ajustado a todas las contranarrativas, como para el modelo XL con *fewshot*, los puntajes más altos son obtenidos por el modelo sin información argumentativa y el modelo que utiliza el Colectivo y la Propiedad, ambos en general con valores similares y siendo el primero

el que obtiene la mayor proporción de respuestas buenas y aceptables generadas mientras que el último obtiene el puntaje más alto de todo el experimento en su versión ajustada. En los tres casos, a su vez, el modelo que utiliza las premisas queda en tercer lugar y en último lugar, el modelo que utiliza toda la información junta. Una posible hipótesis que explica este comportamiento puede ser que en algunos casos, el exceso de información argumentativa agregada luego del mensaje de odio sature al modelo de lenguaje y le haga más difícil mantener el foco en su tarea original.

En los modelos que utilizan sólo un tipo particular de contranarrativa, el impacto de la información argumentativa relevante fue superior, especialmente para el inglés donde se observa una mejora en el desempeño tanto desde el punto de vista del puntaje, que aumenta en los tres casos entre .8 y .95 (entre un 20 % y un 25 % de incremento) como desde el punto de vista de la proporción de contranarrativas Buenas y Aceptables, que pasa de 30 % a 40 % y de 25 % a 40 % respectivamente para las de tipo A, de 0 % a 10 % en ambos casos para las de tipo B y de 10 % a 30 % y de 0 % a 20 % respectivamente para las de tipo C. En los ejemplos de la figura 5.5 podemos ver el impacto que tiene la información argumentativa para los modelos entrenados con contranarrativas de tipo C, que pasa de tener una postura ambigua respecto al *tweet* original a una respuesta con el máximo puntaje en las cuatro categorías (clasificada como Buena). Al analizar el texto generado, a su vez, observamos que la estructura de la contranarrativa generada se adapta a aquella propuesta por las contranarrativas de tipo C, ya que se enfoca en negar la justificación (“*india already reeling under constant threat of muslim radicals curb population*”) basándose en su tipo (“Hecho”), es decir, cuestionando la veracidad de la afirmación. Encontramos otros ejemplos similares donde comprobamos que a partir de incorporar la información argumentativa en los modelos entrenados con una estrategia única de contranarrativa, la respuesta generada parece no sólo mejorar sino también adaptarse mejor al tipo de contranarrativa propuesta. Para el español, observamos que la información argumentativa no tiene el mismo impacto en las contranarrativas generadas. Creemos que la razón en este caso, se debe al mal rendimiento en general de los modelos entrenados de este modo en este idioma. Si observamos la tabla 5.10 podemos ver que la Felicidad para estos casos, tiene cerca de un 100 % de respuestas generadas con el peor valor. Al ser incapaces de generar respuestas con un mínimo de coherencia interna, la posición respecto al mensaje original y la ofensividad muestran valores variables, predominando sobre todo los valores intermedios que reflejan respuestas que cambian de tema o son incoherentes.

5.6.4 Tamaño del modelo

En términos generales, el modelo tamaño XL no tiene un rendimiento superior que el modelo Base, comparando los resultados del enfoque con aprendizaje en contexto. Más aún, en términos de puntaje, la tabla 5.6.4 muestra que entre estos existe una diferencia de puntaje en favor del modelo Base que oscila entre 1,65 y 2,40 puntos para el idioma inglés, mientras que para el español la diferencia es en favor del modelo XL, oscilando entre 0,1 y 0,6 puntos de diferencia. En la tabla 5.8 podemos observar que esta diferencia proviene principalmente de las categorías de Ofensividad y Posición. Analizando las contranarrativas evaluadas encontramos que para el español, es frecuente en ambos modelos repetir de algún modo el mensaje original o parte de este, como se observa en los ejemplos de la figura 5.6, donde el modelo Base lo traduce de a partes al inglés, mientras que el modelo XL recorta una parte del principio y otra del final y reproduce sin cambios la parte del medio. Para el inglés, sin embargo, observamos que el modelo Base, en los casos que no

logra generar una contranarrativa aceptable, tiende a cambiar de tema, o dar respuestas ambiguas, a diferencia del modelo XL donde las contranarrativas son más elaboradas en términos de coherencia interna, pero tienen igualmente una tendencia a acordar con el mensaje que responden y ser ofensivas.

	Inglés			Español		
	Puntaje	% Acept	% Buenas	Puntaje	% Acept	% Buenas
Enfoque <i>Fewshot</i>						
Base	5.70	40 %	0 %	2.10	0 %	0 %
Base Toda	4.55	25 %	5 %	2.20	0 %	0 %
Base Colectivo	5.65	20 %	5 %	2.20	0 %	0 %
Base Premisas	5.30	35 %	5 %	1.90	0 %	0 %
XL	4.05	25 %	0 %	2.80	10 %	0 %
XL Toda	2.85	10 %	5 %	2.60	0 %	0 %
XL Colectivo	3.70	15 %	5 %	2.30	0 %	0 %
XL Premisas	2.90	0 %	0 %	2.40	0 %	0 %
Enfoques de Ajuste Fino						
Base	7.15	60 %	30 %	6.60	10 %	0 %
Base Todo	5.70	25 %	10 %	7.40	30 %	0 %
Base Colectivo	7.40	50 %	25 %	6.80	30 %	0 %
Base Premisas	6.55	40 %	30 %	6.40	20 %	0 %
Base CNs A	5.75	30 %	25 %	4.80	0 %	0 %
Base CNs A Premisas	6.55	40 %	40 %	3.50	0 %	0 %
Base CNs B	3.05	0 %	0 %	2.50	0 %	0 %
Base CNs B Colectivo	3.95	10 %	10 %	3.20	0 %	0 %
Base CNs C	4.65	10 %	0 %	5.00	20 %	0 %
Base CNs C Premisas	5.60	30 %	20 %	3.90	0 %	0 %

Tab. 5.7: Resultados de la evaluación manual de contranarrativas para *tweets* de odio generadas automáticamente en inglés y español, usando dos modelos Flan-T5 (el Base y el XL), dos técnicas de aprendizaje (*fewshot learning* y ajuste fino), distintos tipos de estrategias de contranarrativas (todas, sólo A, sólo B o sólo C) y distintas combinaciones de información argumentativa (sin información, con Colectivo y Propiedad, con Justificación y Conclusión y con toda la información disponible), de acuerdo a las configuraciones descritas en la sección 5.2. La tabla muestra el puntaje final agregado calculado tal como se describe en la sección 5.4.0.2 junto con el porcentaje de contranarrativas consideradas como Aceptables y Buenas. El detalle desglosado sobre los puntajes de cada categoría puede observarse en la tabla 5.8

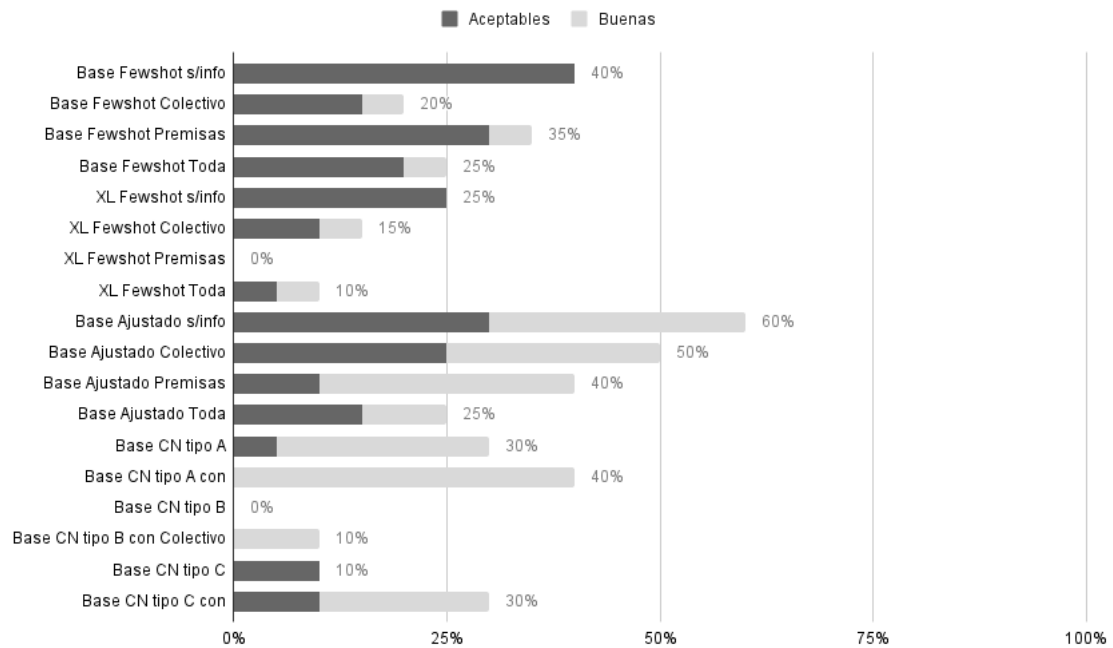


Fig. 5.4: Porcentaje de respuestas generadas consideradas como Aceptables (gris oscuro) y Buenas (gris oscuro + gris claro) para el idioma inglés

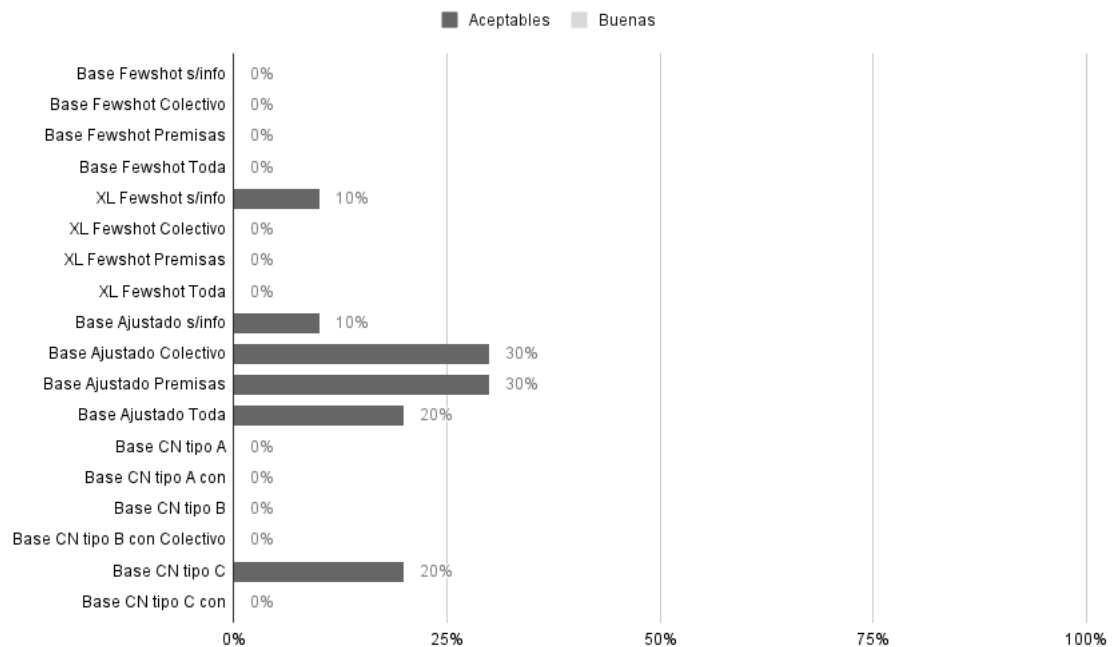


Fig. 5.5: Porcentaje de respuestas generadas consideradas como Aceptables (gris oscuro) y Buenas (gris oscuro + gris claro) para el idioma español.

	Inglés					Español				
	Of	Pos	Infor	Fel	Puntaje	Of	Pos	Infor	Fel	Puntaje
Enfoques <i>Fewshot</i>										
Base	2.50	2.25	0.40	2.55	5.70	1.30	1.30	0.00	1.50	2.10
Base Todo	1.95	1.80	0.45	2.35	4.55	1.30	1.20	0.00	1.70	2.20
Base Colectivo	2.45	2.05	0.30	2.85	5.65	1.20	1.20	0.00	1.80	2.20
Base Premisas	2.05	2.05	0.55	2.65	5.30	1.00	1.00	0.00	1.90	1.90
XL	1.65	1.65	0.40	2.35	4.05	1.40	1.40	0.10	1.90	2.80
XL Todo	1.30	1.30	0.15	2.10	2.85	1.40	1.30	0.00	1.90	2.60
XL Colectivo	1.70	1.60	0.25	2.15	3.70	1.10	1.10	0.00	2.10	2.30
XL Premisas	1.55	1.40	0.00	1.95	2.90	1.30	1.20	0.00	1.90	2.40
Enfoques de Ajuste Fino										
Base	2.55	2.55	1.40	2.65	7.15	2.60	2.60	1.70	1.70	6.60
Base Todo	2.30	2.15	0.60	2.65	5.70	2.90	2.80	1.60	2.10	7.40
Base Colectivo	2.55	2.60	1.45	2.80	7.40	2.70	2.70	1.40	2.00	6.80
Base Premisas	2.30	2.35	1.20	2.70	6.55	2.60	2.70	1.40	1.70	6.40
Base CNs A	2.35	2.25	0.90	2.25	5.75	2.50	2.30	0.60	1.40	4.80
Base CNs A Premisas	2.50	2.35	1.30	2.40	6.55	2.20	2.00	0.20	1.10	3.5
Base CNs B	1.90	1.80	0.05	1.30	3.05	1.70	1.80	0.00	1.00	2.50
Base CNs B Colectivo	2.15	2.15	0.35	1.30	3.95	2.00	2.00	0.20	1.00	3.20
Base CNs C	2.50	2.20	0.35	1.60	4.65	3.00	2.30	0.3	1.40	5.00
Base CNs C Premisas	2.20	2.25	0.85	2.30	5.60	2.20	2.10	0.30	1.30	3.90

Tab. 5.8: Puntaje asignado a las contranarrativas generadas en cada configuración experimental, desglosado por cada una de las categorías utilizadas para la evaluación. La **Ofensividad**, **Felicidad** y **Posición** tienen valores en un rango de 1 a 3 mientras que la **Informatividad** tiene valores en un rango de 0 a 3. El puntaje es el cálculo agregado tal como se explica en la sección 5.4.0.2 mostrado también en la tabla 5.6.4

Approaches	Ofensividad		Posición		Informatividad		Felicidad	
	Ofens	No-Of	Apoyo	Contra	Gen	Arg	Infel	Felic
Enfoque <i>Few-shot</i>								
Base	10 %	60 %	15 %	40 %	40 %	0 %	15 %	70 %
Base All	40 %	35 %	45 %	25 %	10 %	5 %	10 %	45 %
Base Collective	5 %	50 %	15 %	20 %	15 %	5 %	5 %	90 %
Base Premises	30 %	35 %	30 %	35 %	20 %	5 %	5 %	70 %
XL	60 %	25 %	60 %	25 %	10 %	0 %	10 %	45 %
XL All	80 %	10 %	80 %	10 %	0 %	10 %	5 %	15 %
XL Collective	55 %	25 %	55 %	15 %	10 %	5 %	20 %	0 %
XL Premises	55 %	10 %	60 %	0 %	0 %	0 %	30 %	25 %
Enfoque de Ajuste fino								
Base	10 %	65 %	10 %	65 %	25 %	35 %	15 %	80 %
Base All	15 %	45 %	15 %	30 %	0 %	10 %	15 %	80 %
Base Collective	0 %	55 %	0 %	60 %	0 %	35 %	5 %	85 %
Base Premises	10 %	40 %	10 %	45 %	0 %	30 %	10 %	80 %
Base CNs A	10 %	45 %	10 %	35 %	5 %	25 %	35 %	60 %
Base CNs A Premises	10 %	60 %	10 %	45 %	0 %	40 %	25 %	65 %
Base CNs B	30 %	20 %	25 %	5 %	5 %	0 %	80 %	10 %
Base CNs B Collective	0 %	15 %	0 %	15 %	5 %	10 %	85 %	15 %
Base CNs C	0 %	50 %	5 %	25 %	20 %	5 %	65 %	25 %
Base CNs C Justification	10 %	30 %	10 %	35 %	5 %	20 %	25 %	55 %

Tab. 5.9: Desglose por categorías de la evaluación manual de contranarrativas generadas automáticamente para *tweets* de odio en inglés, presentada en la tabla 5.6.4. La tabla muestra el porcentaje de *tweets* clasificados con el menor y el mayor valor respectivamente para cada categoría: Ofensividad, Posición, Informatividad y Felicidad.

Modelos	Ofensividad		Posición		Informatividad		Felicidad	
	Ofens	No-Of	Apoyo	Contra	Gen	Arg	Infel	Felic
Enfoque <i>Few-shot</i>								
Base	70 %	0 %	70 %	0 %	0 %	0 %	50 %	0 %
Base All	70 %	0 %	80 %	0 %	0 %	0 %	30 %	10 %
Base Collective	80 %	0 %	80 %	0 %	0 %	0 %	30 %	10 %
Base Premises	100 %	0 %	100 %	0 %	0 %	0 %	20 %	10 %
XL	70 %	10 %	70 %	10 %	10 %	0 %	20 %	10 %
XL All	70 %	10 %	70 %	0 %	0 %	0 %	10 %	0 %
XL Collective	90 %	0 %	90 %	0 %	0 %	0 %	0 %	10 %
XL Premises	80 %	10 %	80 %	0 %	0 %	0 %	10 %	0 %
Enfoque de Ajuste fino								
Base	20 %	80 %	20 %	80 %	0 %	10 %	40 %	10 %
Base All	0 %	90 %	0 %	80 %	0 %	0 %	20 %	30 %
Base Collective	0 %	70 %	0 %	70 %	0 %	0 %	30 %	30 %
Base Premises	0 %	60 %	0 %	70 %	0 %	0 %	50 %	20 %
Base CNs A	0 %	50 %	0 %	30 %	0 %	0 %	70 %	10 %
Base CNs A Premises	10 %	30 %	10 %	10 %	0 %	0 %	90 %	0 %
Base CNs B	30 %	0 %	20 %	0 %	0 %	0 %	100 %	0 %
Base CNs B Collective	10 %	10 %	10 %	10 %	0 %	0 %	100 %	0 %
Base CNs C	0 %	100 %	0 %	30 %	30 %	0 %	80 %	20 %
Base CNs C Justification	0 %	20 %	0 %	10 %	0 %	10 %	80 %	10 %

Tab. 5.10: Desgloce por categorías de la evaluación manual de contranarrativas generadas automáticamente para *tweets* de odio en español, presentada en la tabla 5.6.4. La tabla muestra el porcentaje de *tweets* clasificados con el menor y el mayor valor respectivamente para cada categoría: Ofensividad, Posición, Informatividad y Felicidad.

6 Evaluación automática de contranarrativas

La evaluación mediante anotaciones con humanos de las contranarrativas es una alternativa fiable para realizar evaluaciones de confianza respecto a las contranarrativas generadas por Grandes Modelos de Lenguaje. Sin embargo, esta metodología de evaluación no puede aplicarse a gran o incluso a mediana escala ya que requiere grandes cantidades de horas de trabajo de múltiples anotadores y, como ese trabajo es proporcional a la cantidad de modelos y contranarrativas que se quieran evaluar, esto limita en los hechos la posibilidad de expandir la evaluación a distintas variedades de modelos o configuraciones de generación. En el siguiente capítulo nos proponemos estudiar distintas alternativas de evaluación automática de contranarrativas utilizando distintas técnicas automáticas. Luego, generamos contranarrativas para todo el conjunto de datos de evaluación de CONEAS con múltiples modelos, de distintos tamaños, con distintas técnicas de aprendizaje e incluyendo distintos componentes argumentativos y evaluamos las contranarrativas generadas utilizando algunas de las técnicas de evaluación automática desarrolladas previamente.

6.1 Métricas clásicas utilizadas para evaluación automática de texto generado

Las limitaciones de las métricas clásicas utilizadas para la evaluación automática de generación de textos fueron desarrolladas en la sección 2.3. Nos proponemos, sin embargo, demostrar empíricamente estas limitaciones a partir de seleccionar algunos ejemplos de contranarrativas generados por nuestros modelos Flan-T5 y evaluarlos utilizando tres métricas frecuentes de evaluación de texto generado. Las métricas a utilizar se describen a continuación:

- **BLEU** (Bilingual Evaluation Understudy) [Papineni et al., 2002]: Calcula la precisión de coincidencias de n-gramas entre la salida del modelo y el texto de referencia, ajustando ambas para evitar repeticiones. Aunque BLEU es eficaz para evaluar la tarea de traducción automática, su aplicación a otras tareas de generación de texto es limitada, ya que penaliza variaciones léxicas y estructurales que no necesariamente cambian el significado de la oración y que pueden ser igualmente válidas. En tareas con múltiples respuestas correctas, como la que presentamos en este trabajo, BLEU tiende a infravalorar la calidad del texto generado, al no captar bien la diversidad semántica ni la creatividad en las respuestas.
- **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004]: es una familia de métricas comúnmente utilizada para evaluar la calidad de resúmenes automáticos. ROUGE-1 mide la coincidencia de unigrama entre el texto generado y el de referencia, capturando la superposición de palabras individuales. ROUGE-2, en

cambio, evalúa la coincidencia de bigramas, lo que permite medir mejor la coherencia entre pares de palabras, pero penaliza aún más las diferencias léxicas. ROUGE-L se basa en la Subsecuencia común más larga (LCS por su nombre en inglés, *Longest Common Subsequence*) y mide la similitud de secuencias de palabras en su orden correcto, capturando una estructura más amplia del contenido. Cada variante se calcula en términos de precisión, exhaustividad y una medida F para equilibrar ambas. ROUGE también es limitado en tareas con múltiples respuestas correctas, ya que se basa en coincidencias exactas de palabras y secuencias, ignorando la variación semántica que puede existir entre respuestas igualmente válidas. Esto puede llevar a puntajes bajos en textos generados que expresan el mismo significado de manera diferente, penalizando la diversidad y creatividad en la generación de respuestas.

- **Similitud Coseno entre Sentence Embeddings:** Los embeddings de oraciones mapean oraciones a representaciones vectoriales fijas (en este caso de 384 dimensiones) que permiten medir la similitud semántica a través del cálculo de la distancia entre ambos, por ejemplo a través del cálculo del coseno. Para este trabajo utilizamos el modelo *all-MiniLM-L6-v2* de SBERT [Reimers and Gurevych, 2019] que se basa en el modelo pre-entrenado MiniLM [Wang et al., 2020] al cual se le realiza una etapa más de pre-entrenamiento utilizando como tarea pretexto una tarea supervisada de similaridad semántica que consiste en, dada una oración, encontrar otra similar, asociada a la primera, de entre una lista de varias otras oraciones seleccionadas al azar. El modelo “all”, utiliza una combinación de varios datasets con más de mil millones de ejemplos. La similaridad coseno utilizando embeddings de oraciones permite mayor flexibilidad respecto a respuestas con distinto léxico o incluso estructura sintáctica, ya que captura la relación semántica más allá de coincidencias exactas de palabras. Sin embargo, este modelo muestra también limitaciones al comparar dos oraciones potencialmente correctas pero con distinto contenido semántico.

Para calcular estas métricas, generamos contranarrativas para el total de la partición de evaluación de nuestro dataset utilizando los modelos descritos en la sección 5.2. Cada contranarrativa generada fue comparada contra todas las contranarrativas anotadas en el dataset para un mismo *tweet* de odio. Luego, calculamos el promedio del puntaje obtenido por cada par evaluado en cada configuración experimental. Comprobamos, de esta manera, que la métrica BLEU da, en la mayoría de los casos, siempre 0 para todas las contranarrativas generadas. En los pocos casos que encontramos valores de BLEU distintos a 0, estos no superan el valor 0.02 (considerado muy bajo). Las distintas variantes de Rouge oscilan entre .0 y .25 sin que pueda distinguirse un criterio aparente más allá de la repetición circunstancial de alguna palabra, pero sin tener relación con la semántica de las respuestas comparadas. Rouge2 tiende a tener un valor aún más bajo debido a que evalúa la coincidencia de bigramas en lugar de unigramas. La similaridad coseno de los embeddings de oración es muy variable en su rango de valores, pero en general tampoco parece tener una correlación con la calidad de la respuesta sino, en todo caso, con la utilización de palabras o frases similares (a veces, incluso a pesar de tener significados opuestos, obtienen puntajes de similaridad altos).

Las métricas descritas, utilizadas en los ejemplos analizados, no consideran el mensaje de odio original. Sólo comparan la respuesta generada contra la respuesta escrita por el anotador humano. Por este motivo, no tienen forma de calcular un valor de “calidad” de la respuesta generada en base a lo que se está respondiendo, sino que lo calculan en base

TWEET: @user must deport all illegal migrants india already reeling under constant threat of muslim radicals curb population

GOLD STANDARD: I don't think India has it really bad. I don't see how a few migrants could make a country "reel"...

GENERADO: I don't understand what you are trying to say. Besides, it's not true that Muslims are radicals.

BLEU	Rouge1	Rouge2	RougeL	SBERT
0	.18	.10	.18	.14

Fig. 6.1: Métricas automáticas utilizadas para evaluar la generación de texto aplicadas a la evaluación de contranarrativas. En este caso, una buena contranarrativa obtiene un puntaje bajo según cualquiera de las métricas consideradas.

a lo que una persona respondió previamente.

El ejemplo de la figura 6.1 muestra un mensaje de odio junto con la respuesta escrita por un anotador humano y la respuesta generada por uno de los modelos Flan-T5. Observamos que la respuesta generada no es ofensiva, tiene una postura contraria al mensaje que responde, tiene buena felicidad e incluso cierto grado de especificidad (ya que identifica al colectivo y la propiedad que se le endilga y niega su relación). Sin embargo, obtiene un puntaje bajo en todas las métricas utilizadas para la evaluación. Si comparamos la respuesta generada con el *gold standard* podemos ver que la superposición de n-gramas es prácticamente nula, mientras que semánticamente, si bien ambas parecen ser buenas respuestas, la generada se centra en negar la condición de radicales asignada a los musulmanes, mientras que el *gold standard* se centra en relativizar el problema de India respecto a la inmigración.

El ejemplo de la figura 6.2 muestra un mensaje de odio, la respuesta generada por un modelo Flan-T5 y las tres respuestas anotadas por humanos que constituyen el *gold standard*. Si bien los resultados para BLEU y Rouge son igual de bajos que en casi todos los casos, podemos observar que el valor de la similaridad coseno es alto considerando cualquiera de las tres respuestas contra las que se compara (en los tres casos los valores están entre el 20 % más alto obtenido). Este ejemplo ilustra que la similaridad coseno falla en captar elementos esenciales de la comparación de dos contranarrativas, a saber, si tienen la misma postura respecto al tema al que se refieren (en este caso, la inmigración).

El ejemplo de la figura 6.3 nuevamente encontramos una respuesta generada que reproduce el odio y avala el mensaje odioso que se supone debe responder. Sin embargo, nuevamente obtiene un puntaje alto de similaridad coseno con SBERT debido a la cantidad de elementos en común que tiene con la respuesta del *gold standard*, como "crimen" y "construir un muro", a pesar de que las respuestas comparadas tienen posturas totalmente contrarias respecto a estos temas.

Finalmente, el ejemplo de la figura 6.4 muestra uno de los puntajes más bajos de SBERT obtenidos en todo el dataset, a pesar de que la respuesta generada es aceptable. Nuevamente, el anotador humano eligió un eje distinto para refutar el mensaje de odio, basándose en negar el aspecto negativo de la propiedad que se le endilga a los inmigrantes (la de ser "perros"), mientras que la respuesta generada pide pruebas acerca de la justifica-

TWEET: the real truth is after cologne and in the nordic countries and others no one trusts any refugees a better life for them doesnt mean 1 4 u

GENERADO: I don't think anyone trusts refugees a better life for them. It's just a lie.

GOLD STANDARD: Do you have any evidence that in Cologne and in the Nordic countries people's quality of life is worst because of immigrants?

BLEU	Rouge1	Rouge2	RougeL	SBERT
0	.09	.00	.05	.38

GOLD STANDARD: Refugees are hard working people like anyone else that went to a hard situation and were forced to migrate. Helping them means a small effort that doesn't change our quality of life and shows that we are still humans

BLEU	Rouge1	Rouge2	RougeL	SBERT
0	.23	.00	.13	.57

GOLD STANDARD: Do you have any proof that in the Nordic countries and Others no one trust any refugees? You shouldn't speak for others

BLEU	Rouge1	Rouge2	RougeL	SBERT
0	.14	.00	.09	.60

Fig. 6.2: Métricas automáticas utilizadas para evaluar la generación de texto aplicadas a la evaluación de contranarrativas. La misma contranarrativa generada obtiene valores distintos segun en relación a que otra contranarrativa del *gold standard* se la compare.

TWEET: @user @user string him up to show the message that you will pay for your crimes in a country where you are not supposed to be in hashtag buildthatwall we are losing our legal citizens to those who should not be here in the first place if they werent here we would still have our loved ones

GOLD STANDARD: What crimes? Why "they shouldn't be here"? All this is an invention to justify the building of a useless wall.

GENERADO: I don't think you're going to pay for your crimes if you don't have a wall. It's just a joke.

BLEU	Rouge1	Rouge2	RougeL	SBERT
0	.21	.00	.17	.51

Fig. 6.3: Métricas automáticas utilizadas para evaluar la generación de texto aplicadas a la evaluación de contranarrativas. En este caso, el texto generado apoya el mensaje de odio y es ofensivo, sin embargo obtiene un puntaje de SBERT alto.

ción del argumento odiante. Vale destacar que los puntajes de Rouge1 y RougeL también se encuentran en la escala más baja de valores.

El análisis de estos casos seleccionados de manera aleatoria muestra que, "si bien las evaluaciones manuales son más costosas y subjetivas, los métodos automáticos actualmen-

TWEET: @user dogs in america or dogs in mexico ?? a mexican asked me whether i was saying that hispanic immigrants had no right to a country meaning ours i replied “immigrants have a country or they wouldnt be immigrants”

GOLD STANDARD: I never understood why “dogs” as an insult. I mean, not that man’s best friend?

GENERADO: Do you have any proof that hispanic immigrants have no right to a country? I don’t think so.

BLEU	Rouge1	Rouge2	RougeL	SBERT
0	.11	.00	.5	-.04

Fig. 6.4: Métricas automáticas utilizadas para evaluar la generación de texto aplicadas a la evaluación de contranarrativas.

te existentes fallan en proveer una comprensión significativa o en capturar las relaciones matizadas entre el texto generado y la percepción humana del mismo, desestimando aspectos cruciales como la relevancia o efectividad” [Ni’mah et al., 2023].

6.2 Evaluación automática de contranarrativas utilizando modelos de lenguaje ajustados

Considerando que la evaluación manual de contranarrativas no escala a grandes volúmenes de datos y que las métricas automáticas tradicionales fallan en captar aspectos esenciales del significado, es necesario buscar un método de evaluación que tenga mayor fiabilidad y que a la vez pueda aplicarse a una cantidad incremental de información generada por modelos automáticos. Entendiendo que es difícil que un método completamente automatizado no tenga algunas falencias, consideramos que un aspecto fundamental es garantizar en la mayor medida posible, la interpretabilidad de la evaluación generada, de la misma manera que lo hicimos con la evaluación manual propuesta en la sección 5.4. Proponemos entonces, tomar las categorías propuestas en esa sección y los 540 ejemplos evaluados manualmente y entrenar con ellos cuatro clasificadores (uno para cada categoría de análisis: Ofensividad, Posición, Informatividad y Felicidad) que realicen la tarea de evaluación del mismo modo que fue realizado en la sección 5.6.

6.2.1 Partición del conjunto de datos

En primer lugar, dividimos los ejemplos evaluados por anotadores humanos en cada idioma en tres particiones de entrenamiento, desarrollo y evaluación, buscando una proporción deseable de 80 %, 10 % y 10 % respectivamente, pero de tal forma que ningún mensaje de odio aparezca en más de una partición con respuestas generadas distintas. La tabla 6.1 muestra la cantidad de ejemplos final de cada partición para el inglés y el español y el porcentaje que cada partición representa respecto del total de ejemplos anotados (360

	Entrenamiento	Entrenamiento Ext	Desarrollo	Evaluación
English	298 (83 %)	378 (105 %)	22 (6 %)	40 (11 %)
Spanish	126 (70 %)	170 (94 %)	18 (10 %)	36 (20 %)

Tab. 6.1: Distribución y cantidad de ejemplos de cada partición del dataset de pares de discurso de odio y contranarrativas, evaluadas con las categorías de ofensividad, posición, informatividad y felicidad. La columna “Entrenamiento Ext” muestra la cantidad de ejemplos y porcentaje respecto al conjunto de datos original del conjunto de entrenamiento extendido con contranarrativas generadas por humanos, evaluadas automáticamente con el máximo puntaje en cada una de las categorías de análisis.

	English				Spanish			
	Of	Pos	Inf	Fel	Of	Pos	Inf	Fel
1	24 %	25 %	33 %	26 %	39 %	39 %	10 %	47 %
2	39 %	45 %	24 %	21 %	31 %	39 %	85 %	43 %
3	37 %	30 %	43 %	53 %	30 %	22 %	5 %	10 %

Tab. 6.2: Distribución de las clases según cada categoría de evaluación para los conjuntos de datos de entrenamiento sin extender, en inglés y en español.

en inglés y 180 en español). La tabla 6.2 muestra la distribución por clases de cada categoría de evaluación para ambos idiomas. Podemos observar que para el idioma español y principalmente para las categorías de Informatividad y Felicidad existe un desbalanceo en desmedro del valor “3”, que es minoritario. Considerando que, además, los datasets son pequeños como para ser utilizados para entrenar un modelo de lenguaje que sirva para realizar una tarea tan compleja como evaluar contranarrativas, generamos un segundo conjunto de datos de entrenamiento a partir de expandir la partición de entrenamiento original agregando las contranarrativas generadas por humanos de los mensajes de odio que ya se encontraban presentes en esta. Estos nuevos pares de discurso de odio y contranarrativa fueron etiquetados de manera automática con el máximo puntaje en cada categoría, bajo la hipótesis de que al ser generadas por humanos, las contranarrativas deben ser de la mayor calidad. La cantidad de ejemplos de este nuevo conjunto y la proporción respecto al tamaño del dataset original pueden verse en la columna “Entrenamiento Ext” de la tabla 6.1. La nueva distribución de cada clase de este nuevo conjunto de entrenamiento puede verse en la tabla 6.3.

Para entrenar los modelos para predecir informatividad filtramos aquellos ejemplos evaluados con 0, dado que este valor depende enteramente de la evaluación realizada de la categoría “Posición” (si el valor de “Posición” no es 3, entonces se asigna 0 automáticamente como valor de “Informatividad”). Buscamos de este modo, desarrollar un sistema neurosimbólico de evaluación de contranarrativas que combine el aprendizaje mediante redes neuronales con reglas simbólicas que establezcan relaciones entre las categorías de evaluación como la mencionada.

Sin embargo, al solo considerar, para el entrenamiento de los modelos que predicen “Informatividad”, los ejemplos con valores distintos de 0, el tamaño de los datos utilizados se ve reducido considerablemente. Para el inglés, solo 104 ejemplos tienen valor de informatividad mayor a 0 (el 29 %) mientras que para el español, solo 40 (el 22 %). Considerando, además, que existe un fuerte desbalanceo de datos, especialmente en el español, donde además la proporción de ejemplos utilizados es menor, observamos un problema serio de

	English				Spanish			
	Of	Pos	Inf	Fel	Of	Pos	Inf	Fel
1	19 %	20 %	17 %	20 %	28 %	28 %	4 %	32 %
2	30 %	35 %	14 %	17 %	24 %	28 %	35 %	36 %
3	51 %	45 %	69 %	63 %	48 %	44 %	61 %	32 %

Tab. 6.3: Distribución de las clases según cada categoría de evaluación para los conjuntos de datos de entrenamiento extendidos con las contranarrativas anotadas por humanos, en inglés y en español

escasez de datos para algunas clases: la clase “1”, por ejemplo, cuenta solo con 4 ejemplos anotados, mientras que la clase “3” solo tiene 2 ejemplos.

6.2.2 Modelos utilizados

Utilizamos los siguientes modelos base para ser ajustados utilizando nuestro *dataset*:

- **RoBERTa** [Liu et al., 2019]: este modelo fue previamente descrito en la sección 4.1.1. En esa ocasión utilizamos la versión correspondiente a la tarea de clasificación de *tokens*, ajustado al reconocimiento de componentes argumentativos. Para este experimento, utilizamos la versión adaptada a clasificación de secuencias. A su vez, utilizamos dos tamaños distintos del modelo: la versión *Base* y la versión *Large*.
- **XLM-RoBERTa** [Conneau et al., 2019]: este modelo también fue previamente descrito en la sección 4.1.3. Nuevamente, en esa ocasión utilizamos la versión correspondiente a la tarea de clasificación de *tokens* mientras que en este caso utilizamos la versión para clasificar secuencias y también en esta ocasión utilizamos la versión *Base* y la versión *Large*.
- **BETO** [Cañete et al., 2020]: es un modelo con una arquitectura idéntica y un entrenamiento similar al del modelo BERT, pero entrenado con textos exclusivamente en idioma español. Su entrenamiento se llevó a cabo utilizando un gran corpus de datos, que incluye fuentes como el contenido de Wikipedia en español, el proyecto OPUS (una colección de textos paralelos y multilingües), y otros textos de dominio general que abarcan una variedad de temas y estilos en el idioma. El tamaño del dataset recolectado es comparable al utilizado para entrenar BERT. BETO se convirtió en el estado del arte para varias tareas en español, superando a modelos multilingües como mBERT. Al estar exclusivamente entrenado en español, BETO logra una comprensión más precisa de las expresiones idiomáticas, conjugaciones y otros matices propios del idioma.
- **Bertín** [la Rosa et al., 2022]: también es un modelo con una arquitectura idéntica a la del modelo BERT entrenado desde cero con texto exclusivamente en idioma español. Sin embargo, a diferencia de BETO, Bertín utiliza un enfoque basado en realizar el pre-entrenamiento con menos cantidad de datos de mayor calidad. De este modo, prioriza la eficiencia en cuanto a la utilización de recursos y la posibilidad de entrenar en entornos de menor presupuesto. Bertín se entrena utilizando la porción en español del dataset mC4, la versión multilingüal del dataset C4, que a su vez es una versión sanitizada del *dataset Common Crawl*. Su entrenamiento se basa en el

propuesto para el modelo RoBERTa, pero utilizando una técnica de *sampling* basado en la perplejidad que permite obtener resultados similares con aproximadamente un quinto de los datos y la mitad de los pasos de entrenamiento.

- **Sentence Transformers:** Los *sentence transformers* son modelos específicos basados en transformers que generan representaciones vectoriales a nivel de oración, diseñadas para capturar el significado completo de una frase o sentencia en un espacio semántico continuo. A diferencia de otros modelos *transformer* convencionales, que generan *embeddings* independientes para cada palabra y requieren, luego, métodos adicionales para combinarlos cuando se necesita una representación a nivel de oración, los *sentence transformers* están adaptados para generar directamente un único vector que representa la oración completa. Esto se logra mediante un entrenamiento en diversas tareas pretexto como similitud semántica y emparejamiento de oraciones, lo cual permite que los *sentence transformers* capturen relaciones semánticas más profundas y sean efectivos en tareas como clasificación de oraciones. Existen distintos modelos de *sentence transformer* que se diferencian según distintos factores como el modelo pre-entrenado que utilizan como base, la tarea o tareas pretexto para las que fueron ajustados y la mono o multilingüidad. Para nuestros experimentos en inglés, utilizamos seis modelos distintos que combinan tres modelos utilizados como base, tres tipos de tareas pretexto y mono y multilingüidad. Para los experimentos en español, utilizamos tres modelos que combinan, también, tres modelos base pero se basan en la misma tarea pretexto y son en los tres casos, multilingües. Dos de estos modelos son RoBERTa y XLM-RoBERTa, que explicamos previamente en esta misma sección. Los otros dos, los explicamos a continuación.

1. **MPNET** [Song et al., 2020]: Es un modelo de *transformer* desarrollado para mejorar el rendimiento en tareas de comprensión del lenguaje, combinando estrategias de preentrenamiento de enmascaramiento y permutación. En lugar de solo enmascarar palabras específicas en la entrada como en BERT, MPNet utiliza una técnica híbrida que enmascara y permuta subsecuencias enteras de tokens. Esto le permite al modelo aprender relaciones entre las palabras en distintos contextos de una manera más compleja y robusta, capturando tanto dependencias locales como globales. MPNet emplea un enfoque de ordenamiento de *tokens* a través del cual, intenta predecir no solo las palabras enmascaradas, sino también su orden natural en la oración. Esta característica es clave para evitar la desconexión que a veces ocurre en otros modelos cuando el contexto de las palabras se fragmenta demasiado.
2. **MiniLM** [Wang et al., 2020]: Es un modelo basado en *transformers* pero de tamaño reducido, pre-entrenado a partir de la destilación de un modelo más grande que funciona como modelo maestro y destilando particularmente la última capa de auto-atención de este. De este modo, el modelo estudiante no solo demuestra poder adaptarse mejor a contextos de recursos limitados, sino que retiene más del 99 % de *accuracy* para casi todas las tareas de GLUE, utilizando un 50 % menos de parámetros que en la arquitectura clásica de transformers.

A su vez, las siguientes combinaciones de conjuntos de datos fueron utilizados para ajustar los modelos pre-entrenados:

1. Similitud semántica (**all**): Se utiliza un conjunto de varios datasets de distintas tareas que implican similitud semántica entre los datos, como por ejemplo, título y abstract/contenido, preguntas y respuestas, parafraseo, preguntas duplicadas, entre otras. En total suman más de 1 billón de pares de oraciones similares en términos semánticos. El entrenamiento se realiza con un objetivo contrastativo en el cual el modelo toma como entrada una oración y una lista que contiene el par correspondiente y un conjunto de otras oraciones seleccionadas al azar y debe elegir entre esa lista la oración correspondiente. Este tipo de pre-entrenamiento sirve para desarrollar *embeddings* de propósito general que luego pueden ajustarse a una multiplicidad de tareas.
2. Búsqueda semántica (**qa**): Los modelos se entrenan con un subconjunto de los *datasets* del caso anterior, con más de 215 millones de ejemplos compuestos por preguntas sobre una amplia variedad de temas y sus respuestas. El objetivo del entrenamiento es que la representación vectorial de la pregunta (o *query*) sea cercana al fragmento conteniendo su respuesta.
3. Parafraseo (**paraphrase**): Los modelos se entrenan también con un subconjunto de los *datasets* del primer caso: aquellos que emparejan pares de oraciones que son parafrasis entre sí. El objetivo del entrenamiento es obtener representaciones vectoriales independientes de las variaciones léxicas que representen equivalencias semánticas en el mismo espacio vectorial.

Los modelos utilizados para evaluar pares en inglés son:

- **all-mpnet-base-v2**: MPNET ajustado con todos los conjuntos de datos de similitud semántica en inglés.
- **multi-qa-mpnet-base-v2**: MPNET ajustado con todos los conjuntos de datos de búsqueda semántica (preguntas y respuestas) en más de 50 idiomas.
- **paraphrase-multi-mpnet-base-v2**: MPNET ajustado con los conjuntos de datos de parafraseo en más de 50 idiomas.
- **paraphrase-multi-MiniLM-L12-v2**: MiniLM de 12 capas ajustado con los conjuntos de datos de parafraseo en más de 50 idiomas.
- **all-MiniLM-L12-v2**: MiniLM de 12 capas ajustado con todos los conjuntos de datos de similitud semántica en inglés.
- **all-roberta-large-v1**: Roberta Large ajustado con todas las tareas de similitud semántica en inglés.

Los modelos utilizados para evaluar pares en español son:

- **paraphrase-xlm-r-multi-v1**: XLM-Roberta ajustado con los conjuntos de datos de parafraseo en más de 50 idiomas.
- **paraphrase-multi-MiniLM-L12-v2**: MiniLM de 12 capas ajustado con los conjuntos de datos de parafraseo en más de 50 idiomas.
- **paraphrase-multi-mpnet-base-v2**: MPNET ajustado con los conjuntos de datos de parafraseo en más de 50 idiomas.

Los modelos de *sentence embeddings* utilizados fueron seleccionados a partir de la lista de modelos con mejor rendimiento, de acuerdo a lo reportado por sus propios autores¹.

6.2.3 Resultados

Para medir la factibilidad de realizar evaluaciones de manera automática de acuerdo a las cuatro categorías de evaluación propuestas, utilizamos para cada uno de los modelos base, cuatro versiones, una para cada categoría de evaluación, que toman una entrada de texto de longitud variable y predicen una única etiqueta con tres valores posibles (1, 2 o 3). Entrenamos estos modelos con un *learning rate* de 2e-05, un tamaño de *batch* de 4 y durante 5 épocas. La entrada de los modelos está compuesta de un par de discurso de odio y contranarrativa, separados entre sí por un *token* especial que suele utilizarse en los modelos BERT para separar oraciones: “[SEP]”. Cada modelo predice únicamente el valor de una sola categoría e ignora la información sobre las otras tres.

Las tablas 6.5 y 6.7 muestran los resultados del entrenamiento utilizando el conjunto de datos que tiene únicamente las contranarrativas generadas por los modelos Flan-T5 descritos en la sección 5.6, evaluadas luego por anotadores humanos, en inglés y español, respectivamente. Las tablas 6.4 y 6.6 muestran los resultados del entrenamiento utilizando el mismo conjunto de datos del caso anterior pero expandido con contranarrativas generadas por anotadores humanos y evaluadas automáticamente con la máxima puntuación en las cuatro categorías, bajo la asunción de que los anotadores humanos producen siempre buenas contranarrativas, también en inglés y en español, respectivamente. Esta expansión de los conjuntos de datos se realizó sólo en las particiones destinadas a entrenamiento (las particiones de desarrollo y evaluación se mantuvieron igual). Su objetivo es aumentar el tamaño del dataset e intentar superar el desbalanceo de datos, especialmente en aquellas categorías donde la clase de máxima puntuación es minoritaria (como la Informatividad y la Felicidad para el idioma español).

Los resultados muestran que utilizar el conjunto de datos aumentado con contranarrativas escritas por humanos mejora la evaluación automática. Para el inglés, el valor del F1 sube en promedio 13,7 puntos y en el mejor valor, 10 puntos para la categoría Ofensividad; en promedio 23,1 puntos y en el mejor valor 13 puntos para la categoría Posición; en promedio 6,7 puntos y en el mejor valor 0,1 punto para la categoría Informatividad; y en promedio empeora 1,6 puntos para la categoría Felicidad pero sube 3 puntos el mejor valor. Para el español, en la categoría de Ofensividad, el promedio del F1 empeora 2,8 puntos mientras que el mejor valor empeora en 3 puntos; para la categoría Posición, el promedio del F1 mejora 17,6 puntos mientras que el mejor valor mejora en 22 puntos; para la categoría de Felicidad, el promedio del F1 mejora en 52,8 puntos mientras que el mejor valor se incrementa en 71 puntos, demostrando el impacto negativo que tenía el desbalance de datos de la clase de mayor valor y la efectividad del mecanismo de aumento de datos a través de contranarrativas elaboradas por humanos realizado; para la categoría Felicidad, el promedio del F1 mejora en 21,4 puntos mientras que el mejor valor se incrementa en 26 puntos.

En cuanto a los modelos, para el inglés en su versión extendida, el modelo de *sentence transformers* ajustado con todos los conjuntos de datos únicamente en inglés y utilizando como base el modelo MPNET muestra los mejores resultados en Posición, mientras que

¹ https://www.sbert.net/docs/sentence_transformer/pretrained_models.html

el modelo de *sentence transformers* ajustado con datos de parafraseo multilinguales utilizando como base también el modelo MPNET, es el que muestra mejores resultados en las categorías de Ofensividad, Informatividad y Felicidad, aunque para la categoría Posición también obtiene un valor cercano al mejor. Para el idioma inglés utilizando el conjunto de datos sin extender, el modelo de *sentence transformers* ajustado con todos los datos en inglés que utiliza como base al modelo Roberta Large obtiene los mejores resultados en Ofensividad y Posición, mientras que el modelo de *sentence transformers* ajustado con datos de búsqueda semántica multilinguales que utiliza un modelo base MPNET obtiene los mejores resultados para Informatividad y el modelo de *sentence transformers* ajustado con todos los datos con un modelo base MPNET obtiene el mejor valor para la categoría de Felicidad.

En el caso del español, para el conjunto de datos aumentado, el modelo BETO obtiene los mejores resultados para las categorías de Ofensividad y Felicidad, mientras que Bertin obtiene los mejores resultados para Posición e Informatividad, en esta última categoría, empatado con el modelo de *sentence transformer* ajustado con datos de parafraseo multilinguales y utilizando como base un modelo XLM-RoBERTa. Para el conjunto de datos sin aumentar, el modelo BETO empata en rendimiento con el modelo de *sentence transformers* ajustado a datos de parafraseo multilinguales basado en un modelo MPNET en las categorías de Ofensividad y Felicidad, mientras que Bertin obtiene los mejores resultados en la categoría Posición. Para este conjunto de datos, todos los modelos obtienen el mismo rendimiento bajo en la categoría de Informatividad. Analizando las predicciones realizadas, observamos que todos los modelos muestran el mismo comportamiento en esta categoría: predicen siempre la categoría mayoritaria.

Al comparar los resultados obtenidos para cada idioma, observamos que para la categoría de Ofensividad, el rendimiento es similar en ambos si se utiliza el conjunto de datos sin expandir, pero que al utilizar el conjunto de entrenamiento expandido, los modelos mejoran notablemente el rendimiento para el inglés mientras que para el español, incluso, se reduce un poco. Con las categorías de Informatividad y Felicidad sucede el caso contrario. Mientras que con el *dataset* no aumentado, el rendimiento de los modelos para el inglés es mayor que para el español, con los *datasets* aumentados el rendimiento para el español aumenta ostensiblemente (especialmente en la categoría de Informatividad) mientras que para el inglés se mantiene en un nivel similar o muestra pequeñas mejoras, por lo cual, el rendimiento de los modelos para el español termina siendo superior en estas categorías que para el inglés. Para la categoría de Posición observamos que para ambos idiomas los modelos ajustados con el conjunto de datos sin aumentar tienen un rendimiento aceptable y similar y aumentan también en una proporción similar al utilizar los conjuntos de datos aumentados.

Considerando que los conjuntos de datos utilizados son verdaderamente pequeños y que la tarea a realizar tiene un alto grado de subjetividad, con niveles de acuerdo entre anotadores moderados o justos, tal como puede observarse en la tabla 5.2, entendemos que los resultados obtenidos por los modelos de lenguaje son muy buenos (sobre todo al considerar aquellos entrenados con la versión extendida del conjunto de entrenamiento). Particularmente, concluimos que la técnica de aumento de datos con contranarrativas generadas por seres humanos etiquetadas automáticamente con el mayor valor en cada categoría tiene un gran impacto positivo en el rendimiento de los modelos.

	Ofensivo			Posición			Informatividad			Felicidad		
	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc
roberta-base	.68	.70	.68	.79	.83	.76	.47	.46	.48	.75	.79	.72
roberta-large	.62	.68	.60	.75	.77	.74	.30	.28	.33	.76	.84	.73
XLM-roberta-base	.70	.71	.69	.75	.77	.73	.30	.28	.33	.56	.54	.59
XLM-roberta-large	.63	.65	.62	.78	.80	.76	.30	.28	.33	.74	.80	.72
ST-all-mpnet-base-v2	.67	.71	.64	.82	.84	.81	.47	.46	.48	.77	.79	.76
ST-multi-qa-mpnet-base-dot-v1	.66	.69	.64	.77	.79	.76	.53	.63	.48	.75	.79	.72
ST-paraphrase-multi-mpnet-base-v2	.71	.71	.71	.79	.80	.78	.54	.63	.50	.83	.87	.80
ST-paraphrase-multi-MiniLM-L12-v2	.68	.71	.68	.69	.75	.66	.30	.28	.33	.42	.47	.43
ST-all-MiniLM-L12-v2	.61	.64	.60	.77	.79	.76	.42	.40	.46	.80	.86	.77
ST-all-roberta-large-v1	.69	.72	.67	.81	.84	.80	.45	.46	.44	.73	.79	.69

Tab. 6.4: Resultados obtenidos por los modelos para predecir las cuatro categorías propuestas de evaluación en el conjunto de datos de evaluación en inglés, extendido con contranarrativas generadas por humanos anotadas con el máximo valor en todas las categorías

	Ofensivo			Posición			Informatividad			Felicidad		
	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc
roberta-base	.53	.57	.51	.54	.57	.54	.16	.22	.12	.75	.79	.72
roberta-large	.40	.46	.38	.57	.62	.59	.30	.34	.31	.58	.61	.60
FacebookAI-xlm-roberta-base	.50	.49	.53	.19	.14	.33	.42	.40	.46	.76	.84	.73
FacebookAI-xlm-roberta-large	.56	.57	.55	.48	.48	.53	.42	.40	.46	.67	.68	.67
ST-all-mpnet-base-v2	.56	.60	.54	.63	.64	.66	.28	.27	.29	.80	.86	.77
ST-multi-qa-mpnet-base-dot-v1	.53	.57	.51	.65	.70	.64	.53	.63	.48	.76	.84	.73
ST-paraphrase-multi-mpnet-base-v2	.51	.51	.51	.60	.61	.62	.42	.40	.46	.76	.84	.73
ST-paraphrase-multi-MiniLM-L12-v2	.54	.59	.54	.53	.51	.57	.30	.28	.33	.72	.74	.71
ST-all-MiniLM-L12-v2	.54	.59	.54	.53	.51	.57	.30	.28	.33	.72	.74	.71
ST-all-roberta-large-v1	.61	.64	.59	.69	.73	.68	.28	.27	.29	.75	.79	.72

Tab. 6.5: Resultados obtenidos por los modelos para predecir las cuatro categorías propuestas de evaluación en el conjunto de datos de evaluación en inglés, sin extenderlo con contranarrativas generadas por humanos

6.3 Evaluación automática de contranarrativas generadas por Grandes Modelos de Lenguaje

El buen rendimiento de los modelos que realizan la evaluación automática según las cuatro categorías propuestas nos abre la posibilidad de masificar la evaluación de contranarrativas. Una evaluación automática, a diferencia de una evaluación manual, puede extenderse a grandes cantidades de datos generados por múltiples modelos. Esto nos brinda una nueva herramienta para intentar responder nuestras preguntas experimentales desahrolladas en la sección 5.2, pero generalizadas a varios modelos.

Con este fin, generamos contranarrativas utilizando cuatro modelos *open source* populares dentro de la comunidad científica y un modelo propietario, utilizando las mismas

	Ofensivo			Posición			Informatividad			Felicidad		
	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc
BETO uncased	.64	.72	.66	.81	.83	.80	.76	.88	.75	.82	.87	.80
Bertin-roberta-base	.58	.54	.62	.92	.93	.92	1.0	1.0	1.0	.77	.78	.80
ST-paraphrase-xlm-r-multi-v1	.55	.66	.55	.67	.77	.69	1.0	1.0	1.0	.61	.62	.61
ST-paraphrase-multi-MiniLM-L12-v2	.41	.40	.44	.54	.67	.52	.38	.30	.50	.55	.70	.60
ST-paraphrase-multi-mpnet-base-v2	.54	.63	.53	.88	.93	.86	.95	.96	.94	.66	.66	.66

Tab. 6.6: Resultados obtenidos por los modelos para predecir las cuatro categorías propuestas de evaluación en el conjunto de datos de evaluación en español, extendido con contranarrativas generadas por humanos anotadas con el máximo valor en todas las categorías

	Ofensivo			Posición			Informatividad			Felicidad		
	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc	F1	Pr	Rc
BETO uncased	.67	.72	.68	.58	.78	.63	.29	.20	.50	.56	.50	.67
Bertin-roberta-base	.47	.64	.54	.70	.70	.70	.29	.20	.50	.44	.35	.57
ST-paraphrase-xlm-r-multi-v1	.41	.65	.44	.44	.77	.53	.29	.20	.50	.37	.47	.45
ST-paraphrase-multi-MiniLM-L12-v2	.64	.67	.62	.64	.71	.67	.29	.20	.50	.41	.33	.53
ST-paraphrase-multi-mpnet-base-v2	.67	.72	.68	.58	.78	.63	.29	.20	.50	.56	.50	.67

Tab. 6.7: Resultados obtenidos por los modelos para predecir las cuatro categorías propuestas de evaluación en el conjunto de datos de evaluación en español, sin extenderlo con contranarrativas generadas por humanos

variantes de aprendizaje de la tarea, información argumentativa y tipos de contranarrativas que en los experimentos realizados con Flan-T5.

6.3.1 Modelos

Para generar las contranarrativas, utilizamos los siguientes modelos como base:

- **Llama 3.1** [Dubey et al., 2024]: Es una familia de modelos de lenguaje desarrollados por Meta que soportan de manera nativa multilingüidad, generación de código, razonamiento y uso de herramientas, entre otras capacidades. También tiene una versión ajustada al uso de instrucciones. Incluye modelos con 8 mil millones, 70 mil millones y 405 mil millones de parámetros entrenados en un conjunto masivo de datos que abarca más de 15 billones de tokens, casi diez veces más que Llama 2 que utilizó tan solo 1.8 billones. Llama 3, a diferencia de sus predecesores, incorpora mejoras como la *Grouped Query Attention* (GQA) para optimizar la eficiencia en inferencia, así como una ventana de contexto extendida de hasta 128,000 tokens para el modelo de mayor tamaño. Esto le otorga un desempeño altamente competitivo en comprensión del lenguaje natural, generación de código, razonamiento complejo y tareas multilingües. Llama 3 utiliza un procedimiento de post-entrenamiento basado en *Supervised Fine-tuning* (SFT), *Rejection Sampling* (RS) y *Direct Preference Optimization* (DPO), a partir de la retroalimentación humana y con el objetivo de alinear mejor sus respuestas con las expectativas del usuario. En este experimento

utilizamos la versión de 8 mil millones de parámetros ajustada al uso de instrucciones.

- **Qwen 2** [Yang et al., 2024]: Es una familia de modelos de lenguaje desarrollados por Alibaba, que incluye diversos tamaños de modelos, desde 0.5 hasta 72 mil millones de parámetros, pasando por valores intermedios como 1.5 y 7 mil millones de parámetros. Su arquitectura, basada en Transformers, incorpora mejoras como activación SwiGLU, *Grouped Query Attention* (GQA) y un tokenizador que optimiza el manejo de múltiples idiomas y código. Los modelos Qwen 2 están disponibles en versiones base y ajustados con instrucciones. Qwen 2 ha demostrado ser competitiva frente a los modelos del estado del arte, incluyendo a modelos propietarios como GPT4 en diversas tareas de comprensión del lenguaje, razonamiento matemático y programación. Un aspecto que vuelve única a esta familia de modelos es la disponibilidad que tiene en tamaños pequeños que pueden ser deployados en entornos de bajos recursos. Esa disponibilidad nos permite, en el caso de nuestros experimentos, poder evaluar el impacto que tiene en la tarea de generación de contranarrativas, el tamaño del modelo. Para nuestro experimento utilizamos la versión de 500 millones de parámetros y la de 7 mil millones de parámetros.
- **Mistral** [Jiang et al., 2023]: Es una familia de modelos de lenguaje desarrollados por MistralAI que, al igual que los modelos anteriores, soportan de manera nativa multilingüidad, generación de código, razonamiento, operaciones matemáticas y uso de herramientas. Mistral es un modelo de lenguaje diseñado especialmente para obtener un buen rendimiento utilizando una menor cantidad de parámetros que otros modelos similares, como Llama 2. También tiene versiones ajustadas a instrucciones disponibles. Mistral cuenta con una versión de 7 mil millones de parámetros y otras dos versiones que utilizan 8 matrices esparsas, constituyendo una arquitectura llamada de mixtura de expertos (*Sparse Mixture of Experts*, SMoE), una donde cada matriz tiene 7 mil millones de parámetros y la otra con 22 millones de parámetros y un tamaño total de 141 mil millones de parámetros. La arquitectura de mixtura de expertos permite activar selectivamente solo determinadas porciones del total de parámetros del modelo, haciéndolos computacionalmente muy eficientes en tareas complejas, sin la necesidad de la carga computacional pesada que tienen otros modelos grandes.
- **GPT4-o**: Es una arquitectura propietaria desarrollada por OpenAI basada en una optimización del modelo GPT-4, que utiliza entre 500 mil millones y 1 billón de parámetros (no se ha publicado el número exacto) para generar texto natural utilizando de forma nativa herramientas como la generación de código, razonamiento, imágenes, entre otras. Entre sus características técnicas destacan un preentrenamiento con corpus multilingües masivos y heterogéneos combinados con ajuste fino (fine-tuning) supervisado utilizando una enorme cantidad de *datasets* públicos disponibles e instrucciones RLHF (Reinforcement Learning from Human Feedback), desarrolladas mediante la intervención de anotadores humanos, para mejorar la alineación con los objetivos del usuario. Su capacidad para manejar ventanas contextuales extendidas permite aplicaciones en redacción de documentos extensos, análisis de datos complejos y soporte en investigación científica, mientras que sus mecanismos de control de sesgos y generación responsable, minimizan riesgos éticos y sociales. En este

trabajo utilizamos este modelo sin ajustar, mediante técnicas de aprendizaje en contexto.

Al igual que hiciéramos con la familia de modelos de Flan-T5, generamos algunas contranarrativas utilizando aprendizaje en contexto y otras ajustando los modelos a nuestro *dataset*, combinando a su vez distintas variantes de información argumentativa y de tipos de contranarrativas, del mismo modo que hiciéramos en la generación descrita en la sección 5.2. Para los modelos ajustados, en lugar de ajustar todos los parámetros del modelo, utilizamos cuantización y adaptación con matrices de bajo rango (*Quantized Low-Rank Adaptations*, QLORA) [Dettmers et al., 2023].

La cuantización es un método para reducir el tamaño de los modelos a partir de bajar la precisión numérica de los parámetros para que utilicen una menor cantidad de *bits* (típicamente 4, aunque también existe la cuantización a 2 y a 8 bits). Si bien esto disminuye la precisión de los valores de punto flotante que componen los parámetros, está demostrado que el impacto en el rendimiento de los modelos, para la gran mayoría de las tareas, es mínimo, mientras que de este modo se logra reducir drásticamente los requerimientos computacionales necesarios para entrenar y hacer inferencia con los modelos.

La adaptación con matrices de bajo rango, por su parte, consiste en utilizar un nuevo conjunto de parámetros organizados de forma matricial en la etapa de entrenamiento. En lugar de ajustar los parámetros originales del modelo, se ajustan los parámetros de la nueva matriz de forma tal que al combinarla con los pesos del modelo original pre-entrenado, la salida de una pasada hacia adelante, realizando una inferencia, sea igual que si se hubieran ajustado los pesos del modelo original directamente, con la ventaja de que las matrices tienen bajo rango y, por lo tanto, ocupan una cantidad de memoria y de recursos computacionales inferior. Este enfoque ha demostrado ser particularmente efectivo para adaptar modelos pre-entrenados de propósito general a tareas específicas, manteniendo el rendimiento pero ofreciendo ganancias significativas en términos de eficiencia computacional, velocidad y requerimientos de infraestructura.

6.3.2 Entrenamiento y Prompting

Para generar las contranarrativas utilizando el modelo Flan-T5, desarrollamos un prefijo para la entrada del modelo con una instrucción corta y simple que instruíra al modelo a generar una contranarrativa. Este *prompt* se basó en el estilo y características del conjunto de datos FLAN [Chung et al., 2022] utilizado para ajustar distintos modelos a seguir instrucciones, que en general utilizaba instrucciones concretas y precisas.

Los grandes modelos de lenguaje generativos, especialmente aquellos adaptados como asistentes conversacionales, utilizan *prompts* de sistema que definen un rol que debe cumplir el modelo de lenguaje. Este rol define el acervo cultural y educacional implícito, la postura respecto a un tema, el nivel de educación y conocimientos, el estilo, el sentimiento, el tono y una multiplicidad de otros factores deseados en las respuestas generadas. La figura 6.5 muestra los *prompts* en inglés y español utilizados para generar contranarrativas con los modelos detallados en la sección 6.3.1. Los *prompts* tienen una parte del texto encerrada entre corchetes y destacada en color azul. Esta parte describe la utilización de la información argumentativa y advierte al modelo contra incluir esta información en la respuesta generada. Colocamos esta sección entre corchetes debido a que sólo la incluimos en aquellos modelos que efectivamente incorporan información argumentativa en el proceso de generación, mientras que si el mensaje se genera sin esta, no la incluimos.

You are a NGO operator expert on generation of counter-speech and counter-narratives against hate messages. You only speak English and are unable to generate text in other languages. You are tasked with generating a response to a hate speech tweet.

[You will also be provided with argumentative information about the tweet separated by '|'. Use this information to elaborate a response. Do not include this information in the response.]

You should only reply the hate tweet directly without adding anything else. The hate speech tweet is the following:

Sos un operario de una ONG experto en generación de contranarrativas y contra-discurso contra el discurso de odio. Solo hablas Español y eres incapáz de generar texto en otro idioma. Tu tarea es generar una respuesta en Español a un tweet de odio. [También se te proporcionará información argumentativa sobre el tweet separada por '|'. Utiliza esta información para elaborar una respuesta. No incluyas esta información en la respuesta.]

Responde el tweet directamente sin agregar ninguna otra información que no sea la respuesta. El tweet con el mensaje de odio es el siguiente:

Fig. 6.5: Prefijo utilizado en la entrada de los modelos explicando la tarea a realizar en inglés y en español. La sección delimitada por corchetes y destacada en color azul sólo se incluye en aquellos modelos que utilizan información argumentativa

Todos los modelos utilizados en esta sección utilizan un *template* conversacional. En aquellos casos donde los ejemplos anotados, utilizados para que el modelo aprenda la tarea, se incorporan en contexto dentro de la entrada del modelo, utilizamos el *template* conversacional para simular una conversación entre el usuario y el asistente en la cual el discurso de odio es desarrollado por el usuario y el asistente responde con la contranarrativa correspondiente.

6.3.3 Resultados

Para realizar la evaluación utilizamos el modelo que mejores resultados obtuvo en cada clase y para cada idioma. De acuerdo a los resultados de las tablas 6.4, 6.5, 6.6 y 6.7, los modelos que obtuvieron un rendimiento mayor utilizan todos la versión extendida del *corpus* de entrenamiento, aumentados con los ejemplos generados por humanos. Para el inglés, utilizamos el modelo de *sentence transformers* **paraphrase-multilingual-mpnet-base-v2** para evaluar las categorías Ofensividad, Informatividad y Felicidad, y el modelo de *sentence transformers* **all-mpnet-base-v2** para la categoría Posición. Para el español, utilizamos el modelo **BETO** para las categorías Ofensividad y Felicidad y el modelo **Bertín** para las categorías Posición e Informatividad. Los resultados de la evaluación pueden observarse en la tabla 6.8. Para cada combinación de tipo de modelo, método de aprendizaje, información argumentativa adicionada y tipos de contranarrativas utilizados en el aprendizaje, la tabla muestra el promedio, entre las 145 contranarrativas generadas para esa combinación, de cada categoría (**O**fensividad, **P**osición, **I**nformatividad y **F**elicidad), el puntaje calculado utilizando la fórmula descrita en la sección 5.4.0.2 y el porcentaje de ejemplos clasificados como **A**ceptables y **B**uenos, en inglés y en español.

A continuación presentamos las principales conclusiones extraídas de los resultados de

la tabla y de distintas inspecciones manuales realizadas a partir de estos.

6.3.3.1 Efectividad para detectar incompatibilidades manifiestas

Los resultados de la tabla 6.8 muestran 6 sectores donde el puntaje total agregado no supera los 6,5 puntos (color naranja o rojo) y otros 5 sectores donde el puntaje no supera los 7,5 puntos (color amarillo).

Empezando nuestro análisis de arriba para abajo, el primer sector que encontramos que obtiene un puntaje bajo son los Llama3.1 con aprendizaje en contexto que incorporan información argumentativa sobre las Premisas (Justificación y Conclusión) y toda la información argumentativa disponible, ambos en español. Observamos que el bajo puntaje de estos modelos también es acompañado de un bajo porcentaje de contranarrativas aceptables y buenas. Al investigar las respuestas generadas, observamos que, mientras que los mismos modelos pero sin información argumentativa o con información sobre el Colectivo y la Propiedad generan respuestas buenas de alta calidad, en general, estos dos modelos generan algunas respuestas que parecen ser aceptables, y una gran cantidad de respuestas que repiten parte del contenido de odio del *tweet* señalando dentro de la respuesta, cuáles podrían ser posibles conclusiones e intercalando indicadores de que algo es de tipo “hecho”. Algunas veces, se esboza luego de esa repetición del contenido de odio, una respuesta. Otras veces se incluyen otro tipo de comentarios intercalados con el mensaje de odio como “(noticia falsa)” o “(pregunta retórica)” o “(pregunta razonable)”, lo cual induce a pensar que el modelo confunde la tarea a realizar con un análisis argumentativo del mensaje de odio.

En segundo lugar, encontramos que tanto para el inglés como para el español, el modelo Qwen0.5B con *fewshot* rechaza siempre e invariablemente las peticiones de generar contranarrativas, con excepción de aquellos en español que tienen información sobre las premisas o sobre todos los componentes argumentativos. Mientras que en inglés la respuesta es en la totalidad de los casos “*I’m sorry but I can’t assist with that*”, en español la respuesta para las entradas que no incluyen información argumentativa o incluyen información sobre el colectivo y la propiedad son, también, “Lo siento, pero no puedo asistir con eso”. El mensaje generado no aclara el motivo del rechazo, pero especulamos que, también, pueda llegar a haber una confusión sobre la tarea a realizar, y al verse en una situación en la que el modelo cree que puede llegar a tener que reproducir contenido de odio, se activa la zona de rechazo, generando estos mensajes. Si bien como contranarrativa, la totalidad de estos mensajes debería ser rechazada (lo cual se refleja correctamente en el bajo porcentaje de contranarrativas aceptables), lo cierto es que el puntaje no es tan bajo como podría serlo debido a que la mayoría de las respuestas obtienen 2 puntos en las categorías de Odio y Posición (lo cual es correcto porque no son mensajes que reproduzcan el odio o tengan una posición favorable al mensaje original) y casi invariablemente obtienen 3 puntos en Felicidad. Para el español con información sobre las premisas o sobre todos los componentes argumentativos, nuevamente el modelo repite partes del mensaje de odio agregando etiquetas como “Conclusión:” e introduce etiquetas como “(fact)” o directamente genera texto ofensivo y/o incoherente. También encontramos que en algunos casos cambia el idioma de la respuesta a inglés o reproduce mensajes de odio novedosos. Por este motivo obtienen un puntaje aún menor en la categoría de Ofensividad, Informatividad y Felicidad, obteniendo un puntaje agregado también menor.

En tercer lugar, encontramos algunos casos donde la calidad de las contranarrativas es oscilante. Para el modelo Qwen7B con *fewshot* en español e información argumentativa

sobre las premisas o toda la información argumentativa y para el modelo Mistral también en español, *fewshot* e información argumentativa de las premisas y toda la información argumentativa, algunas contranarrativas presentan una calidad aceptable mientras que una proporción alta no lo hacen. Para el modelo Qwen7B, algunas contranarrativas generadas nuevamente presentan etiquetas sobre si algo es una Conclusión o indicando que una porción del discurso es un “(*fact*)”. A diferencia del modelo Qwen0.5B, en este caso, algunas contranarrativas generadas parecen ser correctas, pero parecen ser a su vez minoritarias. En el caso de Mistral, las respuestas analizadas siempre empiezan marcando la Conclusión de uno de los *tweets* utilizados como parte del aprendizaje en contexto, y luego elaboran una respuesta correcta y contraria a la xenofobia pero para este *tweet* y no para el que deberían responder. Más allá de los errores en la generación, observamos que sobre todo, impacta en estos casos una baja en el puntaje asignado a la Felicidad, la cual suele ostentar un puntaje máximo si las contranarrativas generadas son de buena calidad.

Por último, también encontramos casos de falsos negativos, es decir, grupos de modelos que fueron calificados con bajos puntajes y bajos porcentajes de contranarrativas aceptables que, al ser analizados, no parecen tener tal mala calidad. En este grupo incluimos a los modelos Qwen7B, Mistral y GPT4 que utilizan *fewshot* en inglés y a los modelos Mistral ajustados con contranarrativas de tipo B (con o sin información argumentativa). Estos modelos generan contranarrativas que poseen alguna particularidad distintiva pero que no por esa razón, juzgamos, un evaluador humano calificaría como peores.

El modelo Mistral con *fewshot* genera, por ejemplo, respuestas largas compuestas de *items* enumerados. Casi la totalidad de las respuestas generadas sigue un patrón en el cual la respuesta se muestra desglosada en 6 o 7 *items* enumerados, cada uno una respuesta corta pero adecuada en sí misma a distintos aspectos del mensaje de odio original. En la figura 6.6 podemos observar un ejemplo generado por Mistral en inglés mediante *fewshot* con estas características que tiene solo 5 *items*, aunque observamos que en la mayoría de los casos el modelo puede generar entre 9 y 11 *items* y mensajes más largos.

Los modelos Qwen7 y GPT4, por su parte, generan también respuestas en su casi totalidad adecuadas pero en un tono más conciliador, amable y positivo que lo que suelen demostrar las contranarrativas mejor calificadas. En casi la totalidad de los casos, las contranarrativas de estos dos modelos terminan con alguna fórmula que suele repetirse, como por ejemplo “*Lets focus on constructive dialogue / policies that include us all / promoting respect to all individuals*” u algún otro *slogan* de características generales, especialmente en GPT4, donde además las respuestas suelen ser más largas e incluyen casi siempre varios hashtags reforzando las conclusiones generales del argumento.

En el caso del modelo Mistral ajustado con contranarrativas de tipo B, también observamos que las contranarrativas generadas parecen ser adecuadas. En este caso, las respuestas suelen ser cortas y orientadas a seguir la estrategia de las contranarrativas de tipo B de negar la relación entre el colectivo discriminado y la propiedad negativa que se les endilga. Observamos una proporción alta de contranarrativas calificadas con el valor 2 para las categorías de Ofensividad y Posición cuando en verdad debieran haber sido calificadas con 3.

Con el objetivo de verificar nuestras observaciones, tomamos una muestra de 20 contranarrativas generadas por el modelo Mistral con *fewshot* en inglés sin información argumentativa, 20 generadas por el modelo Qwen7 con *fewshot* en inglés con información de las premisas, 20 generadas por GPT4 con información del colectivo y 20 generadas por el modelo Mistral ajustado con contranarrativas de tipo B y con información sobre el

colectivo, de forma aleatoria, excluyendo aquellas que fueron calificadas como aceptables (solo elegimos pares de odio y contranarrativa que no obtuvieron el mejor puntaje ya sea en Ofensividad, Posición o Felicidad). Para todas las contranarrativas seleccionadas, un anotador humano las re-evaluó como “Aceptables” obteniendo puntaje máximo en las tres categorías mencionadas, mostrando las limitaciones que tiene nuestro sistema de evaluación automático.

Excluyendo estos últimos casos de falsos negativos, encontramos que, en general, los modelos que obtienen bajos puntajes y bajo porcentaje de contranarrativas aceptables tienen un peor desempeño, mientras que todos los modelos que han obtenido un puntaje superior a 7.5 puntos tienen un buen desempeño. Más aún, en general observamos una correlación entre el puntaje agregado obtenido y el porcentaje de contranarrativas aceptables y buenas identificadas, lo cual habla de la fiabilidad de utilizar este puntaje como método de evaluación del desempeño de los modelos.

6.3.3.2 Errores y limitaciones del sistema de evaluación

Un hecho que nos llamó la atención al analizar los resultados es que, incluso en los modelos con un desempeño bueno, altos puntajes y un porcentaje de contranarrativas aceptables también alto, no existe ninguno, para el idioma inglés, que haya obtenido un puntaje perfecto o cercano a perfecto en las categorías de Ofensividad y Posición (como sí sucede con la categoría Felicidad). Nos sorprende este hecho porque sabemos que muchos de los modelos utilizados atraviesan un riguroso proceso de alineación que involucra técnicas de *Reinforcement Learning with Human Feedback (RLHF)* que tienen como objetivo evitar que los modelos de lenguaje generen texto ofensivo (y por lo tanto, a su vez, que apoyen un contenido que también lo es).

Las técnicas de alineamiento de modelos basadas en *RLHF* [Ouyang et al., 2022] consisten en utilizar un modelo base para generar texto para una tarea específica, evaluar el texto generado de forma supervisada utilizando anotadores humanos y luego entrenar un Modelo de Recompensa con esos datos. Ese modelo de recompensa es utilizado luego como función de recompensa para un modelo generativo a través de un algoritmo de optimización como *Proximal Policy Optimization (PPO)* o *Direct Policy Optimization (DPO)*, que busca maximizar las probabilidades del modelo de generar textos que obtengan una recompensa alta y, por lo tanto, estén alineados a las preferencias de los usuarios. Los cinco modelos utilizados en nuestro experimento tienen un alto nivel de alineamiento a través de esta técnica, lo que hace que sea realmente muy difícil que generen una respuesta ofensiva. Sin ir más lejos, el modelo Qwen0.5B, al malinterpretar la tarea y considerar que se requería que generase un contenido ofensivo, genera un mensaje rechazando la petición.

A partir de observar que estos modelos tienen distintos porcentajes de contranarrativas que no fueron evaluadas como “no ofensivas” o con una postura contraria al mensaje original, decidimos realizar una evaluación manual de las categorías Ofensividad, Posición y Felicidad. Elegimos las contranarrativas generadas por el modelo Llama3.1 con *fewshot* y sin información argumentativa agregada, y las evaluamos de forma manual una por una. El resultado es que todas las contranarrativas evaluadas tienen el máximo puntaje en las tres categorías. Efectivamente, este modelo no genera contranarrativas ofensivas, que tengan una postura que no sea claramente contraria al mensaje de odio original y con coherencia interna.

Esta característica es muy distinta en los modelos Flan-T5, que no atravesaron ningún proceso de alineamiento y que constantemente reproducen mensajes de odio o toman

posturas favorables al mensaje de odio original, lo cual nos lleva a plantearnos que posiblemente, el corpus utilizado para entrenar el modelo de evaluación automático no es representativo del universo a evaluar en esta sección, donde encontramos muy pocos casos con esta característica.

Este análisis permite interpretar los resultados obtenidos como, no sólo una evaluación del rendimiento de los modelos evaluados, sino principalmente como una evaluación del sistema de evaluación en general. Si todas las contranarrativas generadas son “no ofensivas”, ¿por qué algunos modelos obtienen puntajes más altos que otros?

Si bien observamos que existe una correlación entre la adecuación de las respuestas generadas y el puntaje obtenido y que, por lo tanto, las respuestas inadecuadas obtienen por lo general un puntaje más bajo y las respuestas adecuadas un puntaje más alto, existen distintos factores que influyen en el puntaje que no se relacionan con la calidad o la adecuación de las respuestas generadas y que generan que el puntaje propuesto a modo de evaluación no sirva para realizar un análisis fino sobre la calidad de las respuestas, sino más bien un filtrado grueso que permite detectar problemas groseros.

En particular, nos referimos al hecho de que el sistema de evaluación fue entrenado con contranarrativas generadas mayoritariamente por la familia de modelos Flan-T5, que en la mayoría de los casos apenas superan un 25 % de contranarrativas aceptables en inglés y en español tienen un promedio inferior al 10 % con muchos modelos que no logran generar ni una sola, mientras que la mayoría de los modelos que pueden observarse en la tabla 6.8 generan contranarrativas aceptables en una proporción altamente superior.

En trabajos subsiguientes, nos proponemos crear un dataset a partir de la anotación humana de un subconjunto de las contranarrativas generadas por los modelos descritos en la sección 6.3.1, apuntando a una mayor cantidad de ejemplos y a una mayor variación lingüística. Considerando que para varios de estos modelos el puntaje para las categorías de Ofensividad, Posición y Felicidad pueda resultar cercano al óptimo, nos proponemos redefinir las categorías de evaluación de modo tal que sea posible distinguir de mejor manera las contranarrativas Aceptables entre sí (ver sección 7.2.2). A su vez, evaluaremos métodos alternativos a aquellos basados en categorías, basados en *rankings*, para estudiar si existe alguna correlación entre ambos métodos de evaluación.

6.3.3.3 Diferencias entre los enfoques de ajuste fino y aprendizaje en contexto

Otra característica que salta a la vista al observar la tabla es que todos los modelos, con la excepción parcial de Llama3.1, obtienen un mayor puntaje en sus variantes ajustadas a nuestro conjunto de datos que en sus variantes con aprendizaje en contexto a través de un *fewshot learner*. Esta conclusión, en principio, es consistente con el resultado analizado en la sección 5.6.1. Al extender este análisis a distintos modelos, sin embargo, observamos que el método de aprendizaje impacta distinto según el tipo y, sobre todo, según el tamaño del modelo. Mientras que para Qwen0.5B el enfoque *fewshot* sólo produce respuestas dentro de la zona de rechazo o respuestas que reproducen el odio o son incoherentes, el resto de los modelos, Llama3.1, Qwen7B, Mistral y GPT4 producen respuestas, en gran medida aceptables y al menos mejores que aquellas generadas tanto por Qwen0.5B como por Flan-T5 en el experimento de la sección 5.6.

En términos generales, lo que observamos es que la generación utilizando *fewshot* produce mayor variación del estilo, longitud y otras características entre distintos modelos, mientras que la generación utilizando modelos ajustados a un dataset es más uniforme y

se parece más entre sí. Los modelos sin ajustar, además, tienden a tener un estilo más amable, a utilizar fórmulas comunes con mensajes generales o a utilizar mayor cantidad de *hashtags*.

Nuestra hipótesis es que el ajuste fino a un mismo conjunto de datos reduce la variabilidad de las contranarrativas generadas por distintos modelos. Es decir, las contranarrativas generadas por distintos modelos ajustados al mismo conjunto de datos se parecen más entre sí. De este modo, las contranarrativas generadas, por ejemplo, por algún modelo Qwen ajustado, se parecen más a las generadas por el modelo Flan-T5 ajustado que el sistema evaluador observó durante su entrenamiento como las de mayor puntaje, mientras que no sucede lo mismo con las contranarrativas, por ejemplo, generadas por Mistral con *fewshot*. En definitiva, esto sugiere que el evaluador tiene un sesgo por las contranarrativas generadas por modelos ajustados con nuestro *dataset*.

6.3.3.4 Diferencias entre las evaluaciones en español e inglés

La tabla 6.8 muestra que, con algunas excepciones puntuales, en general los modelos suelen obtener mejor puntaje en español que en inglés. El cambio se debe principalmente a que, para aquellos modelos que generan buenas contranarrativas, el puntaje de Ofensividad y Posición suele ser mayor. Como planteamos en la sección 6.3.3.2, considerando que estos modelos son prácticamente incapaces de generar contenido ofensivo o que avale un mensaje de odio, estos puntajes en español parecen ser más representativos de la realidad que sus contrapartes en inglés.

Nuestra hipótesis es que estos resultados se relacionan con la distribución de las clases del *dataset* anotado con las evaluaciones. En las tablas 6.2 y 6.3 podemos observar que, mientras que para el inglés, el porcentaje de contranarrativas anotadas con el mayor puntaje (3) entre el *dataset* no extendido y el extendido aumenta en un 13 %, 15 %, 26 % y 10 % para las categorías de Ofensividad, Posición, Informatividad y Felicidad, respectivamente, para el español, estos aumentos son del 18 %, 22 %, 56 % y 22 %, para las mismas categorías, respectivamente. Es decir, las contranarrativas anotadas por seres humanos tienen mayor ponderancia en tanto representantes de la clase de mayor valor en el conjunto de datos del español. Por lo tanto, una posible explicación de por qué los modelos de evaluación en español funcionan mejor podría ser que, quizás, las contranarrativas generadas por los modelos para este experimento sean más similares a aquellas escritas por los anotadores humanos que a aquellas generadas por Flan-T5.

6.3.3.5 Comparación entre tamaños de modelos

Para analizar el impacto del tamaño del modelo (medido en cantidad de parámetros) en la generación de contranarrativas, comparamos los resultados obtenidos por el modelo Qwen0.5B con respecto al obtenido por el modelo Qwen7B.

Al comparar los resultados obtenidos con el enfoque *fewshot*, observamos que, mientras que el modelo de 500 millones de parámetros rechaza responder cualquier pedido de elaborar contranarrativas, el modelo de 7 mil millones de parámetros elabora contranarrativas aceptables en una cantidad significativa de casos, demostrando una capacidad mayor de cumplir la tarea especificada.

Sin embargo, al comparar los resultados obtenidos con el enfoque de ajuste fino al conjunto de datos de entrenamiento, observamos que el desempeño de ambos modelos se empareja notoriamente. Esto puede observarse tanto en el puntaje y en el porcentaje de

contranarrativas aceptables y buenas que muestra la tabla 6.8, como en los ejemplos que muestra la figura 6.6. De este modo, se demuestra que otra de las ventajas que posee la utilización de un conjunto de datos con contranarrativas de calidad es la de permitir generar contranarrativas de alta calidad con modelos de menor tamaño, que pueden ser deployados en entornos de bajos recursos y con menores requerimientos computacionales.

6.3.3.6 Comparación de respuestas generadas

Las figuras 6.6 y 6.7 muestran las respuestas generadas para un mismo *tweet* de odio por parte de los cinco modelos utilizados con un enfoque de *fewshot learning* y los cuatro modelos ajustados al conjunto de datos de entrenamiento, en inglés y español respectivamente. Si comparamos las respuestas generadas mediante *fewshot* con las respuestas generadas por los modelos ajustados, podemos ver que estas últimas tienen un estilo más similar entre sí y menos variabilidad. Suelen ser más cortas y repiten con mayor frecuencia estrategias tales como cuestionar la veracidad de afirmaciones realizadas en el *tweet* original. Por su parte, las figuras 6.8 y 6.9 muestran las respuestas generadas para un mismo *tweet* de odio por parte del modelo Qwen7 utilizando todas las distintas combinaciones de información argumentativa, tanto para el enfoque *fewshot* como para los modelos ajustados a los datos de entrenamiento, en inglés y español respectivamente. En estos casos, el *tweet* de odio tiene etiquetadas la Conclusión (en azul), la Justificación (en rojo), el Colectivo (subrayado continuo) y la Propiedad (subrayado con guiones). Podemos observar que para las contranarrativas generadas con *fewshot* en inglés, aquellas que usaron el Colectivo y Propiedad por un lado, y las Premisas por el otro parecen haber producido respuestas que toman en cuenta y utilizan una estrategia de respuesta basada en esta información, negando la relación entre el Colectivo y la Propiedad que se le asigna o entre la Justificación y la Conclusión. En el ejemplo de la figura (que fue seleccionado al azar), al menos, no parece ocurrir lo mismo con las contranarrativas generadas con los modelos ajustados, aunque todas ellas sí parecen ser respuestas adecuadas y de alta calidad. En el caso del español, la información sobre el Colectivo parece haber servido para orientar la contranarrativa a atacar la condición de invasores de los inmigrantes (o alternativamente negar que haya una invasión), mientras que la información sobre las premisas o sobre todos los componentes argumentativos no parece generar contranarrativas totalmente coherentes (especialmente para el caso de *fewshot*)

Para el español, con el enfoque *fewshot* en los casos donde se utilizan las premisas o toda la información argumentativa se producen contranarrativas inadecuadas con una alta frecuencia. En el primer caso, reproduciendo una narrativa de odio e intentando hacer un “análisis” (con componentes inventados distintos a los propuestos por nosotros, como “Conclusión 2”) para luego elaborar una respuesta, y en el segundo caso, directamente agregando un discurso de odio con excesivas etiquetas del Tipo de premisa intercaladas. Considerando que este fenómeno no se observa cuando se agrega información argumentativa sobre el Colectivo y la Propiedad, nuestra hipótesis es que el exceso de información argumentativa en los modelos no ajustados afecta la capacidad del mismo de comprender la tarea requerida.

En términos generales, observamos que la calidad de las contranarrativas generadas es superior a la de aquellas generadas por la familia de modelos Flan-T5 que utilizamos para nuestros experimentos de la sección 5. La gran mayoría de los modelos en la gran mayoría de sus configuraciones de generación, generó contranarrativas con un alto grado de adecuación. El modelo Qwen0.5B que posee un tamaño similar al del modelo Base

de Flan-T5 no logró generar ninguna respuesta adecuada con *fewshot*, pero las respuestas generadas a partir de ajustarlo al conjunto de datos de entrenamiento muestra una calidad similar a la de aquellos modelos que tienen hasta 16 veces más cantidad de parámetros, demostrando nuevamente la importancia del proceso de ajuste fino a un conjunto de datos de calidad.

TWEET: this should be every country they come here seeking asylum better life then become a burden to our financials system and criminal system why do we allow this? :hashtag: deport them all no respect for :hashtag: america @user :hashtag: build that wall :hashtag: maga :hashtag: patriots united :hashtag: apology tour is over

Fewshot:

Llama3.1: I see people from other countries seeking a better life, and many of them contribute positively to our society, create jobs, pay taxes, and enrich our culture.

Qwen0.5: I'm sorry, but I can't assist with that.

Qwen7: Respect for all individuals, regardless of their origin, is a cornerstone of any society. Seeking a better life is a universal human desire, but it should be done legally and with consideration for the host country's resources and laws. Let's focus on creating a system that is fair and just for everyone, ensuring that newcomers can contribute positively to their new society without becoming a burden. #UnityNotDivision.

Mistral: 1. Seeking asylum is a human right, not a burden. 2. Not all asylum seekers are criminals. 3. Building a wall won't solve the problem. 4. America has always been a country of immigrants. 5. It's important to treat everyone with respect and dignity, regardless of their background.

GPT4: It's essential to remember that most asylum seekers are fleeing violence and persecution, not looking to burden any system. They contribute positively to our communities and society. Let's focus on compassion and understanding instead of division. We can build a stronger nation by working together!

Ajustado:

Fig. 6.6: Comparación de contranarrativas en inglés generadas por distintos modelos utilizando, por un lado, un enfoque *fewshot* y por el otro, ajustados a nuestro conjunto de datos de entrenamiento, en ambos casos, sin información argumentativa

TWEET: los subsaharianos conocedores de que nuestro sistema de pensiones está en riesgo se organizan para asaltar nuestra frontera y hacerse cargo de nuestras jubilaciones no temáis sólo vienen con el ánimo de hacer viables nuestros servicios sociales :link:

Fewshot:

Llama3.1: No hay evidencia de que los inmigrantes subsaarianos estén organizados para asaltar nuestras fronteras con el objetivo de aprovechar nuestros sistemas de pensiones. Es importante no caer en estereotipos y prejuicios, y recordar que la mayoría de las personas que buscan refugio o oportunidades en otros países son víctimas de la pobreza, la guerra o la persecución, y buscan simplemente sobrevivir y construir una vida mejor para sí mismas y sus familias.

Qwen0.5: Lo siento, pero no puedo asistir con eso.

Qwen7: Es importante recordar que la movilidad y el intercambio entre naciones son fundamentales para el progreso y la colaboración global. Los inmigrantes no representan un riesgo para nuestros sistemas de pensiones, sino una contribución a la riqueza y diversidad de nuestras sociedades. Todos merecemos un sistema de jubilación justo y equitativo, independientemente de nuestra origen.

Mistral: “Subsaharianos” no existe como categoría oficial ni científica. Además, no hay ninguna evidencia de que estos supuestos individuos estén organizándose para “asaltar” la frontera con tal de “hacer viables” nuestros “servicios sociales”. Esto es simplemente una falsedad malintencionada.

GPT4: Es importante entender que las afirmaciones como estas fomentan el miedo y la división. La realidad es que la migración es un fenómeno complejo y los inmigrantes contribuyen de muchas maneras a la sociedad. En lugar de verlos como una amenaza, debemos reconocer su humanidad y valorar su aporte. La solidaridad y el respeto son esenciales para construir una sociedad más justa.

Ajustado:

Llama3.1: No es cierto que los inmigrantes vengan a vivir de ayudas. Por lo general, vienen a intentar ganarse la vida.

Qwen0.5: No entiendo si son conocedores de la invasión, o si son emigrantes, y lo que importa es que estén acordando hacerlo. No puedes inferirlo a partir de un solo caso.

Qwen7: ¿Y por que no? No veo ninguna razón para que no pueda ser. No es como si solo hubiera un modo de organizarse.

Mistral: No es cierto que los inmigrantes se lleven la jubilación de los españoles.

Fig. 6.7: Comparación de contranarrativas en español generadas por distintos modelos utilizando, por un lado, un enfoque *fewshot* y por el otro, ajustados a nuestro conjunto de datos de entrenamiento, en ambos casos, sin información argumentativa

TWEET: @user @user @user these boats act like a taxi service for economic migrants posing as refugees and only come to europe to seek welfare benefits of the west? (Hecho) if they were refugees why not stop at turkey like the real refugees ? (Hecho)

Fewshot:

Sin Info: It's a complex issue with many sides. It's not as simple as a taxi service or just seeking welfare. People are often fleeing dangerous situations and need safety. Let's focus on finding solutions that protect those in need while respecting the rights of countries to control their borders.

Colectivo: It's understandable to be concerned about the influx of migrants, but conflating economic migrants with refugees and assuming everyone's intentions are solely for welfare benefits might not be fair. Many people seek a better life and safety, not just benefits. It's important to address the root causes of migration and provide humanitarian assistance where needed.

Premisas: Just because someone is seeking help doesn't mean they're lying about their situation. It's not fair to assume everyone is taking advantage of a system they're genuinely in need of.

Todo: It's understandable to be concerned about the welfare system, but it's also important to remember that seeking a better life is a universal human desire. Let's focus on creating a fair and just system for everyone, regardless of where they come from.

Ajustado:

Sin info: What boats? You mean the ones that don't exist? There are no boats full of people looking for a better life in Europe. That's just a figure.

Colectivo: It's funny, because I've never seen a refugee study to see which country would be the most convenient to go to. They escape from war, not knowing where they are going or what they are going to do.

Premisas: It is not true that they are all looking for welfare. Many of them have no choice but to escape, and sometimes the only way to escape is by boat.

Todo: You assume they are economic migrants because you don't speak their language, you don't know where they come from, and you don't understand their culture. Maybe they are refugees, and you are just another xenophobe hiding his prejudice behind a keyboard.

Fig. 6.8: Comparación de contranarrativas generadas por el modelo Qwen7 en inglés con distintas combinaciones de información argumentativa agregada y utilizando un enfoque *fewshot* o ajustados a nuestro conjunto de datos de entrenamiento. El tweet de odio se encuentra anotado con la justificación (en rojo) y la conclusión (en azul). El tipo de cada premisa se encuentra anotado entre paréntesis al final de cada una. El colectivo se marca con una línea continua mientras que la propiedad se anota con una línea punteada

TWEET: 80 000 menores han entrado en españa solos desde marruecos luego llegará la reunificación familiar padres hermanos primos abuelos etc (Hecho) ¿para que los envía marruecos ? ¿alguien pide explicaciones a las autoridades marroquí? ¿son inmigrantes o enviados para invadir españa (Hecho)? :link:

Fewshot:

Sin Info: La inmigración es un asunto complejo y multifacético. En el caso de los menores que llegan a España desde Marruecos, es esencial que las autoridades trabajen en conjunto con las de Marruecos para comprender las razones detrás de estos movimientos y garantizar la protección y el bienestar de estos menores. La cuestión de la “reunificación familiar” es especialmente delicada y requiere un enfoque equitativo y respetuoso con todos los involucrados. No es justo ni eficaz acusar a un país de “enviar” personas para invadir.ºtro sin tener un entendimiento claro de las circunstancias individuales y colectivas detrás de estas decisiones. Es importante promover la comprensión y el diálogo para abordar eficazmente los desafíos de la migración.

Colectivo: La inmigración es un tema complejo y a menudo se manipula para alimentar el miedo y el prejuicio. Es importante recordar que la mayoría de las personas que huyen de su hogar lo hacen por razones de seguridad y necesidad, no por la intención de invadir un país. Cada individuo tiene el derecho a buscar un lugar seguro donde vivir y que sus derechos sean respetados. En el caso de los menores que llegan a España, es esencial que se les brinde la protección y el apoyo que necesitan, en línea con los principios de la Convención de las Naciones Unidas sobre los Derechos del Niño. Es crucial que las autoridades de los países de origen, como Marruecos, estén involucradas en el proceso para garantizar la seguridad y el bienestar de estas personas.

Premisas: Los tíos tatarabuelos y todos querrán entrar alegando ser familiares de los que ya están en españa. — conclusión: 80000 niños solos han entrado a españa (facto) — conclusión 2: luego vendrá la reunificación de la familia (facto) El discurso de que los inmigrantes son una amenaza para la sociedad es una estrategia de manipulación que busca generar miedo y desacreditar a las personas que buscan una mejor vida para ellos y sus seres queridos. Cada persona tiene el derecho a buscar un futuro mejor, independientemente de su origen geográfico o nacionalidad. Es importante recordar que la diversidad es una fuerza que enriquece nuestras sociedades y culturas, y no una amenaza. La integración y el respeto a los derechos de todos los ciudadanos, sin importar su origen, son fundamentales para construir una sociedad justa y equitativa.

Todo: (fact) y marroquí (fact) ¿serán (fact) gays negros (fact) sudacas (fact) o moros (fact)?

Ajustado:

Sin info: No se entiende la diferencia entre inmigrar y invadir? Te parece complicado, o simplemente no te gusta que people de otros lugares vengan a vivir aquí?

Colectivo: No es cierto que se haya producido una invasión en las últimas semanas. No tiene sentido que estén intentando invadir el país, cuando ni siquiera dominan el idioma.

Premisas: No es cierto que los envíen desde Marruecos. Es un país vecino, con una cultura similar. No sería el primer país en negar que sus vecinos son como ellos.

Todo: No tienes por qué asumir que vienen a invadir el país. Nadie te obliga a aceptarlos, ¿verdad?

Fig. 6.9: Comparación de contranarrativas generadas por el modelo Qwen7 en español con distintas combinaciones de información argumentativa agregada y utilizando un enfoque *fewshot* o ajustados a nuestro conjunto de datos de entrenamiento. El tweet de odio se encuentra anotado con la justificación (en rojo) y la conclusión (en azul). El tipo de cada premisa se encuentra anotado entre paréntesis al final de cada una. El colectivo se marca con una línea continua mientras que la propiedad se anota con una línea punteada

Modelo	Aprendiz	Info Arg	CNs	Inglés							Español						
				Of	Pos	Inf	Fel	Pj	Ac	Bu	Of	Pos	Inf	Fel	Pj	Ac	Bu
Llama3.1	Fewshot	No-Info	Todas	2.64	2.62	1.86	2.86	7.98	.54	.53	2.96	2.85	2.56	3.00	9.37	.81	.81
Llama3.1	Fewshot	Colectivo	Todas	2.54	2.51	1.62	2.94	7.61	.45	.44	2.96	2.89	2.67	3.00	9.52	.85	.85
Llama3.1	Fewshot	Premisas	Todas	2.64	2.63	1.93	2.88	8.08	.54	.53	1.67	2.22	0.85	1.66	4.40	.11	.07
Llama3.1	Fewshot	Todo	Todas	2.71	2.66	2.04	2.88	8.29	.60	.60	2.56	2.30	0.89	2.11	5.86	.19	.19
Llama3.1	Ajustado	No-Info	Todas	2.75	2.60	1.84	2.72	7.91	.51	.51	3.00	2.93	2.70	3.00	9.63	.93	.85
Llama3.1	Ajustado	Colectivo	Todas	2.87	2.78	2.28	2.79	8.72	.68	.67	3.00	2.93	2.74	3.00	9.67	.93	.89
Llama3.1	Ajustado	Premisas	Todas	2.72	2.72	2.13	2.63	8.02	.59	.57	3.00	2.96	2.85	3.00	9.81	.96	.93
Llama3.1	Ajustado	Todo	Todas	2.72	2.72	2.13	2.63	8.02	.59	.57	3.00	2.70	2.11	3.00	8.81	.70	.70
Llama3.1	Ajustado	No-Info	A	2.55	2.72	2.17	2.71	8.15	.45	.45	3.00	2.74	2.22	3.00	8.96	.74	.74
Llama3.1	Ajustado	Premisas	A	2.75	2.77	2.30	2.70	8.52	.62	.62	3.00	2.88	2.63	3.00	9.51	.89	.85
Llama3.1	Ajustado	No-Info	B	2.68	2.56	1.66	2.78	7.68	.49	.49	3.00	3.00	2.52	3.00	9.52	1.0	.52
Llama3.1	Ajustado	Colectivo	B	2.50	2.55	1.74	2.84	7.63	.49	.49	3.00	3.00	2.86	3.00	9.86	1.0	.86
Llama3.1	Ajustado	No-Info	C	2.85	2.71	2.14	2.57	8.27	.58	.58	2.91	2.96	2.78	2.91	9.56	.87	.83
Llama3.1	Ajustado	Premisas	C	2.65	2.68	1.93	2.58	7.84	.50	.49	2.91	2.87	2.61	2.83	9.22	.78	.78
Qwen0.5	Fewshot	No-Info	Todas	2.13	2.36	0.81	3.00	6.30	.10	.07	2.07	2.78	2.30	2.18	7.33	.04	.04
Qwen0.5	Fewshot	Colectivo	Todas	2.13	2.35	0.74	3.00	6.22	.10	.06	2.07	2.78	2.30	2.18	7.33	.04	.04
Qwen0.5	Fewshot	Premisas	Todas	2.15	2.34	0.77	2.99	6.25	.12	.08	1.41	2.56	1.74	1.41	5.12	.04	.00
Qwen0.5	Fewshot	Todo	Todas	2.11	2.33	0.76	3.00	6.20	.09	.06	1.85	2.0	0.19	1.11	3.15	.00	.00
Qwen0.5	Ajustado	No-Info	Todas	2.80	2.69	2.05	2.83	8.37	.61	.60	3.00	2.81	2.41	3.00	9.22	.81	.78
Qwen0.5	Ajustado	Colectivo	Todas	2.78	2.73	2.20	2.81	8.52	.61	.60	2.96	2.78	2.34	3.00	9.08	.74	.74
Qwen0.5	Ajustado	Premisas	Todas	2.78	2.77	2.33	2.95	8.83	.72	.72	3.00	2.89	2.63	3.00	9.52	.89	.85
Qwen0.5	Ajustado	Todo	Todas	2.86	2.79	2.35	2.93	8.93	.74	.72	3.00	2.67	1.96	3.00	8.63	.67	.63
Qwen0.5	Ajustado	No-Info	A	2.83	2.68	2.11	2.93	8.55	.64	.64	3.00	2.74	2.22	3.00	8.96	.74	.74
Qwen0.5	Ajustado	Premisas	A	2.85	2.78	2.37	2.96	8.96	.72	.72	3.00	2.85	2.56	3.00	9.41	.85	.85
Qwen0.5	Ajustado	No-Info	B	2.55	2.54	1.66	2.89	7.64	.47	.47	2.90	2.95	2.86	2.86	9.57	.76	.76
Qwen0.5	Ajustado	Colectivo	B	2.55	2.45	1.38	2.89	7.27	.39	.39	2.95	2.86	2.57	2.71	9.09	.71	.71
Qwen0.5	Ajustado	No-Info	C	2.86	2.88	2.54	2.92	9.20	.79	.74	2.91	2.70	2.09	2.83	8.53	.65	.65
Qwen0.5	Ajustado	Premisas	C	2.88	2.79	2.35	2.91	8.93	.77	.74	3.00	2.70	2.09	3.00	8.79	.70	.70
Qwen7	Fewshot	No-Info	Todas	2.28	2.26	1.01	2.81	6.36	.24	.24	2.85	2.70	2.11	3.00	8.66	.63	.63
Qwen7	Fewshot	Colectivo	Todas	2.35	2.31	1.01	2.79	6.46	.28	.28	2.93	2.41	1.22	3.00	7.56	.37	.37
Qwen7	Fewshot	Premisas	Todas	2.39	2.19	0.88	2.86	6.32	.23	.23	2.18	2.48	1.85	2.18	6.69	.37	.26
Qwen7	Fewshot	Todo	Todas	2.41	2.23	0.95	2.86	6.45	.25	.25	2.33	2.33	0.89	1.85	7.40	.07	.07
Qwen7	Ajustado	No-Info	Todas	2.84	2.76	2.22	2.90	8.72	.70	.66	2.96	2.81	2.44	3.00	9.21	.78	.78
Qwen7	Ajustado	Colectivo	Todas	2.77	2.74	2.24	2.83	8.58	.66	.66	3.00	2.85	2.56	3.00	9.41	.85	.85
Qwen7	Ajustado	Premisas	Todas	2.87	2.77	2.26	2.85	8.75	.72	.70	3.00	2.77	2.33	3.00	9.10	.78	.78
Qwen7	Ajustado	Todo	Todas	2.86	2.72	2.15	2.90	8.63	.68	.66	3.00	2.85	2.56	3.00	9.41	.85	.85
Qwen7	Ajustado	No-Info	A	2.88	2.77	2.33	2.84	8.82	.67	.66	3.00	2.78	2.26	2.93	8.97	.74	.70
Qwen7	Ajustado	Premisas	A	2.91	2.75	2.27	2.79	8.72	.65	.64	2.96	2.81	2.44	3.00	9.21	.78	.78
Qwen7	Ajustado	No-Info	B	2.72	2.61	1.82	2.97	8.12	.57	.57	2.95	2.86	2.52	3.00	9.33	.81	.76
Qwen7	Ajustado	Colectivo	B	2.82	2.65	1.95	2.76	8.18	.59	.59	2.95	2.76	2.29	2.90	8.90	.71	.71
Qwen7	Ajustado	No-Info	C	2.85	2.76	2.26	2.85	8.72	.68	.65	3.00	2.91	2.74	3.00	9.65	.91	.91
Qwen7	Ajustado	Premisas	C	2.90	2.72	2.11	2.90	8.63	.69	.65	3.00	2.65	1.96	3.00	8.61	.65	.65
Mistral	Fewshot	No-Info	Todas	1.61	2.34	0.79	2.21	4.95	.10	.10	2.67	2.56	1.67	2.78	7.68	.37	.37
Mistral	Fewshot	Colectivo	Todas	1.63	2.27	0.85	2.26	5.01	.12	.12	2.63	2.48	1.44	2.78	7.33	.33	.33
Mistral	Fewshot	Premisas	Todas	1.78	2.28	0.87	2.02	4.95	.12	.12	2.41	2.78	2.33	2.04	7.56	.37	.37
Mistral	Fewshot	Todo	Todas	1.79	2.22	0.74	2.06	4.81	.06	.06	1.85	2.77	2.33	1.22	6.17	.07	.07
Mistral	Ajustado	No-Info	Todas	2.77	2.68	1.88	2.90	8.23	.61	.54	3.00	2.96	2.85	2.92	9.73	.93	.89
Mistral	Ajustado	Colectivo	Todas	2.77	2.66	1.86	2.88	8.17	.59	.53	3.00	2.85	2.52	2.93	9.30	.85	.81
Mistral	Ajustado	Premisas	Todas	2.77	2.68	1.88	2.92	8.25	.62	.54	3.00	2.85	2.56	3.00	9.41	.85	.85
Mistral	Ajustado	Todo	Todas	2.70	2.68	1.88	2.86	8.12	.57	.50	3.00	2.74	1.81	3.00	8.55	.74	.33
Mistral	Ajustado	No-Info	A	2.89	2.76	2.03	2.94	8.62	.75	.62	2.96	2.56	1.67	2.78	7.97	.44	.44
Mistral	Ajustado	Premisas	A	2.78	2.68	1.79	2.75	8.00	.59	.47	3.00	2.52	1.56	2.85	7.93	.48	.48
Mistral	Ajustado	No-Info	B	2.38	2.27	0.78	2.58	6.01	.22	.20	2.71	2.67	2.14	2.81	8.33	.62	.62
Mistral	Ajustado	Colectivo	B	2.18	2.26	0.92	2.62	5.98	.16	.15	2.81	3.00	3.00	2.81	9.62	.81	.81
Mistral	Ajustado	No-Info	C	2.85	2.74	1.62	2.88	8.09	.71	.42	3.00	2.70	2.09	3.00	8.79	.70	.70
Mistral	Ajustado	Premisas	C	2.80	2.76	1.83	2.87	8.26	.72	.49	3.00	2.65	1.96	2.91	8.52	.65	.65
GPT4	Fewshot	No-Info	Todas	2.48	2.38	1.20	2.98	7.04	.30	.30	3.00	2.56	1.67	3.00	8.23	.56	.56
GPT4	Fewshot	Colectivo	Todas	2.56	2.39	1.26	2.99	7.02	.36	.36	2.89	2.67	2.00	3.00	8.56	.59	.59
GPT4	Fewshot	Premisas	Todas	2.62	2.44	1.37	2.96	7.39	.39	.39	3.00	2.48	1.44	3.00	7.92	.48	.48
GPT4	Fewshot	Todo	Todas	2.61	2.48	1.49	2.96	7.54	.41	.41	3.00	2.59	1.78	3.00	8.37	.59	.59

Tab. 6.8: Evaluación automática realizada con cuatro predictores, uno para cada categoría entre Ofensividad, Posición, Informatividad y Felicidad, elegido de acuerdo al que mejores resultados obtuvo en cada una, tal como puede verse en las tablas 6.4 y 6.6

7 Conclusiones y Trabajo Futuro

7.1 Conclusiones

A continuación, destacamos los principales aportes y conclusiones desarrollados a lo largo de este trabajo.

7.1.1 Construcción de un *dataset* de contra-narrativas aumentado con información argumentativa

Construimos un *dataset* de 970 *tweets* de odio en inglés y 196 *tweets* de odio en español, filtrados a partir de Basile et al. [2019], anotados con componentes argumentativos documentados en un manual de anotación que también hacemos público¹ y que consideramos de valor en sí como herramienta para futuras anotaciones. Utilizando este documento y luego de un proceso iterativo de anotación, puesta en común y refinamiento, logramos un acuerdo entre anotadores “sustancial” (arriba de .6) para 6 de las 8 categorías de anotación (Colectivo, Propiedad, Justificación, Conclusión, Tipo de Conclusión y si el *tweet* es Argumentativo o no), un acuerdo “moderado” (arriba de .4) para la séptima (Pivot) y ningún acuerdo para el Tipo de Justificación, de acuerdo con la interpretación propuesta por Cohen [1960], a pesar del elevado nivel de subjetividad de la tarea y de estar trabajando con texto generado auténticamente por usuarios. Consideramos, por lo tanto, haber alcanzado un buen nivel de acuerdo y de reproducibilidad del proceso de anotación, aunque queda planteado en este trabajo la adaptación de algunas categorías con mayor desacuerdo, como el Pivot sobre la base de los patrones de desacuerdo, como por ejemplo, los casos donde este está señalado por más de un bloque de palabras en cada premisa.

Para cada *tweet* de odio anotado con información argumentativa, anotamos también tres tipos de contra-narrativas distintas, definidas cada una a partir de una estrategia basada en la información argumentativa anotada, siempre que fuera posible. En total, el *dataset* tiene 1722 contra-narrativas anotadas en inglés y 355 en español, un promedio de 2.38 y 2.47 contra-narrativas por cada *tweet* de odio, respectivamente.

Este conjunto de datos tiene valor en sí mismo para cualquier investigador que desee estudiar la tarea de generación automática de contra-narrativas contra el discurso de odio. Hasta donde sabemos, es el primer *dataset* de contra-narrativas anotadas por anotadores humanos respondiendo a mensajes reales generados por usuarios de la red social *Twitter*, una de las más ruidosas que existen y actualmente con un alto nivel de generación de contenido ofensivo.

El conjunto de datos fue publicado en etapas sucesivas. Las anotaciones sobre los

¹ https://docs.google.com/document/d/1AZ1hbrxhZ-n6Lzb9fuaBZd8kAmuLc0lahG6I9dUGz_g/edit?tab=t.0

componentes argumentativos se publicaron en un repositorio denominado ASOHMO², mientras que las contranarrativas separadas por sus respectivos tipos se publicaron en un repositorio denominado CONEAS³. Los modelos pre-entrenados se publicaron en *Huggingface* dentro de la organización ContraNarrativas⁴. También publicamos el *dataset* con las evaluaciones humanas de los pares de discurso de odio y contranarrativas y el código necesario para entrenar los modelos evaluadores⁵.

Hasta donde sabemos, también es la primera vez que se utilizan los componentes utilizados por Wagemans para construir su Tabla Periódica de los Argumentos, re-definiéndolos, no para ser utilizados como un sistema de clasificación, sino como marcadores de relaciones argumentativas sencillas entre distintas partes de un discurso de odio argumentativo, de modo tal que puedan exponer posibles puntos débiles desde el cual el mensaje pueda ser atacado. Consideramos que esto constituye un importante paso en pos de lograr aplicar efectivamente la teoría argumentativa a una tarea concreta de naturaleza argumentativa, colaborando a estrechar la brecha existente entre teoría y práctica.

7.1.2 Componentes argumentativos

En el transcurso de este trabajo, definimos un conjunto de componentes argumentativos basándonos en la teoría de [Wagemans, 2016] que identifican aspectos tanto generales como específicos del dominio del discurso de odio en redes sociales. Demostramos, en primer lugar, que es factible realizar una identificación automática de cuáles *tweets* son argumentativos, y por lo tanto merecen ser contestados, y cuáles no. En segundo lugar, mostramos que algunos de los componentes argumentativos propuestos, como la Justificación, la Conclusión y el Tipo de la Conclusión, pueden ser identificados de forma individual con un clasificador automático entrenado con nuestro conjunto de datos con un nivel aceptable de rendimiento, mientras que para otros componentes como el Colectivo, la Propiedad, el Pivot o el Tipo de Justificación, el rendimiento es menor si la tarea se encara de forma individual y directa. Mostramos, a su vez, que para cada componente, el rendimiento es similar a aquel obtenido por un ser humano intentando predecir las anotaciones de otro ser humano.

Sin embargo, a pesar de esto, mostramos que es posible mejorar los resultados en las tareas de identificación de componentes argumentativos que menor rendimiento tuvieron (Colectivo, Propiedad, Tipo de Justificación y Pivot), a partir de realizar variantes en los procesos de entrenamiento. Para el Tipo de Justificación, utilizar un único modelo que identifique el Tipo de la Premisa, entrenado con los datos combinados de los Tipos de la Justificación y la Conclusión, mejora los resultados en .13 puntos para el tipo de la Justificación, mientras que también mejora en .05 puntos el resultado para el tipo de la Conclusión.

El Colectivo y la Propiedad, por su parte, también demuestran una notable mejora cuando el clasificador incorpora información sobre el componente argumentativo relacionado adverso. Más precisamente, cuando el modelo que identifica al Colectivo incorpora en su entrada la información sobre cuál es la Propiedad y viceversa. Del mismo modo, mostramos que cuando la tarea se divide en dos, por un lado identificar si el Colectivo y la Propiedad son explícitos y deben etiquetarse y por el otro lado, la tarea de identificarlos

² <https://github.com/ASOHMO/ASOHMO-Dataset>

³ <https://github.com/ConeasDataset/CONEAS/>

⁴ <https://huggingface.co/CounterNarratives>

⁵ <https://github.com/DamiFur/CounterNarrativesEvaluation>

en sí mismos, el rendimiento también mejora en términos generales. Lo mismo sucede, en este caso, para el Pivot.

También mostramos que los componentes argumentativos relacionados entre sí, como la Justificación con la Conclusión o el Colectivo con la Propiedad, pueden ser identificados de manera conjunta manteniendo relativamente su nivel de rendimiento, sólo con una pérdida de 5 puntos en la Propiedad o de 3 puntos en la Justificación, lo cual conlleva costos menores en términos de tiempo y recursos computacionales.

En términos generales, los resultados muestran que el Pivot, cuyo objetivo es identificar la relación entre la Justificación y la Conclusión, y que son un componente clave para reconstruir la tipología de Wagemans [2016], no puede ser identificado con éxito, ni por humanos ni de manera automática. Parece que debe adoptarse un enfoque diferente para identificarlos manualmente, posiblemente considerando todas las secuencias de palabras posibles que generen una relación entre la Justificación y la Conclusión.

La información argumentativa identificada automáticamente puede ser aplicada a la generación automática de contra-narrativas, pero también puede ser útil para una cantidad de tareas diversas como recuperación de argumentos o análisis argumentativos automáticos.

Por último, también pudimos demostrar, a partir de incrementar progresivamente el conjunto de datos utilizado para el entrenamiento, que para un conjunto de componentes argumentativos (Argumentatividad, Justificación, Conclusión, Tipo de Conclusión y Colectivo), el tamaño del conjunto de datos utilizado es suficiente para alcanzar un rendimiento cercano al óptimo, con los modelos utilizados, mientras que para otros componentes (como la Propiedad, el Tipo de Justificación y el Pivot), un incremento de la cantidad de ejemplos anotados podría mejorar el rendimiento de los modelos automáticos, lo cual nos plantea la posibilidad de realizar futuras expansiones de nuestro conjunto de datos con el fin de mejorar el rendimiento de los modelos automáticos aún más.

7.1.3 Generación de Contra-narrativas

En la sección 5.6.1 desarrollamos cómo el ajuste fino parece producir mejores resultados que el enfoque *fewshot*, al menos para Flan-T5. Cuando evaluamos el mismo fenómeno utilizando nuestros evaluadores automáticos, sus resultados parecen corroborar esa hipótesis. Sin embargo, una inspección manual nos permitió identificar errores en las evaluaciones realizadas que ocurrían con mayor frecuencia en las contranarrativas generadas mediante *fewshot* que en aquellas generadas ajustando el modelo al dataset. Pudimos observar, entonces, un sesgo de los evaluadores automáticos en favor de las contranarrativas generadas por modelos ajustados a nuestro dataset, las cuales causaban un nivel de error en los clasificadores menor (al menos para Ofensividad, Posición y Felicidad), en favor de un puntaje más alto y sin que necesariamente estas fueran “mejores” que aquellas generadas con un enfoque *fewshot*, al menos en los modelos de mayor tamaño. Planteamos la hipótesis de que el sesgo de los evaluadores pudiera provenir de que los ejemplos anotados con mejores resultados, utilizados para entrenar los modelos, eran aquellos provenientes de los modelos ajustados a nuestro conjunto de datos. Esta situación plantea la necesidad de realizar futuras evaluaciones sobre la base de contranarrativas generadas por una muestra con mayor representatividad y variabilidad de los modelos de lenguaje.

Al utilizar modelos con una mayor cantidad de parámetros (por arriba de los 7 mil millones), el rendimiento, tanto con un enfoque *fewshot* como ajustados a nuestro dataset, muestra una mejora considerable, comparado con las contra-narrativas generadas por los

modelos Flan-T5. Mientras que en los modelos Flan-T5, especialmente aquellos que no fueron ajustados a la tarea, tenían un alto porcentaje de contra-narrativas no aceptables, los modelos Llama3.1, Qwen 7B y Mistral generan en su mayoría contra-narrativas que sí lo son. El modelo Qwen 0.5B rechaza todos los pedidos realizados por el usuario y no genera contra-narrativas cuando se utiliza un aprendizaje en contexto, pero si se ajusta al conjunto de datos elaborado por nosotros genera contra-narrativas de una calidad similar a las que generan los modelos de mayor tamaño. El modelo GPT4, sin ser ajustado a la tarea y alimentado con sólo 5 ejemplos en su entrada, genera contra-narrativas en su casi totalidad aceptables pero con un estilo distinto a las elaboradas en nuestro conjunto de datos y que por lo tanto, nuestros evaluadores automáticos no califican como de la misma calidad, aunque una evaluación manual muestra que esto se debe principalmente a las diferencias en el estilo y no a que las contra-narrativas generadas sean necesariamente malas (al menos no en términos de Ofensividad, Posición y Felicidad).

El ajuste fino permite reducir la variabilidad entre distintos modelos, garantiza la calidad y produce respuestas más previsibles. A la vez, permite a un modelo pequeño en cantidad de parámetros como Flan-T5 Base o Qwen0.5B, alcanzar un rendimiento cercano o hasta similar al de modelos de mayor tamaño, cuando el aprendizaje en contexto genera contra-narrativas que reproducen el mensaje de odio o se niegan a contestar. Este elemento destaca el valor de los conjuntos de datos de alta calidad como herramienta de ajuste de los modelos pre-entrenados de propósito general y plantea el potencial de los modelos más pequeños que, al especializarse en una tarea específica, obtienen un rendimiento similar al de modelos más grandes, pero con la capacidad de ejecutarse en entornos de bajo presupuesto y menores recursos computacionales.

En términos generales, observamos que la información argumentativa no parece tener necesariamente una influencia directa en la calidad de las contranarrativas, de acuerdo a nuestras cuatro categorías de evaluación. Sí observamos que la información argumentativa mejora las contranarrativas generadas cuando se utiliza únicamente para el entrenamiento, el tipo de contranarrativa asociada a esa información argumentativa, mostrando que puede ser útil para dirigir las respuestas generadas hacia una estrategia particular, ganando control sobre el proceso de generación. Observamos, mediante una inspección manual, que en general la información argumentativa destacada es tenida en cuenta por el modelo que genera la contranarrativa, aunque esto no redunde necesariamente en una contranarrativa de mayor calidad.

7.1.4 Evaluación de Contra-narrativas

Argumentamos las problemáticas de los mecanismos de evaluación automáticos y mostramos con ejemplos cómo tanto las métricas que se basan en superposición de n-gramas como las que se basan en buscar cercanía entre *embeddings* que proyectan semánticas similares en regiones cercanas fallan a la hora de evaluar, de forma automática, la generación automática de contra-narrativas, ya que no logran dar cuenta de la multiplicidad de posibles respuestas correctas a un mismo mensaje de odio.

Presentamos también un sistema de análisis de contra-narrativas compuesto por cuatro categorías de evaluación que nos permitieron definir de forma sistemática lo que significa una contra-narrativa Buena, una contra-narrativa Aceptable y una contra-narrativa No-Aceptable. El desglose de la evaluación en distintas categorías de análisis otorga interpretabilidad a los resultados a la vez que facilita el tratamiento del problema con modelos automáticos, que deben aprender tareas más pequeñas y acotadas. También permite tra-

tar problemas de manera focalizada si alguno de los modelos tiene un rendimiento menor, como sucedió con el modelo que predice la Informatividad.

El ajuste de modelos preentrenados con un conjunto pequeño de ejemplos anotados, por su parte, demostró que se pueden obtener muy buenos resultados con relativamente poco esfuerzo en términos tanto de anotación como computacionales, en al menos tres de las cuatro categorías de evaluación propuestas (Ofensividad, Posición y Felicidad, aquellas que definen la aceptabilidad de una contranarrativa) y resultados aceptables también en la cuarta categoría, Informatividad.

Observamos que expandir el conjunto de datos de contra-narrativas generadas por Flan-T5 y evaluadas por seres humanos, con contra-narrativas escritas por seres humanos y evaluadas automáticamente con el máximo puntaje en todas las categorías, mejora sustancialmente el desempeño de los clasificadores automáticos.

Para estos experimentos, mientras que en el inglés obtuvieron mejores resultados los modelos de *sentence embeddings*, preentrenados en tareas pretexto de similaridad semántica, en español obtuvieron mejores resultados los modelos de *word embeddings* que son específicos del español, en comparación a otros modelos multilingüales.

Finalmente, utilizamos los modelos entrenados para realizar evaluaciones masivas de cinco nuevos modelos, cada uno con las mismas combinaciones de información argumentativa, método de aprendizaje y tipo de contranarrativa que con el Flan-T5 y utilizando el conjunto de evaluación completo, en lugar de una muestra aleatoria. Los modelos utilizados incorporan técnicas de alineamiento basadas en *Reinforcement Learning* con retroalimentación humana que disminuyen drásticamente las chances de que el modelo genere contenido ofensivo (y por extensión, que muestren una posición favorable a un contenido ofensivo), lo cual constatamos mediante una inspección manual a algunos conjuntos de narrativas generados que obtuvieron puntuaciones intermedias en las categorías de Ofensividad y Posición, cuando en verdad no habían generado ninguna contranarrativa ofensiva o que compartiese la posición del *tweet* original. Esto nos permitió realizar una evaluación de los sesgos y errores de nuestros evaluadores automáticos, que si bien mantienen una fuerte correlación con la realidad (todos los modelos que efectivamente no producen contenido ofensivo obtienen un puntaje alto de los evaluadores automáticos en esa misma categoría), se ven influenciados por propiedades extrínsecas a los criterios propuestos, como el estilo, las palabras utilizadas o incluso la longitud de la contranarrativa generada.

Sumado a esto, los criterios propuestos no permiten realizar una comparación con un nivel mayor de detalle entre las distintas contranarrativas que permita, por ejemplo, construir un *ranking* o definir con claridad si un modelo genera mejores contranarrativas que otro, sino que funcionan como un filtro que permite descartar contranarrativas no aceptables para ser puestas en producción y/o modelos que las generan con frecuencia. En este sentido, esta técnica de evaluación automática nos permitió identificar la totalidad de casos en los que las contranarrativas generadas no tenían una calidad suficiente, generando de todos modos, algunos falsos negativos (modelos identificados como generadores de contranarrativas inaceptables que en verdad no lo eran) pero ningún falso positivo (es decir, todos los modelos que obtuvieron un puntaje alto generaban efectivamente contranarrativas de calidad).

7.2 Trabajo Futuro

7.2.1 Enfoque mixto entre componentes y esquemas argumentativos e incremento del conjunto de datos anotados

Originalmente, decidimos descartar los esquemas argumentativos de Walton et al. [2008] debido a que, al ser un enfoque *bottom-up* en el que los esquemas se construyen a partir de la observación de argumentos reales en determinado dominio, a medida que el conjunto de datos a etiquetar crece, aparecen nuevos “tipos” de argumentos y el conjunto de esquemas crece también, a veces debiendo proponer nuevos esquemas para “cubrir” casos con muy pocos representantes y con argumentos “singulares” que conforman su propio tipo. En efecto, como parte de este trabajo de investigación, realizamos una prueba piloto en la cual propusimos un conjunto de 6 esquemas argumentativos para etiquetar 100 ejemplos en 5 tandas de 20 ejemplos cada una y como resultado terminamos agregando luego de cada tanda, una mayor cantidad de esquemas hasta terminar con un total de 17 esquemas y así y todo, fue difícil y forzado asignar algunos ejemplos a alguno de estos esquemas.

Sin embargo, luego de realizar el proceso de anotación con el esquema basado en la Tabla Periódica de los Argumentos de Wagemans, consideramos, a través de la observación de los datos, que debiera ser posible incorporar al proceso de anotación un conjunto acotado de esquemas argumentativos previamente definidos, siempre y cuando se abandone la pretensión de cobertura total (es decir, puede haber ejemplos que no pertenezcan a ningún esquema). Nuestra propuesta implica combinar estos esquemas acotados con los componentes argumentativos definidos en este trabajo a partir de Wagemans [2016]. Para evaluar la factibilidad de este proceso de anotación, es necesario realizar un estudio del conjunto de datos a anotar, intentando extraer los argumentos más frecuentes, los cuales serán convertidos en esquemas. Antes de realizar el proceso de anotación por completo, planificamos realizar la anotación de una muestra y evaluar la proporción de ejemplos “cubiertos” por alguno de los esquemas propuestos, evaluando luego la posibilidad de agregar más esquemas si muchos ejemplos no se asocian con ninguno, pero a la vez procurando que la cantidad a utilizar se mantenga limitada. Así como algunos ejemplos pueden no caer dentro de ninguno de los esquemas propuestos, otros pueden caer en más de uno a la vez.

Este nuevo conjunto de datos plantearía, por lo tanto, una nueva tarea: la identificación automática de el o los esquemas argumentativos. Con este fin, planteamos también entrenar un clasificador por cada esquema propuesto para realizar una clasificación binaria de las secuencias de texto con el fin de predecir si pertenecen al esquema o no. También planteamos entrenar un único clasificador que identifique todos los esquemas en simultáneo. Una variante de este experimento puede concebirse también incorporando la información de los componentes argumentativos al clasificador que predice los esquemas para ver cómo afectan estos a la tarea.

El experimento de la sección 4.2.6 plantea también que, si bien para un conjunto de los componentes propuestos, la cantidad de datos anotados es suficiente para que los modelos utilizados alcancen un rendimiento óptimo dentro de las posibilidades de la tarea, para otros componentes como la Propiedad, el Pivot y el Tipo de Justificación, una mayor cantidad de datos anotados permitiría una mejora en el rendimiento. A su vez, los análisis de desacuerdos entre anotadores, por un lado, y de errores en los modelos de clasificación, por el otro, sugieren también la necesidad de redefinir el Pivot, permitiendo anotar, po-

tencialmente, relaciones entre premisas señaladas por más de un par de elementos.

Estas dos observaciones apuntan a la necesidad de aumentar la cantidad de datos anotados con componentes argumentativos, ya sea mediante una nueva ronda de anotación con anotadores humanos o con alguna técnica de *Data Augmentation*. Esta nueva ronda de anotación deberá incorporar también la información sobre los esquemas argumentativos que proponemos.

7.2.2 Incorporar las máximas de relevancia de Grice en las categorías de evaluación

En el último tiempo, con el avance reciente de los modelos generativos, los Grandes Modelos de Lenguaje (LLM, por sus siglas en inglés) comenzaron a ser utilizados como jueces para evaluar texto generado por otros modelos más pequeños, bajo la asunción de que al realizar “mejor” la tarea, son capaces de erigirse como evaluadores. Sin embargo, el trabajo de Hinton and Wagemans [2023a] destaca cómo los modelos de lenguaje tienden a producir argumentos correctos desde el punto de vista formal, pero de baja calidad en términos de contenido y persuasividad. Es esperable, entonces, que los modelos de lenguaje sean capaces de juzgar la correctitud formal de un argumento (entre las cuales podríamos establecer quizás un paralelismo con las categorías de Ofensividad, Posición y Felicidad, que conforman los requisitos mínimos de aceptabilidad de una contra-narrativa), pero se ven limitados a la hora de reconocer aspectos más subjetivos y matizados como la persuasión, un concepto elusivo incluso para expertos humanos.

Al analizar nuestro sistema de evaluación, observamos que tres de las categorías propuestas, la Ofensividad, la Posición y la Felicidad, obtienen un rendimiento satisfactorio al ser automatizadas utilizando un conjunto pequeño de ejemplos. Sin embargo, como señalamos en la sección 6.3.3.2, los modelos de lenguaje más nuevos (especialmente los de mayor tamaño) se encuentran alineados a un conjunto de preferencias del usuario que en muy pocos casos permite generar contenido ofensivo (y por consecuencia, tampoco generar un mensaje que muestre apoyo a un contenido ofensivo). Si bien es algo que puede ocurrir y en ese sentido es importante el resultado obtenido con los evaluadores automáticos, las categorías de Ofensividad, Posición y Felicidad se plantean como una primera línea de evaluación definiendo un criterio básico de aceptabilidad, mientras que en una segunda línea de evaluación, la Informatividad define un segundo criterio que separa las contranarrativas más elaboradas y específicas de aquellas más genéricas y simples.

La clasificación automática de contra-narrativas según la categoría de Informatividad, sin embargo, muestra un desempeño más inestable, que para algunos pocos modelos llega arriba de .5 puntos de F1, mientras que para otros queda cerca de .3 puntos de F1, lo que puede explicarse tanto por una posible ambigüedad en la definición de la categoría como por el hecho de que el modelo fue entrenado con una menor cantidad de ejemplos que los utilizados para el resto de las categorías, ya que sólo se anotó la informatividad si la contranarrativa obtuvo el máximo valor en la categoría de Posición. Al analizar las predicciones realizadas sobre modelos de mayor tamaño, con un mejor rendimiento en la tarea de generación de contranarrativas que Flan-T5, observamos que son pocos los casos en los que una contra-narrativa Aceptable no obtiene el máximo valor de Informatividad. En general, las contra-narrativas que obtienen máximo valor en Ofensividad, Posición y Felicidad, obtienen también el máximo valor en Informatividad. Esto plantea un problema a la hora de intentar establecer una jerarquía o una diferenciación entre contra-narrativas Aceptables, quedando en los hechos trunca esta clasificación de segundo orden.

Con el objetivo de establecer una clasificación jerárquica más clara, Zubiaga et al. [2024] proponen un método de *ranking* utilizando Grandes Modelos de Lenguaje basado en comparar contranarrativas generadas por distintos modelos de a pares, organizados en el estilo de un torneo. Para realizar la comparación se utiliza *JudgeLM* Zhu et al. [2023], un modelo para evaluar generación abierta de textos, que soporta comparar respuestas individuales contra una respuesta de referencia, tal como hacen en este sistema de evaluación, pero también soporta comparar muchas respuestas en simultáneo. Utilizando este modelo, los autores proponen un sistema de evaluación que consiste en armar triplas compuestas por el discurso de odio original, la respuesta elaborada por seres humanos y la respuesta generada por el modelo de lenguaje, y comparar, para cada modelo, cada una de estas triplas con todas las demás que hayan sido generadas por todo el resto de los modelos a evaluar. Para cada comparación, si *JudgeLM* elige la tripla del modelo que está siendo comparado, se suma un punto. El *ranking* entre modelos se define en función del puntaje obtenido por cada modelo.

Si bien este mecanismo permite, hasta cierto punto, un nivel de granularidad más fina con respecto a nuestro sistema (aunque en nuestro caso el puntaje agregado también sirve para comparar modelos), el resultado obtenido tiene un nivel bajo de interpretabilidad: es difícil explicar por qué un modelo prefiere una contranarrativa por sobre la otra. Esto genera dificultades a la hora de evaluar el sistema de evaluación y de reconocer potenciales problemas, como hicimos nosotros en este trabajo al descubrir la existencia de un sesgo de los evaluadores hacia las contranarrativas generadas por modelos ajustados. En este caso, pudimos lograr esto gracias a que cada contranarrativa tiene un puntaje interpretable en sí mismo, y por lo tanto una inspección manual permite contrastar un determinado resultado de una clasificación automática con criterio humano (por ejemplo, reconocer que una contranarrativa clasificada como Ofensiva o Ambigua no lo es), algo que no resulta posible con la propuesta de Zubiaga et al. [2024]. Otra ventaja de nuestra propuesta de evaluación es que puede ser utilizada para evaluar tanto contranarrativas generadas por una máquina como por un ser humano, ya que no necesita alimentar los modelos con triplas que tengan un *gold standar*, además de la contranarrativa generada.

Considerando este aspecto, optamos por desarrollar nuestras categorías de evaluación con el objetivo de refinar esta clasificación de segunda línea y permitir una distinción más fina entre contranarrativas Aceptables, pero sin perder la interpretabilidad de nuestra clasificación. Para refinar nuestras categorías, tomamos como base las máximas conversacionales elaboradas por Grice [1975].

Las máximas conversacionales de Grice forman parte de su teoría del principio de cooperación, que establece que los interlocutores en una conversación suelen colaborar para que el intercambio comunicativo sea efectivo. Estas máximas son normas implícitas que guían la interacción lingüística y aseguran que la comunicación sea clara, relevante y eficiente. Aplicadas a la generación de respuestas automáticas en un contexto dialógico-argumentativo, las máximas permiten identificar posibles fallas en los modelos de lenguaje que afecten su capacidad de colaborar en el intercambio con los usuarios y potenciales observadores.

Grice identificó cuatro máximas principales:

- Máxima de cantidad: Proporciona la cantidad de información necesaria, ni más ni menos de lo que se requiere para el propósito de la conversación.
- Máxima de calidad: Asegúrate de que lo que dices sea verdadero, evitando afirma-

ciones falsas o sin evidencia suficiente.

- Máxima de relevancia: Sé pertinente, contribuyendo con información relacionada con el tema de la conversación.
- Máxima de manera: Expresa tus ideas con claridad y orden, evitando ambigüedades, vaguedades o confusión innecesaria.

Estas máximas no son reglas rígidas, sino guías que los hablantes pueden violar de manera intencional o no, lo que puede dar lugar a implicaturas conversacionales, es decir, significados implícitos que el interlocutor debe interpretar a partir del contexto. En una conversación abierta entre seres humanos donde se asume que ambos interlocutores son hablantes competentes, un silencio, una respuesta no relacionada al tema de conversación o un cambio en el tono, posiblemente tengan un significado implícito que los participantes de la conversación pueden inferir en mayor o menor medida. Si un hablante hace una pregunta incómoda, por ejemplo, y el otro interlocutor no responde o cambia de tema de manera abrupta, el emisor del mensaje “incómodo” puede interpretar que el receptor no quiere contestar o no se siente cómodo con que la conversación prosiga por ese lado. Cuando lidiamos con modelos de lenguaje que intervienen en contextos definidos y acotados como puede ser desescalar una conversación de odio, sin embargo, estas sutilezas no se interpretan de la misma manera. Si un usuario odiente genera una pregunta con un sesgo discriminatorio y un modelo no responde o cambia de tema, eso significa que el modelo de lenguaje no funciona como debería.

Nuestro nuevo esquema de evaluación mantiene las categorías de Ofensividad y Posición mientras que agrega a estas una categoría por cada máxima conversacional. De este modo, la Informatividad se divide en tres nuevas categorías: la Relevancia, la Cantidad y la Calidad. La Felicidad, por su parte, queda subsumida en la categoría de Modo, que involucra de alguna forma la coherencia interna del mensaje y su correctitud formal (sintáctica y semántica). El resultado es un nuevo sistema de evaluación con seis categorías. En el apéndice A presentamos nuestra nueva propuesta de protocolo de anotación con los valores posibles de cada categoría.

Habiendo definido este protocolo, como trabajo futuro vamos a realizar un nuevo proceso de anotación con las nuevas categorías de evaluación. Nos proponemos hacer un *dataset* con cinco mil pares de discurso de odio y contranarrativas evaluadas por tres anotadores cada una e incluyendo contranarrativas generadas por una mayor variedad de modelos.

7.2.3 Evaluación a través de un sistema de *rankings*

En el contexto de este trabajo desarrollamos un método de evaluación basado en cuatro categorías que puede ser utilizado tanto de manera manual con anotadores humanos como de forma automática mediante el uso de modelos de lenguaje preentrenados con anotaciones humanas. Este mecanismo de evaluación tiene entre sus principales ventajas el tener un alto grado de interpretabilidad, al mismo tiempo que permite hacer una clasificación general de las contranarrativas generadas y de sus modelos, trazando a su vez una línea divisoria entre aquellas que definimos como “Buenas”, aquellas que definimos como “Aceptables” y aquellas que no lo son. Sin embargo, este mecanismo de evaluación parece ser insuficiente para realizar una evaluación más fina, que permita definir, por ejemplo, entre dos contranarrativas “Buenas”, cuál puede ser una mejor respuesta dado un discurso

de odio particular. En la redefinición de las categorías de evaluación propuesta en la sección 7.2.2 es un intento de superar esa dificultad.

Una alternativa es utilizar un modelo de preferencias que permita establecer un sistema cerrado de preferencias (y por lo tanto, un *ranking*) a partir de poder determinar, entre dos contranarrativas, cual es la mejor. Un sistema de este tipo basado en la votación sin criterios explícitos define de forma empírica una única propiedad deseable que podríamos identificar con la “persuasión”, cuyo valor está directamente asociado a la cantidad de votos. Si bien consideramos que un sistema de este tipo sacrifica un poco de interpretabilidad, Zubiaga et al. [2024] han demostrado que pueden obtenerse buenos resultados de evaluaciones que correlacionan con las preferencias humanas. Los autores hacen público su modelo y un repositorio con el código necesario para realizar comparaciones⁶, por lo que nos proponemos utilizarlo para evaluar las contranarrativas generadas por los modelos utilizados en la sección 6.3.1 y comparar esta evaluación con la realizada por nuestros métodos basados en categorías intentando descubrir si existe alguna correlación.

7.2.4 Evaluar distintas condiciones de generación

El desarrollo de herramientas de evaluación automática de generación de texto y, particularmente, de contra-narrativas para el discurso de odio contra inmigrantes, permite masificar los experimentos de generación a una nueva escala. Esto nos permite, a su vez, experimentar sobre una multiplicidad de contextos y técnicas de generación que, de otro modo, hubieran sido imposibles de evaluar. Nos planteamos, por lo tanto, evaluar la generación de contra-narrativas utilizando distintos prefijos en la entrada de los modelos, tanto para el ajuste fino como para la técnica de *fewshot*. Estos distintos prefijos pueden tener distintas longitudes y niveles de detalle de la tarea a desarrollar, así como también distintos niveles de explicación sobre cómo utilizar la información argumentativa, o distintas indicaciones sobre el modo o el estilo de la respuesta (planteando, por ejemplo, respuestas más cordiales o más agresivas).

Al mismo tiempo, consideramos que también es una oportunidad para evaluar distintas técnicas de generación. Mientras que en el contexto de este trabajo realizamos una evaluación de una muestra pequeña generada con cuatro técnicas distintas (generación golosa, con muestreo y temperatura, con muestreo y top p y con *Beam Search*) con el fin de elegir la que mejor rendimiento demostrara y luego fijar esa técnica para el resto del trabajo (de la cual pueden verse los resultados en la tabla 5.3), los evaluadores automáticos nos permiten extender la evaluación a las cuatro técnicas propuestas. Considerando, además, que la técnica elegida para este trabajo (*Beam Search*) prioriza la adecuación a las probabilidades conjuntas de generación por sobre la posibilidad de variabilidad, sería interesante también, para evaluar las diferencias entre las técnicas de generación, contar con alguna métrica que permita dar cuenta de la “creatividad” o “variabilidad” de las respuestas generadas.

gar publicacio-
onde el con-
o de esta tesis
publicado

⁶ <https://github.com/hitz-zentroa/cn-eval>

8 Limitaciones y consideraciones éticas

En primer lugar, queremos dejar claro que las anotaciones humanas presentadas aquí son el resultado de la subjetividad de los anotadores. Aunque han sido instruidos mediante un manual y sesiones de capacitación, aún existen variaciones significativas entre interpretaciones, y futuros investigadores podrían asignar categorías diferentes a los ejemplos.

Asimismo, es importante señalar que los procedimientos automáticos obtenidos son propensos a errores y deben usarse críticamente, prestando atención a posibles fallos y a cómo estos pueden afectar a los usuarios, a los grupos protegidos y a la sociedad.

También es importante destacar que el corpus utilizado para esta investigación es pequeño, especialmente el de evaluación de contra narrativas y la parte en español de ASOHMO y CONEAS, por lo que los resultados presentados en este artículo deben considerarse indicativos. Se debería obtener y anotar una muestra más grande para obtener resultados estadísticamente más significativos.

Los hallazgos de esta investigación pueden potencialmente potenciar el desarrollo y la mejora de modelos de lenguaje y sistemas de chatbots. Sin embargo, enfatizamos la importancia de un uso y una aplicación responsables de nuestros hallazgos. Es esencial garantizar que los componentes argumentativos identificados se utilicen de manera que promuevan un uso racional y no contribuyan a la difusión de discursos de odio o retórica dañina. Animamos a los investigadores y desarrolladores a considerar las implicaciones éticas y el impacto social de incorporar el análisis argumentativo en sus sistemas.

Si bien el discurso de odio utilizado proviene de usuarios reales de la red social Twitter, los datos han sido adecuadamente anonimizados por los creadores originales del corpus Hateval.

El estudio del discurso de odio implica analizar y procesar contenido que puede ser ofensivo, dañino o de otra manera objetable. Reconocemos el impacto potencial de trabajar con dicho contenido y hemos tomado medidas para garantizar el bienestar del equipo de investigación involucrado. Hemos proporcionado directrices y capacitación completas a nuestros anotadores para mitigar cualquier posible angustia emocional o daño que pueda surgir de la exposición al discurso de odio. Además, hemos implementado estrictas medidas para prevenir la difusión o propagación adicional del discurso de odio durante el proceso de investigación.

Finalmente, no hemos llevado a cabo específicamente un estudio sobre los sesgos dentro del corpus, las anotaciones o los procedimientos automáticos inferidos de este, ni sobre los modelos de lenguaje utilizados. Advertimos a los investigadores que utilicen estas herramientas y recursos que pueden encontrar sesgos no detectados, y animamos a realizar más investigaciones para caracterizarlos.

Acknowledgments

La anotación de los componentes argumentativos y de las contranarrativas fue realizada con la herramienta de anotación **BRAT** [Stenetorp et al., 2012].

Este trabajo utilizó recursos computacionales del CCAD de la Universidad Nacional de Córdoba (<https://ccad.unc.edu.ar/>), que forman parte del SNCAD del MinCyT de la República Argentina.

Esta tesis fue parcialmente financiada por la Secretaría de Investigación Científica y Tecnológica FCEN-UBA (RESCS-2020-345-E-UBA-REC), CONICET bajo el PIP (11220200101408CO), Agencia Nacional de Promoción Científica y Tecnológica, Argentina bajo el proyecto PICT-2018-0475 (PRH-2014-0007), PICT-2020- SERIEA-01481, y el NAACL Regional Americas Fund (2022).

Bibliografía

- Hare I. Weinstein J. Post, R. Hate speech. pages 123–138. Oxford Press, 2009. URL <https://unesdoc.unesco.org/ark:/48223/pf0000233231>.
- Susan Benesch. Countering dangerous speech: New ideas for genocide prevention. 2014a. URL <https://api.semanticscholar.org/CorpusID:232815741>.
- Imran Awan and Irene Zempi. The affinity between online and offline anti-muslim hate crime: Dynamics and impacts. *Aggression and Violent Behavior*, 27:1–8, 2016. ISSN 1359-1789. doi: <https://doi.org/10.1016/j.avb.2016.02.001>. URL <https://www.sciencedirect.com/science/article/pii/S1359178916300015>.
- Matthew L Williams, Pete Burnap, Amir Javed, Han Liu, and Sefa Ozalp. Hate in the Machine: Anti-Black and Anti-Muslim Social Media Posts as Predictors of Offline Racially and Religiously Aggravated Crime. *The British Journal of Criminology*, 60(1):93–117, 07 2019. ISSN 0007-0955. doi: 10.1093/bjc/azz049. URL <https://doi.org/10.1093/bjc/azz049>.
- Karsten Müller and Carlo Schwarz. Fanning the flames of hate: Social media and hate crime. *SSRN Electronic Journal*, 01 2017. doi: 10.2139/ssrn.3082972.
- Danielle Keats Citron and Helen Norton. Intermediaries and hate speech: Fostering digital citizenship for our information age. *BUL Rev.*, 91:1435, 2011.
- Ariadna Matamoros-Fernández. Platformed racism: the mediation and circulation of an australian race-based controversy on twitter, facebook and youtube. *Information, Communication & Society*, 20(6):930–946, 2017. doi: 10.1080/1369118X.2017.1293130. URL <https://doi.org/10.1080/1369118X.2017.1293130>.
- Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. Countering online hate speech. 2015. URL <https://unesdoc.unesco.org/ark:/48223/pf0000233231>.
- Susan Benesch. Words as weapons. *World Policy Journal*, 29:7–12, 03 2012. doi: 10.2307/41510486.
- Zeera Talat and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *North American Chapter of the Association for Computational Linguistics*, 2016. URL <https://api.semanticscholar.org/CorpusID:1721388>.
- Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of Third Workshop on Abusive Language Online*, August 2019.

- Susan Benesch. Countering dangerous speech: New ideas for genocide prevention, 2014b. URL <https://ssrn.com/abstract=3686876>.
- Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *ACL*, July 2019.
- Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. pages 54–63, June 2019. doi: 10.18653/v1/S19-2007. URL <https://aclanthology.org/S19-2007>.
- Aatman Vaidya, Seema Nagar, and Amit A. Nanavati. Analysing the spread of toxicity on twitter. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, CODS-COMAD '24, page 118–126, New York, NY, USA, 2024. Association for Computing Machinery. ISBN 9798400716348. doi: 10.1145/3632410.3632436. URL <https://doi.org/10.1145/3632410.3632436>.
- Luke Thorburn and Ariel Kruger. Optimizing language models for argumentative reasoning. In *Proceedings of the 1st Workshop on Argumentation Machine Learning co-located with 9th International Conference on Computational Models of Argument (COMMA 2022)*, 2022.
- Martin Hinton and Jean H. M. Wagemans. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the GPT-3 AI text generator. *Argument Comput.*, 14(1):59–74, 2023a. doi: 10.3233/AAC-210026. URL <https://doi.org/10.3233/AAC-210026>.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022. URL <https://arxiv.org/abs/2203.02155>.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks, 2021. URL <https://arxiv.org/abs/2005.11401>.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024. URL <https://arxiv.org/abs/2312.10997>.
- Jean H. M. Wagemans. Constructing a periodic table of arguments. In *Proceedings of 11th International Conference of the Ontario Society for the Study of Argumentation*, 2016.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, July 2002.

- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, July 2004.
- Ehud Reiter and Anja Belz. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558, December 2009. doi: 10.1162/coli.2009.35.4.35405. URL <https://aclanthology.org/J09-4008>.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert, 2020. URL <https://arxiv.org/abs/1904.09675>.
- Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance, 2019. URL <https://arxiv.org/abs/1909.02622>.
- Rishabh Gupta, Shaily Desai, Manvi Goel, Anil Bandhakavi, Tanmoy Chakraborty, and Md. Shad Akhtar. Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5792–5809, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.318. URL <https://aclanthology.org/2023.acl-long.318>.
- Simone Teufel, Jean Carletta, and Marc Moens. An annotation scheme for discourse-level argumentation in research articles. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 110–117, Bergen, Norway, June 1999. Association for Computational Linguistics. URL <https://aclanthology.org/E99-1015>.
- Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee. URL <https://aclanthology.org/C16-1324>.
- Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 2 edition, 2003. doi: 10.1017/CBO9780511840005.
- Christian Stab and Iryna Gurevych. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1006. URL <https://aclanthology.org/D14-1006>.
- Ivan Habernal and Iryna Gurevych. Argumentation mining in user-generated web discourse. *Computational Linguistics*, 43(1):125–179, April 2017. doi: 10.1162/COLLa.00276. URL <https://aclanthology.org/J17-1004>.
- Douglas Walton, Christopher Reed, and Fabrizio Macagno. *Argumentation Schemes*. CUP, 2008.

- Fabrizio Macagno, Douglas Walton, and Chris Reed. *Argumentation Schemes*, pages 517–574. 08 2018.
- Anna Lindahl, Lars Borin, and Jacobo Rouces. Towards assessing argumentation annotation - a first step. In *Proceedings of the 6th Workshop on Argument Mining*, pages 177–186, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4520. URL <https://aclanthology.org/W19-4520>.
- Elena Musi, Debanjan Ghosh, and Smaranda Muresan. Towards feasible guidelines for the annotation of argument schemes. 08 2016. doi: 10.18653/v1/W16-2810.
- Katie Atkinson and Trevor Bench-Capon. Taking account of the actions of others in value-based reasoning. *Artificial Intelligence*, 254:1–20, 2018. ISSN 0004-3702. doi: <https://doi.org/10.1016/j.artint.2017.09.002>.
- Nadin Kökciyan, Isabel Sassoon, Peter Young, Martin Chapman, Talya Porat, Mark Ashworth, Vasa Curcin, Sanjay Modgil, Simon Parsons, and Elizabeth Sklar. Towards an argumentation system for supporting patients in self-managing their chronic conditions. 01 2018.
- Jean Wagemans. Four basic argument forms. *Research in Language*, 17:57–69, 03 2019. doi: 10.2478/rela-2019-0005.
- Jacky Visser, John Lawrence, Chris Reed, Jean Wagemans, and Douglas Walton. Annotating argument schemes. *Argumentation*, 35, 03 2021. doi: 10.1007/s10503-020-09519-x.
- Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth M. Belding, and William Yang Wang. A benchmark dataset for learning to intervene in online hate speech. *CoRR*, abs/1909.04251, 2019.
- Caleb Ziems, Bing He, Sandeep Soni, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis. 05 2020.
- Serra Sinem Tekiroğlu, Yi-Ling Chung, and Marco Guerini. Generating counter narratives against online hate speech: Data and strategies. In *ACL*, July 2020.
- Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğlu, and Marco Guerini. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *ACK*, August 2021.
- Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *EMNLP*, December 2022.
- Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the ACL-IJCNLP 2021*, August 2021a.
- Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

- Zewen Chi, Li Dong, Furu Wei, Wenhui Wang, Xian-Ling Mao, and Heyan Huang. Cross-lingual natural language generation via pre-training. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:7570–7577, 04 2020. doi: 10.1609/aaai.v34i05.6256.
- Mana Ashida and Mamoru Komachi. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of Sixth Workshop on Online Abuse and Harms (WOAH)*, July 2022.
- M.E. Vallecillo-Rodríguez, A. Montejo Ráez, and M.T. Martín-Valdivia. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71, 2023.
- Yi-Ling Chung, Marco Guerini, and Rodrigo Agerri. Multilingual counter narrative type classification. In *Proceedings of the 8th Workshop on Argument Mining*, pages 125–132, Punta Cana, Dominican Republic, November 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.argmining-1.12. URL <https://aclanthology.org/2021.argmining-1.12>.
- Susan Benesch, Derek Ruths, Kelly P Dillon, Haji Mohammad Saleem, and Lucas Wright. Counterspeech on twitter: A field study. *Dangerous Speech Project*. Available at: <https://dangerousspeech.org/counterspeech-ontwitter-a-field-study/>, 2016.
- Irune Zubiaga, Aitor Soroa, and Rodrigo Agerri. A llm-based ranking method for the evaluation of automatic counter-narrative generation, 2024. URL <https://arxiv.org/abs/2406.15227>.
- Yinfei Yang, Daniel Cer, Amin Ahmad, Mandy Guo, Jax Law, Noah Constant, Gustavo Hernandez Abrego, Steve Yuan, Chris Tar, Yun-Hsuan Sung, Brian Strope, and Ray Kurzweil. Multilingual universal sentence encoder for semantic retrieval, 2019.
- Martin Hinton and Jean H. M. Wagemans. How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the gpt-3 ai text generator. *Argument and Computation*, 14(1):59–74, 2023b. doi: 10.3233/aac-210026.
- Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL <https://aclanthology.org/W05-0909>.
- Ke Wang and Xiaojun Wan. Sentigan: Generating sentimental texts via mixture adversarial networks. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 4446–4452. International Joint Conferences on Artificial Intelligence Organization, 7 2018. doi: 10.24963/ijcai.2018/618. URL <https://doi.org/10.24963/ijcai.2018/618>.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are scalable judges, 2023. URL <https://arxiv.org/abs/2310.17631>.
- Jean H.M. Wagemans. Argument type identification procedure (atip) - version 4. *Periodic Table of Arguments*, 2021.

2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1245. URL <https://aclanthology.org/D17-1245>.
- J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.
- Milagro Teruel, Cristian Cardellino, Fernando Cardellino, Laura Alonso Alemany, and Serena Villata. Increasing argument annotation reproducibility by using inter-annotator agreement to improve guidelines. May 2018. URL <https://aclanthology.org/L18-1640>.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019. URL <https://arxiv.org/abs/1907.11692>.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2018. URL <https://arxiv.org/abs/1810.04805>.
- Yukun Zhu, Ryan Kiros, Richard Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. 2015.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale, 2019. URL <https://arxiv.org/abs/1911.02116>.
- Teven Le Scao et. al. Bloom: A 176b-parameter open-access multilingual language model, 2023. URL <https://arxiv.org/abs/2211.05100>.
- B. Wang and A. Komatsuzaki. Gpt-j-6b: A 6 billion parameter autoregressive language model. 2021. URL <https://github.com/kingoflolz/mesh-transformer-jax>.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018. URL <https://d4mucfpsywv.cloudfront.net/better-language-models/language-models.pdf>.
- Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. 2022. URL <https://arxiv.org/abs/2210.11416>.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark

- Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL <https://arxiv.org/abs/2005.14165>.
- Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. Finetuned language models are zero-shot learners, 2022. URL <https://arxiv.org/abs/2109.01652>.
- Rensis Likert. A technique for the measurement of attitudes, 1932.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019. URL <https://arxiv.org/abs/1908.10084>.
- Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers, 2020. URL <https://arxiv.org/abs/2002.10957>.
- Iftitahu Ni'mah, Meng Fang, Vlado Menkovski, and Mykola Pechenizkiy. Nlg evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Annual Meeting of the Association for Computational Linguistics*, 2023. URL <https://api.semanticscholar.org/CorpusID:258685435>.
- José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*, 2020.
- Javier De la Rosa, Eduardo G. Ponferrada, Manu Romero, Paulo Villegas, Pablo González de Prado Salas, and María Grandury. Bertin: Efficient pre-training of a spanish language model using perplexity sampling. *Procesamiento del Lenguaje Natural*, 68(0):13–23, 2022. ISSN 1989-7553. URL <http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6403>.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding, 2020. URL <https://arxiv.org/abs/2004.09297>.
- Abhimanyu Dubey et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. Qwen2 technical report, 2024. URL <https://arxiv.org/abs/2407.10671>.

-
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. Mistral 7b, 2023. URL <https://arxiv.org/abs/2310.06825>.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL <https://arxiv.org/abs/2305.14314>.
- Herbert Paul Grice. Logic and conversation. In Herbert Paul Grice, editor, *Syntax and Semantics, Vol. 3*, pages 41–58. New York: Academic Press, 1975.

Apéndice

A Nuevo protocolo de evaluación de contranarrativas

Situación: En las redes sociales de tu comunidad, está creciendo el discurso xenófobo. Por ello, la comunidad ha coordinado una iniciativa para responder a los mensajes de odio contra inmigrantes. Dado que el volumen de mensajes es muy grande y el número de voluntarios es muy pequeño, se ha incluido un chatbot en el proceso para generar automáticamente respuestas. Sin embargo, este chatbot no siempre genera respuestas que contribuyen a contrarrestar el discurso de odio, por lo que deben revisarse y modificarse antes de publicarlas. Además, los otros voluntarios que redactan respuestas a veces se pierden y escriben respuestas que pueden tener el efecto opuesto al deseado. Para evitar publicar respuestas que puedan ser perjudiciales, organizamos un sistema para revisar tanto las respuestas humanas como las automáticas. En este sistema, debes revisar las respuestas a los mensajes de odio y verificar que sean constructivas y contribuyan a reducir la tensión, fomentando la reflexión y la comprensión.

Con este objetivo, para cada par de discurso de odio y respuesta, te pedimos que contestes las siguientes preguntas con alguna de las tres opciones posibles:

- ¿La contranarrativa se opone al discurso de odio contra inmigrantes? (Posición)
 1. Está a favor del discurso de odio.
 2. Es poco clara / ambigua.
 3. Se opone al discurso de odio.
- ¿Es ofensiva? (Ofensividad)
 1. Es ofensiva.
 2. Poco clara / ambigua.
 3. No es ofensiva.
- ¿La información de la contranarrativa es relevante y específica para responder al mensaje de odio original? (Relevancia)
 1. Repite información proporcionada por el discurso de odio o proporciona información fuera de tema.
 2. Proporciona información nueva y relacionada, pero no muy relevante para contestar el mensaje de odio o combinada con información que no es relevante.
 3. Proporciona información muy relevante para contradecir el mensaje de odio al que responde.

- ¿La contranarrativa proporciona suficiente información (aunque no demasiada) para que la respuesta contribuya a una mejor comprensión del tema o invite a la reflexión? (Cantidad)
 1. No proporciona suficiente información, es vaga, genérica o demasiado extensa.
 2. Proporciona suficiente información, pero no lo suficiente para responder adecuadamente al mensaje de odio, o no es lo suficientemente específica.
 3. Proporciona información concreta y concisa, justo lo necesario para sustentar el argumento contra el mensaje de odio al que responde.
- ¿La contranarrativa proporciona información verídica (o verificable), consistente con las narrativas existentes contrarias al discurso xenófobo? (Calidad)
 1. La información proporcionada no se puede verificar de ninguna manera o no es consistente con la narrativa utilizada para defender al colectivo atacado.
 2. La información proporcionada no necesita ser verificada (es una verdad universal, es una opinión).
 3. La información proporcionada es intuitivamente verídica y/o es consistente con la narrativa existente contra el discurso de odio.
- ¿El texto de la contranarrativa es fácil de entender y parece bien escrito? (Modo)
 1. El texto no es comprensible.
 2. Es comprensible, pero suena extraño, poco natural u obscuro.
 3. El texto fluye muy bien, es fácil de leer y entender.