Universidad de Buenos Aires
Facultad de Ciencias Exactas y Naturales
Departamento de Computación

Técnicas y recursos para la detección automática de lenguaje discriminatorio en redes sociales

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

Juan Manuel Pérez

Director:	Dr. Franco Luque
Director Adjunto:	Dr. Agustín Gravano
Consejero de Estudios:	Dr. Diego Fernández Slezak
Lugar de Trabajo:	Departamento de Computación Facultad de Cs. Exactas y Naturales Universidad de Buenos Aires

Buenos Aires, 2022
Fecha de defensa: 8 de junio de 2022

TÉCNICAS Y RECURSOS PARA LA DETECCIÓN AUTOMÁTICA DE LENGUAJE DISCRIMINATORIO EN REDES SOCIALES

El discurso discriminatorio (también conocido como discurso de odio) puede describirse como aquel discurso en clave de intenso aborrecimiento, denigración y enemistad que ataca a un individuo o un grupo de individuos por poseer –o aparentar poseer– cierta característica protegida por tratados internacionales como el sexo, el género, la etnia, etc. En los últimos años, este tipo de discurso ha tomado gran relevancia en redes sociales y otros medios virtuales debido a su intensidad y a su relación con actos violentos contra miembros de estos grupos. A raíz de esto, estados y organizaciones supranacionales como la Unión Europea han sancionado legislación que insta a las empresas de redes sociales a moderar y eliminar contenido discriminatorio, con particular foco en aquel que insta a la violencia física.

Debido a la enorme cantidad de contenido generado por usuarios en las redes sociales, es necesario contar con cierta automatización en esta tarea, bien para su análisis o para su moderación. Desde la óptica del procesamiento de lenguaje natural, la detección de discriminación puede entenderse como un problema de clasificación de texto: dado un texto generado por un usuario, predecir si es o no contenido discriminatorio. Así mismo, puede ser de interés predecir otras características: por ejemplo, si el texto contiene un llamado a la acción violenta, si está dirigido contra un individuo o un grupo, o el tipo de característica ofendida, entre otras.

Una de las limitaciones de los enfoques actuales para la detección del lenguaje discriminatorio es la falta de contexto en el mensaje. La mayoría de los estudios y recursos están hechos sobre datos fuera de contexto; es decir, mensajes aislados sin ningún tipo de contexto conversacional o del tema del cual se habla. Esto restringe la información disponible –tanto para un humano como para un sistema– para poder discernir si un texto social es discriminatorio. Otra información usualmente faltante es la característica atacada: es común que los datasets estén anotados de manera poco granular, no brindando información acerca de si la agresión es por motivos de sexo, género, clase social, etc. Por último, una limitación puntual del español es la poca disponibilidad de recursos para esta tarea.

En esta tesis pretendemos abordar algunas de las limitaciones marcadas. Por un lado,

analizamos el impacto de agregar contexto a la detección de lenguaje discriminatorio en redes sociales. Para ello, construimos un conjunt de datos de tweets en base a las respuestas de los usuarios a los posteos de medios periodísticos en Twitter. Esto nos permite obtener dos tipos de contextos: uno "conversacional" al tener una respuesta a un tweet anterior, y otro más extenso al obtener el texto de la noticia en cuestión. El corpus fue recolectado sobre noticias relacionadas a la pandemia de COVID-19, en idioma español mayormente en su variedad dialectal rioplatense y anotado por hablantes nativos de ese dialecto con un nuevo modelo de etiquetado, que es granular respecto de las características ofendidas.

Sobre los comentarios de este dataset realizamos experimentos de detección de discurso de odio planteando dos tareas: detección binaria del lenguaje discriminatorio, donde sólo predecimos una etiqueta binaria indicando presencia de lenguaje discriminatorio; y detección granular, donde predecimos las características ofendidas. Usando técnicas del estado del arte, obtuvimos mejoras significativas en ambas tareas al agregar contexto como entrada de cada instancia, tanto en su forma corta (sólo el titular/tweet de la noticia) como en su forma larga (titular y cuerpo de la noticia). Así mismo, observamos que un clasificador entrenado para la tarea granular mejora levemente su performance al ser evaluado para la tarea binaria, obviando los posibles errores de motivos discriminatorios. Combinando la adición de contexto y granularidad, un clasificador para la detección de lenguaje discriminatorio obtiene mejoras considerables sobre un BERT en español que sólo consume el texto del comentario.

Considerando la detección de discurso de odio dentro del área más abarcativa de clasificación de documentos en dominios sociales, analizamos también algunos aspectos generales de tareas relacionadas como el análisis de sentimiento y la detección de emociones, entre otras. En particular, analizamos el desempeño de varias técnicas modernas de representación al ser entrenadas en dominios sociales. Comúnmente, los modelos de representación son entrenados a partir de textos de dominios formales, como pueden ser Wikipedia u otras fuentes similares. En esta tesis observamos que –desde los word embeddings hasta los modelos pre-entrenados basados en transformers– las representaciones generadas son robustas y mejoran la performance en un conjunto de tareas de clasificación en textos sociales. Sobre los modelos pre-entrenados, estudiamos el impacto de entrenarlos desde cero en textos sociales o efectuar una adaptación a este dominio.

Todos los estudios y recursos presentados en esta tesis fueron realizados en el idio-

ma español. Como un objetivo secundario, pretendemos contribuir a mitigar la enorme asimetría de recursos existente en el área del procesamiento del lenguaje natural.

Palabras claves: Hate Speech, Natural Language Processing, Abusive Language Detection, Domain Adaptation, Social NLP.

TECHNIQUES AND RESOURCES FOR THE AUTOMATIC DETECTION OF HATE SPEECH IN SOCIAL NETWORKS

Hate speech can be described as speech containing intense hatred, denigration, and enmity that attacks an individual or a group of individuals because of possessing –or pretending to possess– any characteristic protected by international treaties such as gender, ethnicity, religion, language, among others. In recent years, this type of discourse has gained great relevance in social networks and other virtual media due to its intensity and its relationship with violent acts against members of these groups. As a result, states and supranational organizations –such as the European Union– have enacted legislation that urges social media companies to moderate and remove discriminatory content, with particular focus on that which promotes physical violence.

Due to the enormous amount of user-generated content on social media, it is necessary to have some degree of automation in this task, either for analysis or for moderation. From a natural language processing (NLP) perspective, hate speech detection can be understood as a text classification problem: given a text generated by a user, predict whether it is discriminatory content. Likewise, it may be of interest to predict other features: for example, if the text contains a call to violent action; if it is directed against an individual or a group; or the offended characteristic, among others.

One of the limitations of current approaches to hate speech detection is the lack of context. Most studies and resources are performed on data without context; that is, isolated messages without any type of conversational context or the topic being discussed. This restricts the information available –both for a human and for an automated system– to discern if a social text is hateful or not. Other information usually lacking is the offended characteristic: datasets are usually annotated with a low level of granularity, failing to provide information about whether the offending message attacks the individual or group due to their gender, social class, race, or whatsoever. Finally, a specific limitation of Spanish is the limited availability of resources for this task.

In this thesis, we intend to address some of the marked limitations. On the one hand, we analyze the impact of adding context to hate speech detection in social networks. To do this, we built a tweet dataset based on user responses to news media posts on Twitter.

This provided us two types of contexts: a conversational context, given by the tweet and its answer, and another context given by the text of the news in question. This dataset was collected on news related to the COVID-19 pandemic, in the Spanish language in its Rioplatense dialectal variety. Native speakers of this dialect annotated the comments with a novel labeling model that is granular regarding the offended characteristics.

Using this dataset, we carried out hate speech detection experiments, proposing two tasks: "binary" detection of discriminatory language, where we only predict a binary label indicating the presence of discriminatory language; and "granular" detection, where we predict the attacked characteristics (n-binary classification tasks at the same time). Using state-of-the-art techniques, we obtained significant improvements in both tasks by adding context as input for each instance, both in its short form (only the headline/tweet of the news article) and in its long-form (headline and body of the news article). We also observed that a classifier trained for the "granular" task slightly improves its performance when being evaluated for the "flat" task, ignoring possible errors of discriminatory motives. Combining the addition of context and granularity, a classifier for the detection of discriminatory language obtained considerable improvements over a BERT in Spanish that only consumes the text of the comment.

Considering hate speech detection within the most comprehensive area of document classification in social domains, we further explored some general aspects of related tasks such as sentiment analysis and emotion detection, among others. In particular, we analyzed the performance of various modern representation techniques when trained in social domains. Commonly, NLP researchers train representation models on texts from "formal" domains, such as Wikipedia or other similar sources. We observed that –from word embeddings to pre-trained models based on transformers– the representations generated are robust and improve performance in a set of classification tasks in social texts. On the pre-trained models, we studied the impact of training them from scratch in social texts versus performing domain-adaptation on the language models.

All of the studies and resources presented in this thesis were carried out in the Spanish language. As a secondary objective, we aim to mitigate the enormous asymmetry of resources in the area of NLP.

Keywords: Hate Speech, Natural Language Processing, Abusive Language Detection, Domain Adaptation, Social NLP.

AGRADECIMIENTOS

En primer lugar, quiero agradecer a mis directores Franco y Agustín, que guiaron mi trabajo y me formaron como investigador. Realizar un doctorado es una labor bastante dura, en el cual uno se encuentra muchas veces perdido en el camino. La función que ambos cumplieron guiándome en esos momentos de desorientación —pero dándome la libertad de elección en cada momento— ha sido fundamental para llegar hasta acá.

Quiero agradecer a todes mis compañeres del Laboratorio de Inteligencia Artificial Aplicada (LIAA) y del Departamento de Computación de Exactas UBA quienes me ayudaron a transitar este doctorado, compartiendo conocimiento, charlas — a veces simplemente catarsis. Edgar Altszyler, Pablo Brusco, Ramiro Gálvez, Bruno Bianchi, Damián Furman, Lara Gauder, Jazmín Vidal, y todos los que me falten en esta lista. A Viviana Cotik, que me ayudó de gran manera en los momentos más críticos de este trabajo.

A todes les integrantes del Proyecto Interdisciplinario de la UBA sobre marginaciones sociales (PIUBAMAS), que fueron fundamentales en los segmentos más importantes de esta tesis.

A mis compañeres de activismo y militancia, particularmente a los compañeros de la Asociación Gremial de Docentes de la UBA (AGD-UBA) y Jóvenes Científicxs Precarizados (JCP). Luchar por nuestros derechos y reconocimiento como trabajadores ha sido sin dudas parte de mi formación.

A mis amigos que me vieron poco estos años. A Víctor, Pablo, Tamara, Silvina, Andrés, Nico, Chudi, Tomás, Pigre, Joe. A Mariela Rajngewerc, con quien atravesamos paralelamente las dificultades de la academia. A Nina Pardal, con quien compartimos caminatas y charlas en Exactas.

A mi familia. A mi hermano Fer, a Graciela, a Julio. A mis primos Nico, Héctor y Meli. A mis viejos, dondequiera que estén, por impulsar mi curiosidad desde pequeño y siempre apoyarme en el estudio. Cada uno, a su manera, me fue llevando por este camino.

Finalmente quiero agradecer a Valeria, mi compañera de vida, que me apoyó en todo momento y soportó el estado de desborde emocional permanente que atraviesa todo doctorando. Realmente hubiera sido imposible sin vos.

A Valeria
A mis viejos
A quienes luchan cada día por hacer este mundo más justo

Índice general

Parte I Introducción 1

1.. ¿Por qué interesa la detección automática de discurso de odio? . . . 3
1.1. Algunos casos resonantes . . . 4
1.1.1. Atentados en Charlottesville . . . 4
1.1.2. Matanza en Sinagoga de Pittsburgh . . . 5
1.1.3. Masacre Rohingya en Myanmar . . . 6
1.2. Avances en el procesamiento del lenguaje natural . . . 6
1.2.1. Asimetría de recursos . . . 8
1.3. Detección de discurso de odio y sus limitaciones . . . 8
1.4. Aportes de este trabajo . . . 9

2.. Preliminares . . . 11
2.1. Aprendizaje supervisado . . . 11
2.2. Redes Neuronales . . . 12
2.2.1. Redes neuronales recurrentes . . . 13
2.3. Técnicas de representación . . . 14
2.3.1. Embeddings a nivel oración . . . 15
2.4. Transfer Learning y modelos pre-entrenados . . . 16
2.4.1. ELMo y ULMFiT . . . 16
2.4.2. Traducción automática y atención . . . 17
2.4.3. Transformers . . . 19
2.4.4. GPT, BERT y amigos . . . 21

Parte II Extracción de opiniones de redes sociales y detección de discurso de odio 25

3.. Extracción de opiniones de textos sociales . . . 27
3.1. Motivación . . . 27
3.2. Clasificación de textos sociales . . . 28
3.3. Trabajo previo . . . 29
3.4. Tareas analizadas . . . 30
3.5. Normalización y preprocesamiento . . . 32
3.6. Modelos de clasificación . . . 33
3.7. Resultados . . . 35
3.8. Discusión . . . 36
3.9. **pysentimiento**: un paquete de python para Análisis de Sentimiento . . . 37
3.10. Conclusiones . . . 37
3.11. Notas . . . 38

4.. Detección de discurso de odio . . . 39
4.1. ¿Qué es el discurso de odio? . . . 39
4.1.1. Abordaje desde una perspectiva legal y de los Derechos Humanos . . . 40
4.1.2. Definiciones utilizadas desde NLP . . . 43

4.2. Trabajo previo . . . 44
4.3. Descripción del dataset utilizado . . . 46
4.4. Tareas de clasificación . . . 46
4.5. Método . . . 48
4.5.1. Preprocesamiento . . . 48
4.5.2. Modelos de clasificación . . . 48
4.6. Resultados . . . 50
4.6.1. Análisis de Error . . . 52
4.7. Discusión . . . 54
4.8. Conclusiones . . . 58

Parte III Detección contextualizada de discurso de odio 61

5.. Construcción de un dataset de discurso de odio contextualizado . . . 63
5.1. Trabajo previo . . . 64
5.2. Esquema del conjunto de datos . . . 66
5.3. Proceso de construcción . . . 68
5.4. Recolección de datos . . . 68
5.4.1. Método de recolección . . . 69
5.4.2. Datos recolectados . . . 70
5.5. Selección de datos a anotar . . . 71
5.5.1. Selección en base a artículos . . . 72
5.5.2. Selección en base a comentarios . . . 72
5.5.3. Muestreo de comentarios . . . 73
5.6. Anotación . . . 74
5.6.1. Definición de discurso de odio y manual de etiquetado . . . 74
5.6.2. Modelo de etiquetado . . . 76
5.6.3. Etiquetadores . . . 77
5.6.4. Esquema de anotación . . . 78
5.6.5. Herramienta de etiquetado . . . 81
5.6.6. Asignación . . . 81
5.7. Resultados . . . 83
5.7.1. Co-ocurrencia de características ofendidas . . . 84
5.7.2. Análisis por característica . . . 85
5.8. Discusión . . . 90
5.9. Conclusión . . . 91
5.10. Notas . . . 91

6.. Experimentos de detección contextualizada de discurso de odio . . . 93
6.1. Trabajo previo . . . 93
6.2. Tareas de clasificación propuestas . . . 96
6.3. Modelos de clasificación . . . 97
6.3.1. Adaptación de dominio . . . 99
6.3.2. Rendimiento humano en la tarea . . . 100
6.4. Resultados . . . 102
6.5. Comparación de clasificadores y análisis de error . . . 105
6.6. Discusión . . . 109

6.7. Conclusiones 111

Parte IV Adaptación de Dominio 113

7.. Adaptación de dominio 115
7.1. Trabajo previo 116
7.2. Modelo pre-entrenado sobre tweets 118
7.2.1. Recolección de tweets 119
7.2.2. Arquitectura y entrenamiento 119
7.2.3. Evaluación 120
7.2.4. Resultados 121
7.3. Adaptación de modelos pre-entrenados 123
7.3.1. Metodología 123
7.3.2. Resultados 124
7.4. Adaptación de dominio para detección contextualizada de discurso de odio . 126
7.5. Discusión 127
7.6. Conclusiones 128
7.7. Notas 129

8.. Conclusiones 133

Apéndice 137

A.. Discurso de odio 139
A.1. Tratados internacional sobre libertad de expresión y discurso de odio 139
A.1.1. Libertad de expresión 139
A.1.2. Discurso de odio 140
A.2. Incidencia de keywords en el dataset 140

B.. Construcción de dataset contextualizado de discurso de odio 141
B.1. Distribución de datos recolectados 141
B.2. Recursos utilizados 141
B.3. Manual de criterios de anotación 141
B.3.1. Presencia de lenguaje discriminatorio 141
B.3.2. Llamado a la acción 143
B.3.3. Características protegidas 143

C.. Detección contextualizada de discurso de odio 149
C.1. Análisis comparativo entre clasificadores granulares y binarios 149

D.. Adaptación de dominio 151
D.1. Algunos detalles técnicos 151
D.2. Tabla completa de resultados 151
D.3. Evaluación multilingual 151
D.3.1. Resultados 152

Parte I

INTRODUCCIÓN

1. ¿POR QUÉ INTERESA LA DETECCIÓN AUTOMÁTICA DE DISCURSO DE ODIO?

El discurso de odio o discriminatorio [1] puede describirse como un discurso en clave de intenso aborrecimiento, denigración y enemistad que ataca a un individuo o un grupo de individuos por poseer –o aparentar poseer– cierta característica protegida por tratados internacionales como el género, la etnia, la creencia religiosa, el idioma hablado, entre otras. Si bien no hay un consenso generalizado sobre qué configura exactamente discurso de odio [7], un punto de contacto entre las distintas definiciones es su tendencia a generar un ambiente de hostilidad contra grupos o individuos, incitando a la violencia colectiva contra ellos.

En los últimos años, este tipo de discurso ha tomado gran relevancia en redes sociales y otros medios virtuales debido a su intensidad y a su relación con actos violentos contra miembros de estos grupos. A raíz de esto, estados y organizaciones supranacionales como la Unión Europea han sancionado legislación que insta a las empresas de redes sociales a moderar y eliminar contenido discriminatorio. Para citar un ejemplo, desde 2016 *Twitter* tiene en sus términos y condiciones:

> Conductas de incitación al odio: No se permite fomentar la violencia contra otras personas ni atacarlas o amenazarlas directamente por motivo de su raza, origen étnico, nacionalidad, pertenencia a una casta, orientación sexual, género, identidad de género, afiliación religiosa, edad, discapacidad o enfermedad grave. Tampoco permitimos la existencia de cuentas cuyo objetivo principal sea incitar la violencia contra otras personas en función de las categorías antes mencionadas.
>
> Imágenes y nombres de usuario que incitan al odio: No puedes usar imágenes o símbolos de incitación al odio en la imagen o el encabezado de tu perfil. Tampoco puedes usar tu nombre de usuario, nombre visible o biografía de perfil para participar en comportamientos abusivos, como realizar acosos dirigidos o expresar odio contra una persona, un grupo o una categoría protegida. (Política relativa a las conductas de incitación al odio, Twitter)

La enorme cantidad de texto generado por usuarios en las redes sociales –alrededor de 500 millones de tweets por día son posteados a nivel mundial– hace imposible que el análisis de este contenido sea realizado de manera enteramente manual. En este escenario de creciente preocupación que genera la proliferación de este discurso, se hace necesario el desarrollo de herramientas que automaticen la detección de discurso de odio en redes sociales, bien sea para el estudio y monitoreo de estas manifestaciones discriminatorias o bien para la moderación.

Desde el procesamiento de lenguaje natural, la detección de discurso de odio puede entenderse en su forma más básica como un problema de clasificación de texto: dado un texto generado por un usuario, predecir si es o no contenido discriminatorio. Así mismo, puede ser de interés predecir otras características: por ejemplo, si el texto contiene un

[1] Usamos de manera indistinta estas expresiones. Para una discusión sobre sus diferencias, ver la Sección 4.1

llamado a la acción violenta, si está dirigido contra un individuo o un grupo, el tipo de característica ofendida, entre otras. Poder identificar estas características puede ayudar a delimitar las formas más peligrosas de este fenómeno, como incitaciones a la violencia contra un grupo o individuo.

1.1. Algunos casos resonantes

Para hacer énfasis en la necesidad de desarrollar herramientas que puedan ayudar a la detección de contenido discriminatorio, comentamos algunos casos puntuales que han tenido lugar en los últimos años en los cuales han co-ocurrido [2] picos de discurso de odio en redes sociales –mayormente racista o xenófobo– con eventos de extrema violencia en la vida real. Aún cuando estos ejemplos relatan escenarios en sus formas más brutales, se ha observado que la mera exposición a este discurso en medios virtuales genera un profundo impacto negativo en la psiquis de sus objetivos [144], a la vez que prepara un terreno hostil y de deshumanización contra grupos vulnerados, como inmigrantes, minorías religiosas y sexuales [19], algo que ya ha sido estudiado a lo largo de décadas antes de la aparición de las redes sociales e Internet.

1.1.1. Atentados en Charlottesville

En Agosto del 2017, una gran movilización organizada por varios movimientos de ultraderecha y supremacistas blancos tuvo lugar en la ciudad de Charlottesville, Virginia, Estados Unidos. Esta concentración fue llamada en el medio del intento de universitarios y el movimiento Black Lives Matter (BLM) de remover estatuas de militares confederados pro-esclavitud de la Guerra de Secesión a lo largo de todo el territorio de Estados Unidos. En este caso puntual, se intentaba remover la estatua de Robert Lee ubicada en el campus de la Universidad de Virginia, durante los primeros meses de mandato de Donald Trump.

Numerosos grupos de ultraderecha, neonazis, neo-confederados (entre otros) convocaron a la marcha "Unite the Right"(UtR) para no permitir que se elimine la estatua de Robert Lee, organizando esta marcha como una campaña militar durante varios meses antes de su concreción. Blout and Burkart [23] describen la experiencia de Charlottesville como la de un "terrorismo inmersivo" ya que generaron un ámbito de terror en varios "teatros" (como lo llaman los autores, usando jerga militar). Principalmente, el teatro físico, con la marcha y enfrentamientos con contra-movilizaciones, la intimidante marcha de antorchas, y el asesinato de Heather Heyer atropellada por un manifestante neo-nazi. Paralelamente, el teatro "virtual" situado en las redes sociales sirvió para generar un clima de intimidación antes, durante, y luego del evento mencionado, desplegando –entre otras consignas de carácter racista y xenófobo– una campaña antisemita contra el alcalde de Charlottesville, de ascendencia judía, y el vicemayor, de ascendencia afroamericana.

Blout and Burkart [23] llegan a la conclusión de que el evento fue organizado de manera centralizada, tanto en su planificación como despliegue en un intento de ejercicio militar. También concluyen que la propaganda y la información diseminada por los organizadores sirvió para publicitar y reclutar a simpatizantes como también para aterrorizar a la población de Charlottesville. Esta propaganda se difundió tanto por medios impresos (por ejemplo, posters pegados en las calles) como por redes sociales como *Facebook, Twitter*

[2] Nótese que utilizamos la palabra co-ocurrir y no causar

Fig. 1.1: Último post de Robert Bowers, tirador en la masacre de Pittsburgh, en la red social Gab.

o *Discord*. Klein [85] analiza los intercambios en Twitter entre los dos bandos (manifestantes de ultraderecha y los contramanifestantes) y muestra que, en el caso de quienes se encontraban del lado de la marcha de UtR se identifica como enemigos a los musulmanes, liberales o izquierdistas, a miembros de la comunidad LGBTI [3], judíos, entre otros, dando cuenta del sesgo discriminatorio de este grupo.

1.1.2. Matanza en Sinagoga de Pittsburgh

En Octubre de 2018, un hombre fuertemente armado entró a la sinagoga "El Árbol de la Vida" en Pittsburgh, Pensilvania, Estados Unidos. Luego de gritar "muerte a los judíos", abrió fuego contra la multitud matando 11 personas y dejando decenas de heridos, la matanza más grande de judíos en EEUU de la que se tenga registro.

El tirador, Richard Bowers, era usuario activo de Gab [4], una red social que nació en 2016 bajo la égida de la defensa de la "libertad de expresión" a raíz de la creciente moderación de Twitter y Facebook a discursos discriminatorios. Desde entonces, ha sido el refugio de activistas de la derecha alternativa, supremacistas raciales, grupos conspiracionistas y otros elementos reaccionarios. El asesino en cuestión posteaba frecuentemente contenido antisemita en dicha red social [105], particularmente contra la HIAS (Sociedad Hebrea de auxilio de inmigrantes). En su último post en dicha red social, horas antes de la masacre, Bowers posteó una amenaza (ver figura 1.1) diciendo que no podía tolerar ver a su gente ser asesinada (por judíos) y que iba a tomar acciones al respecto.

A raíz de esto, Gab –llamada popularmente como el "Twitter racista"– estuvo de baja durante cierto tiempo al serle negado alojamiento web debido a este atentado. Desde entonces, diversos trabajos han recopilado y analizado el contenido discriminatorio en esta red social [82, 105].

[3] Lesbianas, Gays, Bisexuales, Transexuales, Intersexuales. Diversas variantes de estas siglas agregan más identidades a este colectivo, como LGBTIQ+.

[4] `https://gab.com/`

1.1.3. Masacre Rohingya en Myanmar

Entre 2016 y 2017 fue perpetrada una matanza de la etnia Rohingya, un grupo étnico musulmán, en la República de Myanmar (ex Birmania). Cerca de 25 mil personas fueron masacradas y un éxodo de más de 700 mil personas tuvo lugar hacia la lindante Bangladesh, conformando el campamento de refugiados más grande del mundo en la actualidad. La ONU y algunos estados nacionales han calificado lo ocurrido como un "genocidio" y como una "limpieza étnica".

Si bien el sometimiento de este pueblo tiene lugar hace décadas, en los últimos años tuvo un gran recrudecimiento motorizado desde las altas esferas gubernamentales y militares birmanas, que niegan cualquier estatus legal a la población rohingya. En ese punto, las redes sociales han jugado un rol de difusor y catalizador de incitaciones a la violencia y noticias falsas alrededor de esta etnia. Según un informe solicitado por Facebook acerca de la situación en Myanmar [163], gran parte de este problema se debe a un déficit en el "alfabetismo digital"(sic) de la población de este país, que usa casi exclusivamente Internet a través de dicha red social. Enviados de las Naciones Unidas han acusado directamente a Facebook de haber servido como intermediario de discurso de odio a través de su plataforma [5], y que ha tenido un "rol determinante" en este genocidio.

Organizaciones de derechos humanos de ese país han instado a la empresa de Mark Zuckerberg a invertir en el control del discurso de odio, particularmente aquel que insta a la violencia física [115]. A finales de 2021, un grupo de refugiados rohingya denunció a Facebook por 150 mil millones de dólares [6] por haber promovido la violencia contra esta etnia, luego de que en 2018 responsables de la empresa admitieran que no se hizo lo suficiente para detener la proliferación del discurso xenófobo contra los Rohingya en Myanmar.

Este hecho cuenta con una particularidad: apunta a un idioma –el birmano, idioma oficial en Myanmar– que dispone de pocos recursos en el área del procesamiento del lenguaje natural. La mayoría de los recursos y estudios están dedicados al idioma inglés, ignorando las particularidades de cada idioma y el componente cultural de algunas tareas, como en este caso la detección de discurso de odio. Además, según Reuters, para finales de 2018 Facebook no contaba con ningún empleado en Myanmar [7] ni tampoco quedaba claro que alguno de sus empleados dedicados a la tarea del monitoreo sea hablante nativo de birmano.

1.2. Avances en el procesamiento del lenguaje natural

En los últimos 10 años, el área de la Inteligencia Artificial ha sido sacudida por la irrupción de las redes neuronales. Desde el campo de Visión por Computadora, un conjunto de factores han potenciado el éxito de esta técnica de aprendizaje estadístico: datasets de gran tamaño como ImageNet [44], la utilización de dispositivos de gran poder de cómputo como las GPUs, y el desarrollo de mejores algoritmos para su entrenamiento (de optimización, funciones de activación, entre otras cosas). Esta combinación posibilitó que las redes neuronales obtengan mejoras considerables en el desempeño de tareas de reconocimiento de imágenes, trasladándose esto a otras áreas como procesamiento de habla, y a todas

[5] https://www.reuters.com/article/us-myanmar-rohingya-facebook/u-n-investigators-cite-facebook-role-in-myanmar-crisis-idUKKCN1GO2PN

[6] https://www.bbc.com/news/world-asia-59558090

[7] https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/

las áreas de aprendizaje automático en general, con particular foco de aquellos datos no estructurados como imágenes, sonido, y otras señales.

Este boom inicial tuvo su primera repercusión de magnitud en NLP cerca del año 2013 con el desarrollo de los word-embeddings. La técnica de *word2vec* [108] permitió generar representaciones de palabras de manera eficiente sobre grandes cantidades de datos no etiquetados. Estas representaciones de las palabras (podemos pensarlas como vectores de dimensión fija asignadas a cada token) han sido la "salsa secreta" que permitió el éxito de las redes neuronales en el área, permitiendo una mejora en las tareas de reconocimiento de entidades nombradas (NER), POS tagging, parsing, clasificación de textos, entre otras. Otro componente de este éxito de las redes neuronales ha sido el uso de redes recurrentes como las Long Short-Term Memory (LSTM) [74] o las Gated Recurrent Units (GRU) [34], que permiten codificar secuencias de manera autorregresiva. Un caso de éxito particular utilizando estas redes recurrentes ha sido el de la traducción automática mediante la arquitectura sequence-to-sequence (seq2seq) [157]. Estas redes permitieron atacar los problemas de aprendizaje de secuencia a secuencia, como la traducción automática o resumen automático de texto, reemplazando sistemas realmente complejos y de difícil mantenimiento (como los de Statistical Machine Translation) por diseños más simples y con una muy superior performance.

En 2017, Vaswani et al. [160] propusieron una arquitectura que elimina la estructura recurrente: los *Transformers*. Este modelo utiliza únicamente múltiples capas de autoatención para el problema de traducción automática. Al eliminar los pasos recurrentes, permitió la paralelización del cálculo y el entrenamiento de arquitecturas verdaderamente profundas para tareas de NLP, como ya hace tiempo se utilizaban en el área de Visión por Computadora. En conjunto a la aplicación del pre-entrenamiento utilizando la tarea de modelado de lenguaje (que introdujeron Howard and Ruder [77] con ULMFiT, entre otros trabajos) supusieron un cambio rotundo en el modo en que abordamos tareas de aprendizaje automático sobre textos: en lugar de entrenar una red neuronal casi desde cero –quizás sólo con una capa de embeddings con pesos iniciales pre-calculados– la idea es ahora sólo ajustar (*fine-tune*) una gran red neuronal pre-entrenada sobre un dataset de entrenamiento con alguna tarea de modelado de lenguaje. *GPT*, *BERT* y otros personajes de Plaza Sésamo son algunos de los rutilantes nombres en el zoológico de modelos pre-entrenados que son hoy día el estado del arte de NLP. Este nuevo enfoque supuso un gran paso adelante en el área, mejorando los desempeños sensiblemente en benchmarks de tareas como GLUE [161] y RACE [89], entre otros.

En suma, todos estos avances han permitido atacar numerosas tareas que eran problemáticas para NLP. Algunas –como resumen o traducción automática– que adolecían de pobres performances o sistemas realmente complejos y difíciles de mantener. Otras, que simplemente estaban fuera del radar del estado del arte, como por ejemplo tareas de Common Sense Reasoning. Dentro de estas tareas, Íla detección de discurso de odio y toxicidad ha sido de aquellas tareas que, si bien en la etapa previa han podido utilizar sistemas basados en la detección de n-gramas, eran muy frágiles y susceptibles ante el ruido típico de este tipo de texto. Con el advenimiento de modelos neuronales y representaciones más robustas, han mejorado notablemente su performance frente a escenarios más complejos, y se han propuesto nuevos desafíos para atacar este discurso, como la respuesta automática a mensajes discriminatorios [36].

1.2.1. Asimetría de recursos

Un problema no menor en el área de NLP es la enorme asimetría de recursos entre idiomas. La inmensa mayoría de datasets, corpus, modelos y –consecuentemente– estudios han sido realizados en inglés. En particular, el caso de los modelos pre-entrenados basados en Transformers introducidos en los últimos años vienen a agravar esta asimetría ya que estos modelos necesitan muchos recursos computacionales para ser generados, usualmente no disponibles por fuera de algunos pocos laboratorios que centran sus estudios en inglés. Esto ocasiona, por un lado, la creencia generalizada de que muchas técnicas son *independientes del lenguaje*, omitiendo importantes diferencias entre lenguajes y dialectos. Por otro lado, teniendo en cuenta que el objeto de este trabajo es mayormente sociolingüístico, esto dificulta la posibilidad de realizar estudios posteriores utilizando los recursos generados en forma de modelos o datasets.

Para atacar este problema, es entonces necesario desarrollar recursos y estudios en los distintos idiomas, atendiendo sus particularidades lingüísticas y culturales. La "Regla de Bender" describe de manera elegante esto:

> Do state the name of the language that is being studied, even if it's English. Acknowledging that we are working on a particular language foregrounds the possibility that the techniques may in fact be language specific. Conversely, neglecting to state that the particular data used were in, say, English, gives a false veneer of language-independence to the work. (Regla de Bender, Bender [14])

Si bien el español puede considerarse de los idiomas dentro del grupo de los de "altos recursos" [13], aún así la disparidad al compararse con el inglés es abrumadora. En particular, para el área de interés de esta tesis –la detección de discurso de odio–, los recursos son muy escasos y, en la mayoría de los casos, "réplicas" de trabajos hechos en inglés, sin ninguna novedad adicional.

1.3. Detección de discurso de odio y sus limitaciones

Como dijimos anteriormente, la detección del discurso de odio puede pensarse como una tarea de clasificación binaria sobre un texto generado por un usuario. Muchos trabajos en los últimos años han abordado la tarea desde esa perspectiva, desarrollándose numerosos recursos en workshops para varios idiomas, y herramientas muy utilizadas para la moderación de contenido tóxico [8] como Perspective API [9], desarrollada por Jigsaw y Google.

Si bien predecir únicamente si un comentario contiene o no lenguaje discriminatorio tiene una innegable utilidad, este enfoque adolece de ciertas limitaciones. Uno de sus problemas es la falta de contexto en los mensajes analizados. Las personas no solemos consumir el lenguaje de manera aislada sino que lo entendemos situado de acuerdo a varios factores: el emisor, la situación y el medio en el que se lo emite, a quién está dirigido, sobre qué hace mención, entre otras cosas. La mayoría de los estudios y recursos, sin embargo, están realizados sobre datos fuera de contexto: mensajes en redes sociales sin ningún tipo de información conversacional o de otra índole. Para ilustrar el problema de la falta de

[8] El contenido tóxico o abusivo es una categoría un poco más general que la de discurso de odio
[9] `https://www.perspectiveapi.com/`

Fig. 1.2: Tweets y respuestas discriminatorias. Leyendo únicamente el texto de los comentarios resulta difícil descifrar su sentido.

contexto, la figura 1.2 muestra un tweet de un medio periodístico que habla sobre una actriz y respuestas a esa noticia [10]. Leyendo los tweets por fuera del contexto es difícil comprender las respuestas con contenido transfóbico dirigidos a la actriz en cuestión.

La falta de contexto restringe la información disponible –tanto para un humano como para un algoritmo– para poder discernir si un comentario de un usuario es o no discriminatorio. Otra información usualmente no disponible y que puede ayudar a enriquecer la detección de discurso de odio es la característica atacada en un texto: es común que los datasets estén anotados de manera poco granular –casi siempre de manera binaria– no brindando información acerca de si la agresión es por motivos de género, religión, etnia, etc. Contar con esta información puede ser de utilidad para que un algoritmo pueda detectar mejor los diferentes tipos de discurso discriminatorio mediante una señal más rica. Así también, los usuarios de estos algoritmos de detección automática de discurso de odio podrían obtener información más concreta acerca del fenómeno detectado: no sólo si es discriminatorio o no sino el motivo por el cual se lo considera.

Por último, una limitación puntual del español es la poca disponibilidad de recursos para esta tarea, algo que mencionamos en la anterior sección como un problema general de NLP. A esto se le suma que los pocos datasets disponibles –tanto para español como otros idiomas– suelen adolecer de cierto déficit de calidad en su construcción: generados por equipos con poca o nula interdisciplina; o bien anotados por sujetos que no son hablantes nativos del idioma en cuestión o no están inmersos en su realidad sociocultural.

1.4. Aportes de este trabajo

En esta tesis nos proponemos hacer un aporte en el sentido de desarrollar mejores mecanismos automáticos de detección de discurso de odio. Si bien el área de NLP ha avanzado enormemente en los últimos años – y esta subdisciplina en particular ha recibido un gran interés– creemos que muchos de los enfoques actuales inhiben un avance cualitativo

[10] El hilo completo puede encontrarse en `https://twitter.com/infobae/status/1242506130213015552`

en la detección de este pernicioso fenómeno en medios sociales.

Para ello, en primer lugar estudiamos técnicas de detección sobre recursos ya existentes, utilizando del estado del arte. En base a la observación de algunos datasets y la literatura en general, planteamos un nuevo problema: la detección *contextualizada* de discurso de odio. Construimos un corpus de discurso de odio sobre comentarios en noticias de medios gráficos argentinos en *Twitter*, siendo este conjunto de datos etiquetado por hablantes nativos. Este dataset es un aporte importante en sí ya que es uno de los primeros que incluyen información contextual, y es el único a nuestro conocimiento en español que tiene esta información. A su vez, fue construido de manera interdisciplinaria y con una metodología clara en su recolección y anotación.

Con este recurso, exploramos la siguiente pregunta: ¿pueden los métodos actuales basados en modelos pre-entrenados aprovechar información adicional de contexto para mejorar la detección de discurso de odio? Este punto ha sido poco estudiado en la literatura y consideramos que es una pregunta de interés para atravesar los límites de la clasificación basada en una única fuente de información (el comentario analizado). En base a los experimentos realizados, encontramos evidencia de que el contexto puede brindar información útil para detectar este fenómeno, mejorando el desempeño de los algoritmos para esta tarea. Particularmente, observamos que para los mensajes de odio contra ciertos grupos –por ejemplo, contra la comunidad LGBTI– el contexto puede ser aún más útil para su detección.

Finalmente, realizamos un estudio más en general sobre la *adaptación de dominio* en tareas de clasificación de redes sociales. Para ello, generamos un modelo de lenguaje pre-entrenado sobre textos sociales en español al que bautizamos *RoBERTuito*, el primero disponible y a gran escala en este idioma. Comparamos el desempeño de *RoBERTuito* contra otros pre-entrenados sobre textos formales pero ajustados al dominio social. Esta comparación es de interés ya que el ajuste de dominio es relativamente económico frente al enorme costo de entrenamiento que tiene construir modelos como *RoBERTuito*. Observamos que para todas las tareas, *RoBERTuito* obtiene una performance del estado del arte, pero el ajuste de dominio recorta considerablemente la brecha contra otros modelos.

Un aporte en general de esta tesis es que todos los estudios y recursos han sido realizados en español. Vista la enorme asimetría que hay con otros idiomas, y teniendo en cuenta que el español es el segundo idioma en hablantes nativos del mundo, consideramos necesario mitigar este desbalance de recursos.

2. PRELIMINARES

En esta sección realizamos una breve introducción a algunas técnicas de Machine Learning y NLP que utilizamos a lo largo de esta tesis. Particularmente, ilustramos y describimos a grandes rasgos los últimos avances de Deep Learning para el área, desde *word2vec* [108] hasta la arquitectura Transformers [160].

2.1. Aprendizaje supervisado

Muchas de las tareas de las áreas de procesamiento de lenguaje natural, visión por computadora, procesamiento del habla (entre otras) pueden convertirse a problemas de aproximar una función $f : D \to O$, donde D es el *dominio* y O el codominio o posibles salidas. Las funciones que nos interesan aproximar son desconocidas, altamente no lineales y difícilmente expresables de manera analítica. En el caso de que O sea un conjunto finito, diremos que estamos ante un problema de **clasificación** y llamaremos a nuestro aproximador $\widehat{f}$ **clasificador** y a cada uno de las posibles salidas **clases**. En el caso de que O conste de una o más variables continuas, decimos que estamos ante un caso de **regresión**. Para lo que concierne a esta tesis, nos centraremos en problemas de clasificación, así que restringiremos (salvo en los lugares donde se explicite lo contrario) nuestro análisis a este tipo de tareas.

Un ejemplo canónico de clasificación del área de Visión por Computadora es el de, dada una imagen de 28×28 píxeles en blanco y negro, predecir a cual de los 10 caracteres pertenece. En este caso, tenemos que D es un subconjunto de $\{0,1\}^{28\times 28}$ –todas las imágenes posibles de ese tamaño en blanco y negro– restringido a aquellas imágenes que correspondan a caracteres, y que $O = 0, 1, 2, \ldots, 8, 9$ son las posibles salidas de esta función.

En el caso de NLP, uno de los que problemas que más veremos en esta tesis es el de clasificación de textos: dado un texto (un documento, un comentario en una red social) predecir alguna característica discreta de éste. Por ejemplo, si el texto es un comentario de una red social, podemos intentar identificar su polaridad: si es positivo, neutro, o negativo. O si el comentario en cuestión posee algún tipo de discriminación: en este caso tenemos como posibles salidas 0 (marcando no discriminatorio) o 1 en caso de que sí haya discriminación. Otro tipo de tarea es la de inferencia (Natural Language Inference o NLI): dadas dos oraciones de texto, predecir si una es una consecuencia lógica de la otra, si son independientes, o si son contradictorias. En este caso, el dominio son pares de oraciones de texto, y tenemos tres posibles salidas: contradicción, independiente, consecuencia.

Estos problemas que acabamos de describir son realmente difíciles de atacar mediante programas convencionales diseñados a través de heurísticas o reglas prefijadas [20]. En lugar de ello, los abordaremos mediante técnicas de **aprendizaje supervisado**, para lo cual tendremos un conjunto de instancias $x_1, \ldots, x_n$ y sus respectivas etiquetas $y_1, \ldots, y_n$ que usaremos para **entrenar** nuestro aproximador $y \sim f^*(x)$. Este conjunto $\{(x_1, y_1), \ldots (x_n, y_n)\}$ se denomina **conjunto de entrenamiento**.

El proceso de entrenamiento de un estimador consta de seleccionar una función $f\theta$ de un conjunto de candidatos $\{f_\theta : \theta \in \Sigma\}$, donde θ representa los parámetros del clasificador, y Σ representa el conjunto de sus posibles valores. En el caso de un clasificador lineal (como

una Support Vector Machine o una regresión logística) θ constará de un vector de $\mathbb{R}^n$. En el caso de redes neuronales (como las que veremos a continuación), constará de múltiples matrices y vectores correspondientes a sus diferentes capas.

Si bien en algunos capítulos de esta tesis utilizamos técnicas de clasificación lineal como Support Vector Machines y regresiones logísticas, nuestro eje está puesto en los modelos basados en redes neuronales. Estos modelos han logrado el estado del arte en NLP para casi cualquier tarea conocida, y pasamos a continuación a hacer un repaso de sus distintas variantes. Para una descripción de los modelos lineales, referimos a textos clásicos del área como Bishop [20].

2.2. Redes Neuronales

Los **perceptrones multi-capa** (MLP) o **redes feed-forward** (FFN) son la "quintaesencia" de los métodos modernos de Deep Learning, como describen Goodfellow et al. [63]. Una de las primeros acercamientos a esta técnica es la neurona de McCulloch-Pitts [104], que intenta modelar parte del funcionamiento de las neuronas mediante una función:

$$y = H(\theta^T x)$$

donde H es la función de Heaviside o función escalón, que vale 1 si $x \geq 0$, y 0 en otro caso. La neurona de McCulloch-Pitts permite aproximar a dos valores (0 ó 1), a partir de una entrada x y un parámetro θ. El perceptrón, desarrollado en 1958 en Rosenblatt [140], es el primer modelo que utiliza este tipo de modelo de cómputo cuyos parámetros se encuentran mediante un algoritmo. Minsky and Papert [109] demostraron que este tipo de modelos sólo pueden ajustarse a datos linealmente separables, provocando que por largo tiempo no se profundice en la investigación de redes neuronales.

Una forma de sortear estas dificultades planteadas es apilar (stack) estas neuronas para poder ajustar a más tipos de funciones. En términos matemáticos, esto es tan sólo una composición de funciones, tomando ahora $f = f_3 \circ f_2 \circ f_1$, donde f_1 es la primer "capa" de nuestra función correspondiente a la entrada, f_2 es la capa intermedia u oculta, y f_3 es la capa de salida, cada una teniendo sus parámetros $\theta_1, \theta_2, \theta_3$. Si bien este ejemplo consta de 3 capas, se puede generalizar a arbitrarias capas ocultas. Este modelo es el que conocemos como **Perceptrón Multicapa** o **Multi-Layer Perceptron** (MLP por sus siglas en inglés), y provocó el resurgir conexionista de las redes neuronales en los años 80s gracias al desarrollo de algoritmos que permitieron entrenar estos modelos mediante la técnica de backpropagation [143].

Cybenko [41] demostró que para cualquier función continua en el hipercubo de $\mathbb{R}^n$ existe una red neuronal con una función de activación sigmoidal de 3 capas que la puede aproximar infinitamente bien [1]. Sucesivos resultados demostraron con mayor generalidad este resultado, para otras funciones de activación (como las ReLU) y otras arquitecturas. Dicho en términos coloquiales, este teorema asegura que para cualquier función podemos encontrar una MLP que la aproxima. Hay que notar, sin embargo, que este y otros teoremas aseguran existencia pero no son constructivos ni tampoco aseguran que el proceso de backpropagation nos lleve a esa solución [63]. Más aún, tampoco indican qué tan grande tiene que ser la red neuronal para que pueda aproximar adecuadamente a la función objetivo. Teniendo estas salvedades en cuenta, estos resultados (usualmente denominados

[1] La formulación del resultado es un poco más compleja pero escapa los fines de esta tesis

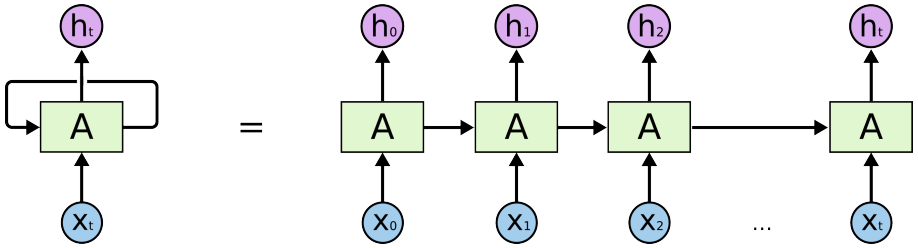


Fig. 2.1: Ilustración del esquema general de una red neuronal recurrente. Fuente: Blog de Christopher Olah

como Teorema de Aproximación Universal de Redes Neuronales) aportan un sustento teórico de la potencialidad de estos algoritmos, refutando en cierto punto lo marcado por Minsky and Papert [109].

2.2.1. Redes neuronales recurrentes

Los problemas descriptos de NLP suelen constar de procesar una secuencia de palabras o tokens $x_1, x_2, \ldots, x_k$ de longitud variable, de manera de ajustar a una función

$$h = f([x_1, \ldots, x_k])$$

Una reformulación de este problema convirtiéndolo a una función que recibe una entrada de largo fijo es el de ajustar una función autorregresiva:

$$h_t = f(x_t, h_{t-1})$$

En este caso, tenemos una salida para cada paso k de tiempo. Si f es una red neuronal, llamamos a este tipo de redes neuronales **recurrentes**, ya que la salida a cada paso (h_t) depende de la salida del paso anterior, h_{t-1}. La Figura 2.1 ilustra este esquema, tomada del excelente artículo de Chirstopher Olah sobre LSTMs [2]. Una primer aproximación a las redes recurrentes es la red de Elman [48], definida por las siguientes ecuaciones

$$h_t = \sigma(W_h x_t + U_h h_{t-1} + b_h) \tag{2.1}$$
$$y_t = \sigma(W_y h_t + b_y) \tag{2.2}$$

donde h_t es normalmente llamado el **estado oculto** en las redes neuronales recurrentes, e y_t es la salida propiamente dicha. Los parámetros a ajustar son W_h, U_h (matrices) y b_h, b_y (escalares). Podemos ver que, a grandes rasgos, este tipo de red recurrente no es nada más que un perceptrón multicapa cuya entrada consta de x_t, la entrada original en el tiempo actual t, y el estado oculto anterior, h_{t-1}.

Para entrenar este tipo de redes recurrentes utilizamos back-propagation through time (BPTT), que consta en desplegar la relación recurrente –como está ilustrado en la Figura 2.1– y aplicar back-propagation de manera normal, poniendo un límite en la cantidad de pasos que tomamos hacia atrás. Las redes recurrentes de Elman sufren de varios problemas, principalmente de **vanishing gradient** y **exploding gradient**. Ambas dificultades pueden observarse ya que el cálculo del gradiente de las ecuaciones 2.2 usando BPTT

[2] `https://colah.github.io/posts/2015-08-Understanding-LSTMs/`

induce la potencia a la n (donde n es el largo de la secuencia) de las matrices W_h y U_h, lo cual puede hacer o que bien el gradiente tienda a cero o a infinito. [3]

Los gradientes que tienden a infinito o **exploding gradient** pueden solucionarse mediante la técnica de **gradient clipping** [63], que consta de reajustar la norma del gradiente para que no exceda cierto valor. Sin embargo, nos queda aún el inconveniente de vanishing gradient. Para ello, se han propuesto otras arquitecturas recurrentes. Hochreiter and Schmidhuber [74] propusieron las **Long Short-Term Memory** (LSTM), una arquitectura basadas en compuertas (gates) que regulan el comportamiento del estado oculto y de la salida, evitando algunos de las dificultades de aprendizaje en las que incurre la red de Elman [4]. Otras arquitecturas como las Gated Recurrent Units [34] usan menor cantidad de compuertas reduciendo la cantidad de parámetros a entrenar.

2.3. Técnicas de representación

Una de las necesidades que tienen las redes neuronales para poder trabajar con textos es el de tener representaciones continuas de cada token o palabra. Las representaciones utilizadas en la época previa de los donde reinaban los modelos lineales –bolsas de palabras/caracteres ponderadas con esquemas como TF/IDF– adolecen de varios inconvenientes. En primer lugar, tienen una altísima dimensionalidad, usualmente del tamaño del vocabulario o algún límite similar. A su vez, no guardan representación semántica de la similaridad de las palabras: dos palabras como silla o banco tienen la misma distancia que perro y nube. Finalmente, sus valores no nulos están concentradas en una o pocas dimensiones y suelen ser discretas.

Latent Semantic Analysis (LSA) [91] es una de las primeras técnicas de representación continua que tuvo cierta popularidad. Para obtener representaciones continuas de las palabras, plantean la factorización una matriz de co-ocurrencia entre tokens y documentos (o contextos) usando la descomposición SVD, obteniendo vectores de dimensión fija para los documentos y términos. LDA (Latent Dirichlet Allocation) [21] es otra técnica basada en modelos gráficos entrenados mediante métodos variacionales, muy utilizada aún en la actualidad ya que genera representaciones latentes de los tópicos de los textos.

Dentro de los métodos neuronales, uno de los más populares ha sido el de Bengio et al. [17], que propone una arquitectura neuronal para un modelo de lenguaje markoviano. En la capa intermedia contiene una tabla de lookup de vectores de las diferentes palabras (también conocido como capa de embeddings) donde se generan las representaciones de las palabras durante la etapa de entrenamiento. Trabajos posteriores (con diferentes variaciones de esta misma idea) como el de Collobert et al. [38] han demostrado que la utilización de este tipo de representaciones es útil para diversas tareas de NLP como POS Tagging, NER, y otras. Más aún, este trabajo tiene una idea que fue utilizada muchos años después con éxito rotundo: la utilización de la tarea de modelado de lenguaje como base para el pre-entrenamiendo de redes neuronales.

Uno de los problemas de los métodos comentados es que no son muy eficientes, sólo pudiéndose entrenar con pocos millones de palabras y con dimensiones reducidas. La técnica *word2vec* [107] permite entrenar representaciones de mayor dimensión y sobre grandes

[3] Esto puede verse usando alguna descomposición de la matriz como la forma normal de Jordan. Sus elementos en la diagonal que sean distintos de 1, o bien tienden a infinito o a cero.

[4] Recomendamos el artículo antes mencionado de Christopher Olah para una muy buena explicación de este tipo de redes

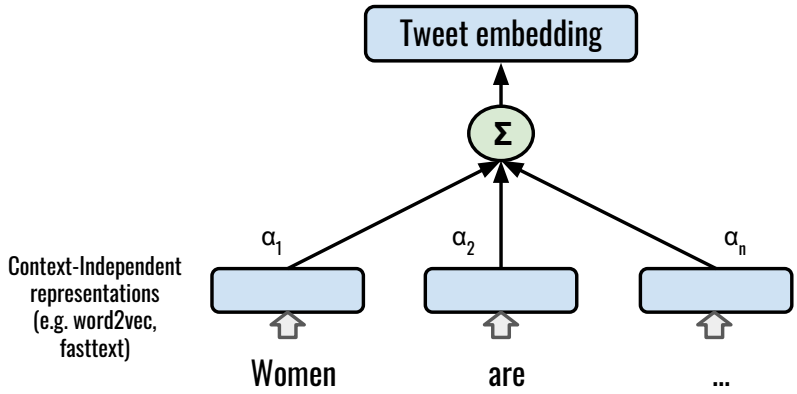


Fig. 2.2: Representación continua de un tweet mediante combinación lineal de las representaciones de cada palabra.

cantidades de textos de manera eficiente. Los vectores de palabras aprendidos guardan cierta estructura lineal y semántica, ilustrado por los autores con algunos ejemplos de analogías de palabras como el ya clásico $v(\text{rey}) - v(\text{hombre}) + v(\text{mujer}) \approx v(\text{reina})$.

Para generar los vectores de *word2vec*, los autores plantean una relajación de la tarea de modelado de lenguaje mediante dos alternativas: Continuous Bag of Words (CBOW) y Skip-Gram. En CBOW se intenta predecir la palabra faltante dada una bolsa de palabras del contexto, mientras que Skip-gram se intenta predecir el contexto dada la palabra central, obteniendo en ambas variantes representaciones intermedias ricas. Mikolov et al. [107] extiende la idea del anterior trabajo proponiendo plantear el problema de skip-gram como uno de distinguir palabras ruido de palabras efectivamente del contexto, haciendo mucho más eficiente el cálculo de estas representaciones. *GloVe* [124] es otra técnica de representación de palabras que combina las ideas de factorización de matrices de LSA mediante un problema de optimización distinto y generando representaciones que superan ligeramente en algunos benchmarks de tareas a los de *word2vec*.

Los métodos mencionados de representación calculan vectores de tamaño fijo sobre cada una de las distintas palabras. En español, por ejemplo, las palabras gato, gata, gatito, gatuno, todas tienen representaciones independientes en *word2vec*, a pesar de tener información morfológica en común. Esto es un problema en varios escenarios: idiomas con muchas inflexiones o aglutinantes (como el turco, alemán o finés) o –lo que es de nuestro interés– texto altamente desnormalizado como el de redes sociales. La técnica *fastText* [24] extiende la idea de *word2vec* mediante la asignación de vectores a secuencias de 3 caracteres (subpalabras), capturando así cierta información morfológica. La representación de una palabra se obtiene mediante una combinación lineal de los vectores de las subpalabras que la componen.

2.3.1. Embeddings a nivel oración

Una forma relativamente simple de obtener una representación de un texto [5] es realizar una combinación lineal de las representaciones obtenidas para cada palabra. Es decir, dada una oración $s = w_1 w_2 \dots w_n$, y representaciones $\overline{w_1}, \overline{w_2}, \dots, \overline{w_n} \in \mathbb{R}^m$, podemos obtener

[5] en nuestra tesis, esto será casi siempre un tweet

una representación

$$\overline{s} = \sum_{i=1}^{n} \alpha_i \overline{w_i} \tag{2.3}$$

con $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$ escalares (dependientes de la oración). De esta manera, obtenemos de n representaciones independientes del contexto una representación para el tweet, sin tener en cuenta posibles interacciones entre los distintos componentes. La Figura 2.2 ilustra esta metodología simple para obtener representaciones de oraciones.

Tenemos entonces dos posibilidades para determinar la combinación lineal: la forma de obtener las representaciones, y la forma de calcular los coeficientes. Para las representaciones, podemos usar varias de las técnicas que ya vimos como *word2vec*, *GloVe*, o *fastText*. Para calcular los coeficientes, consideramos dos posibilidades. La primera, la forma canónica, calculando un promedio de las representaciones, es decir, tomando $\alpha_i = \frac{1}{n}$. Otra es realizar una ponderación usando Smooth Inverse Frequency (SIF) [6], inspirado en TF-IDF. Cada palabra w se pondera con $\frac{a}{a+p(w)}$, donde $p(w)$ es la probabilidad del unigrama y a es un hiperparámetro de suavizado. Los valores altos de a significan más suavizado hacia el promedio simple.

2.4. Transfer Learning y modelos pre-entrenados

2.4.1. ELMo y ULMFiT

Hasta cerca de 2018, la forma canónica de abordar un problema de NLP era entrenar una red neuronal recurrente que consumiera embeddings no contextualizados de los tokens de entrada. Esta arquitectura tiene algunas limitaciones; una de ellas es que, dados dos o más problemas distintos (por ejemplo, análisis de sentimientos y NLI) lo único compartido por ambas redes es la capa más baja –la capa de embeddings – teniendo que entrenar desde cero todo el resto de los parámetros. En términos coloquiales, cada red debe "aprender a leer" sobre cada tarea, ignorando muchas construcciones sintácticas y semánticas comunes del lenguaje.

Uno de los esfuerzos exitosos en sobrepasar este abordaje es ELMo [129]. Este modelo aprende representaciones ya no sobre una única palabra como *word2vec* sino sobre toda una oración, generando representaciones contextualizadas para cada una de ellas. ELMo se entrena sobre una tarea de modelo de lenguaje bidireccional [6] recurrente de varias capas sobre grandes cantidades de texto. En dicho trabajo, utilizan luego una combinación lineal de la salida de cada capa para obtener representaciones contextualizadas de cada token. Esta misma idea es una continuación de Peters et al. [128], y también parcialmente de *CoVe* [103] donde construyen representaciones contextualizadas mediante la tarea de traducción automática.

Alrededor de 2018, este paradigma de entrenar una red desde cero compartiendo su capa más baja –word2vec o bien ELMo– comenzó a cambiar hacia un esquema donde se entrena una red neuronal sobre una tarea genérica para luego ajustarla a la tarea específica, una práctica muy común en el área de Visión por Computadora. Howard and Ruder [77] introdujeron la técnica de ULMFiT(Universal Language Modeling for Fine-tuning for text classification), uno de los trabajos fundamentales de este nuevo enfoque en NLP. ULMFiT consta de pre-entrenar en primer lugar un modelo de lenguaje sobre

[6] En realidad no es estrictamente bidireccional, sino dos LM concatenados

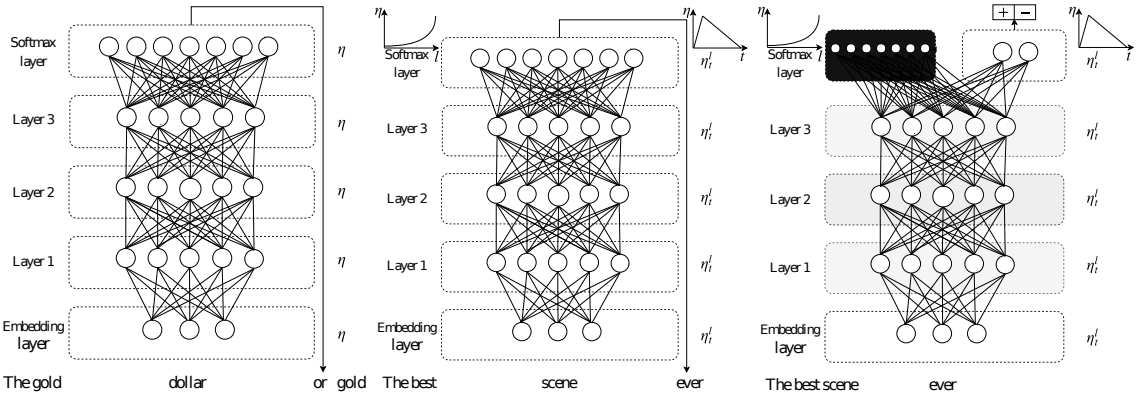


Fig. 2.3: Universal Language Modeling for Text Classification (ULMFiT). Esquema del método planteado: primero se pre-entrena sobre la tarea de modelado de lenguaje sobre un dataset no supervisado. Luego, se corre la misma tarea pero sobre el texto de la tarea ignorando las etiquetas (ajuste de dominio). Finalmente, se agrega una capa de parámetros particulares de la tarea y se entrena la red completa descongelando gradualmente cada capa. Fuente: Howard and Ruder [77]

un gran dataset no etiquetado, y luego utilizar esa misma red (cambiándole la última capa) para ajustarla a una tarea específica. El primer paso, el **pre-entrenamiento** es realizado una única vez, y sus pesos son luego re-utilizados para realizar el ajuste en cada tarea distinta. Este es uno de los primeros esquemas de **transfer learning** exitosos sobre NLP: transferimos conocimiento de la tarea de modelado de lenguaje a las distintas tareas finales que realizamos como POS tagging, análisis de sentimientos, detección de entidades nombradas, etc.

Los autores proponen tres etapas: primero, el pre-entrenamiento sobre la tarea de modelado de lenguaje en un gran dataset de texto (e.g. Wikipedia o Common Crawl); segundo, un ajuste de la tarea de modelado de lenguaje sobre el texto de la tarea en cuestión (LM fine-tuning); y finalmente, el entrenamiento sobre las etiquetas de la tarea (Classifier fine-tuning). La Figura 2.3 ilustra las tres etapas para el problema de clasificación de sentimientos. Entre varias técnicas que utilizan para entrenar estos modelos, vale destacar el uso de *slanted triangular learning rates*, donde el learning rate tiene una etapa de *warmup* donde sube hasta el pico y luego una etapa de *annealing* donde se reduce linealmente hasta 0 por el resto del entrenamiento. Esta técnica es también utilizada por *BERT* y otros modelos de lenguaje basados en transformers.

El modelo de lenguaje utilizado por los autores de *ULMFiT* utiliza una arquitectura *AWD-LSTM* [106]. Estas arquitecturas recurrentes fueron el estado del arte para las tareas de modelado de lenguaje (y consecuentemente, para esquemas de transfer learning como el mencionado) pero fueron sobrepasados a los pocos meses por los modelos de lenguaje basados en *transformers*.

2.4.2. Traducción automática y atención

Hacemos a continuación una pequeña digresión sobre traducción automática y atención, conceptos necesarios antes de hablar de Transformers. El modelo *encoder-decoder* permitió la utilización de redes neuronales para tareas de traducción de secuencias, es decir, donde queremos convertir $x = x_1 \dots x_n$ a $y = y_1 \dots y_m$, dos secuencias de diferente

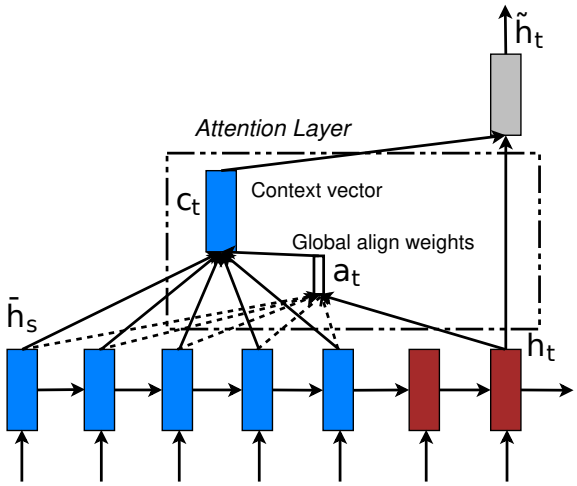


Fig. 2.4: Mecanismo general de atención. En azul, la salida del encoder recurrente de la entrada. En rojo, la salida del decoder recurrente. Fuente: Luong et al. [97]

longitud. Sutskever et al. [157] propusieron el siguiente esquema para abordar esta clase de tareas: en primer lugar, un *codificador* (encoder) consume la cadena de entrada x, convirtiéndola a un vector contextualizado h de dimensión fija; y luego, un *decodificador* (decoder) consume el vector de contexto h para convertirlo a la cadena de salida y. Esto puede resumirse en las ecuaciones:

$$h = \text{encoder}(x)$$
$$y = \text{decoder}(h)$$

Particularmente, los autores plantea un modelo basado en redes recurrentes (LSTMs) tanto para el *encoder* como para el *decoder*. El *encoder* consume paso a paso la entrada, y el *decoder* es esencialmente un modelo de lenguaje recurrente condicionado por el vector de entrada h.

Una de las limitaciones de las redes recurrentes es que sufren sesgo de **localidad** o **secuencialidad** (locality bias) [11]. En palabras coloquiales, las redes recurrentes tienen problemas para aprender dependencias de largo rango en las oraciones, siendo esto producto de su arquitectura autorregresiva donde se construye la salida y_t en base a y_{t-1}. Este sesgo es particularmente dañino en tareas de tranducción de sequencias con la arquitectura encoder-decoder básica ya que a esto se le suma un cuello de botella forzoso por la compresión de toda la secuencia de entrada en el vector de dimensión fija h. Un síntoma de esto es observado en Sutskever et al. [157]: invirtiendo la oración de entrada obtiene mejores resultados para la tarea de traducción automática.

Los mecanismos de *atención* [9] ayudan a mitigar estas dificultades. A grandes rasgos, esta técnica agrega información adicional en cada paso del decoder mediante una combinación lineal de los estados ocultos del encoder. Recordemos que en la arquitectura básica propuesta en Sutskever et al. [157] eran descartados, sólo siendo utilizado el estado final h. Describimos a continuación las ecuaciones de atención, que nos servirán para entender el mecanismo utilizado en la siguiente sección sobre Transformers.

Suponiendo que queremos traducir una secuencia $(x_1, \ldots, x_n)$ a $(y_1, \ldots, y_m)$, y que tenemos los estados ocultos $(\overline{h_1}, \ldots, \overline{h_n})$ generados en el encoder para la entrada, y los

estados ocultos $(h_1, \ldots, h_m)$ y para la salida, el mecanismo de atención [7] consta de calcular para cada paso t de la etapa de decodificación un vector de contexto

$$c_t = \sum_{i=1}^{n} \alpha_i^{(t)} \overline{h_i}$$

donde $\alpha^{(t)}$ es el vector de alineamiento, calculado como

$$\alpha^{(t)} = \text{softmax}(\text{score}(\overline{h_1}, h_t), \text{score}(\overline{h_2}, h_t) \ldots, \text{score}(\overline{h_n}, h_t))$$

Cada $\text{score}(\overline{h_i}, h_t)$ marca una similaridad no normalizada entre sus argumentos. Las alternativas planteadas en Luong et al. [97] son:

$$score(\overline{h_i}, h_t) = \begin{cases} \overline{h_i}^T h_t & \text{dot} \\ \overline{h_i}^T W h_t & \text{general} \\ v^T \tanh(W[\overline{h_i}^T; h_t]) & \textit{concat} \end{cases} \tag{2.4}$$

con W y v parámetros adicionales. En el caso de la atención producto interno podemos reescribir todas las ecuaciones como:

$$C = \text{softmax}(H\widehat{H}^T)\widehat{H} \tag{2.5}$$

donde $\widehat{H}, H$ son los vectores que tienen $(\overline{h_1}, \ldots, \overline{h_n})$ y $(h_1, \ldots, h_m)$ como filas respectivamente, y softmax se calcula fila a fila.

Finalmente, el vector $\widetilde{h_t}$ es calculado como una transformación del estado oculto del decoder h_t y el vector contextual c_t:

$$\widetilde{h_t} = \tanh(W_h[h_t; c_t])$$

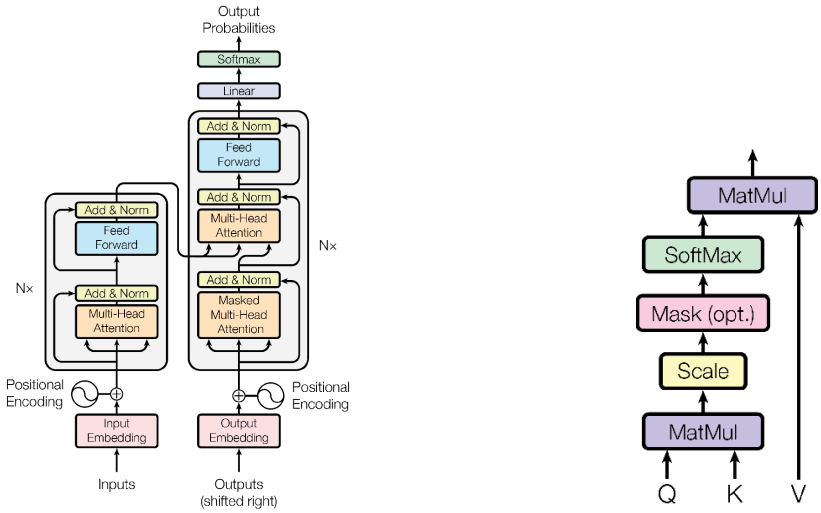
El vector $\widetilde{h_t}$ contiene información de todos los estados ocultos del codificador, atenuando los problemas de localidad de las redes recurrentes. Esta técnica se convirtió en parte integral de las tareas de traducción automática, resumen automático, entre otras. La Figura 2.4 ilustra esta arquitectura.

La técnica de auto-atención o intra-atención [118] (self-attention en inglés) consiste en aproximadamente la misma idea que la atención sólo que teniendo una única secuencia; podemos asumir ecuaciones similares con $\overline{h_i} = h_i$. La auto-atención genera representaciones de los distintos vectores de entrada observando la totalidad de la secuencia, a diferencia de las redes recurrentes que sólo construyen una representación en base al paso anterior. Esta capa es utilizada en arquitecturas para clasificación de texto encima de una capa recurrente para generar representaciones con dependencias sin distinción de la distancia entre los distintos tokens.

2.4.3. Transformers

Mencionamos el sesgo de la secuencialidad como uno de los problemas de las redes recurrentes. Otro de los grandes obstáculos para las arquitecturas autorregresivas es la paralelización. El cómputo secuencial donde h_t se calcula en base a h_{t-1} inhibe un cálculo paralelo, donde las diferentes representaciones puedan ser generadas simultáneamente.

[7] global, en Luong et al. [97] se menciona el mecanismo local que no consideramos



(a) Arquitectura de modelos Transformer *(b)* Mecanismo de scaled self-attention

Fig. 2.5: Modelo de transformador y su versión de auto-atención. La subfigura 2.5a muestra la arquitectura de los codificadores y decodificadores. Fuente: Vaswani et al. [160]

Parikh et al. [118] es uno de los primeros trabajos que proponen una arquitectura para la tarea de inferencia (NLI) enteramente basada en una arquitectura de atención, sin ningún tipo de recurrencia.

Vaswani et al. [160] introdujeron la arquitectura **Transformer** para la tarea de traducción automática. Esta arquitectura no utiliza capas recurrentes ni convolucionales, basándose enteramente en el mecanismo de auto-atención. La Figura 2.5a muestra la arquitectura de los modelos basados en Transformer, organizado en forma de encoder-decoder, con 6 capas de cada uno.

Cada capa del encoder utiliza un mecanismo de auto-atención múltiple seguido de una capa feed-forward punto a punto. Las dos capas de auto-atención o feed-forward están sucedidas por conexiones residuales [72] para facilitar el flujo del gradiente en una arquitectura profunda y una capa de normalización, de manera que la salida se expresa como:

$$\text{Layer}(x) = \text{Norm}(x + \text{subLayer}(x))$$

Las capas decodificadoras son similares, salvo que se les agrega una capa extra de auto-atención donde se combinan las salidas del encoder con las representaciones que genera el decoder. A su vez, las capas de multi-atención están enmascaradas para no poder "ver" las representaciones que se generan en pasos posteriores para guardar su naturaleza secuencial en la tarea.

El cálculo de atención utilizado en este trabajo es similar al visto en la ecuación 2.5, aunque normalizado por $\sqrt{d_k}$, donde d_k es la dimensión de los vectores de entrada:

$$\mathit{Attention}(Q, K, V) = \text{softmax}(\frac{Q^T K}{\sqrt{d_k}})V$$

Cada capa utiliza varias cabezas de auto-atención, cuyas salidas son concatenadas y proyectadas. A su vez, la salida de cada una de las capa pasa por una regularización de tipo dropout [155].

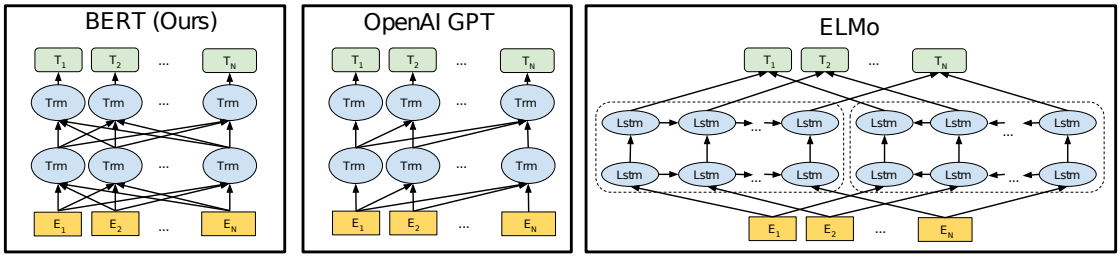


Fig. 2.6: Comparación entre ELMo, GPT y BERT. ELMo genera vectores contextualizados mediante dos modelos de lenguaje recurrentes (uno de izquierda a derecha y el otro al revés), . GPT pre-entrena un modelo de lenguaje basado en Transformers. BERT genera representaciones bidireccionales

Un punto a mencionar de los Transformers es que, siendo que no tienen ningún tipo de recurrencia y convolución, carecen de cualquier ordenamiento de la secuencia de tokens. Para inyectar ese conocimiento en la red, utilizan *vectores de posicionamiento* (positional embeddings) que se suman a los vectores de entrada de la capa de embeddings, como se ilustra en la Figura 2.5a. Estos vectores no son parámetros entrenados (como sí lo son en *BERT*) sino que se calculan mediante funciones sinusoidales.

No nos extenderemos más en la explicación de esta arquitectura, y referimos para más información a los excelentes artículos *Transformers from Scratch* [8], *Annotated Transformer* [9] y *The Illustrated Transformer* [10].

2.4.4. GPT, BERT y amigos

Combinando las ideas de ULMFit –entrenamiento semi-supervisado sobre la tarea de modelado de lenguaje– y la arquitectura Transformer –paralelización del cálculo mediante auto-atención – en Radford et al. [135] se introduce GPT (*generative pre-training*). Esta técnica consiste de un pre-entrenamiento sobre un gran corpus no etiquetado seguido de un fine-tuning discriminativo para cada tarea, muy en la línea de Howard and Ruder [77] introduciendo unos pocos parámetros específicos para cada una de estas. El modelo que usa esta tarea es el de **modelado de lenguaje causal** – es decir, de izquierda a derecha. GPT obtuvo el estado de arte para el benchmark GLUE [161], superando a ELMo y otros.

BERT [45] (Bidirectional Encoder Representations from Transformers) plantea una modificación sobre GPT: en lugar de pre-entrenar sobre la tarea de modelado de lenguaje **causal** –de izquierda a derecha– hacerlo sobre la tarea de modelado de lenguaje **enmascarado**. Esta tarea (usualmente llamada *Cloze task* [158]) consta de enmascarar una cierta cantidad de palabras de una frase, y luego intentar predecir las palabras faltantes. Por ejemplo, en la siguiente frase, consta de reemplazar los dos tokens `[MASK]`:

El `[MASK]` es celeste y el pasto `[MASK]`

A diferencia de la tarea de modelado de lenguaje causal, los autores argumentan que esta tarea permite generar representaciones bidireccionales ricas. La Figura 2.6 muestra una comparación entre los distintos tipos de pre-entrenamiento de GPT.

[8] `http://peterbloem.nl/blog/transformers`
[9] `https://nlp.seas.harvard.edu/2018/04/03/attention.html`
[10] `https://jalammar.github.io/illustrated-transformer/`

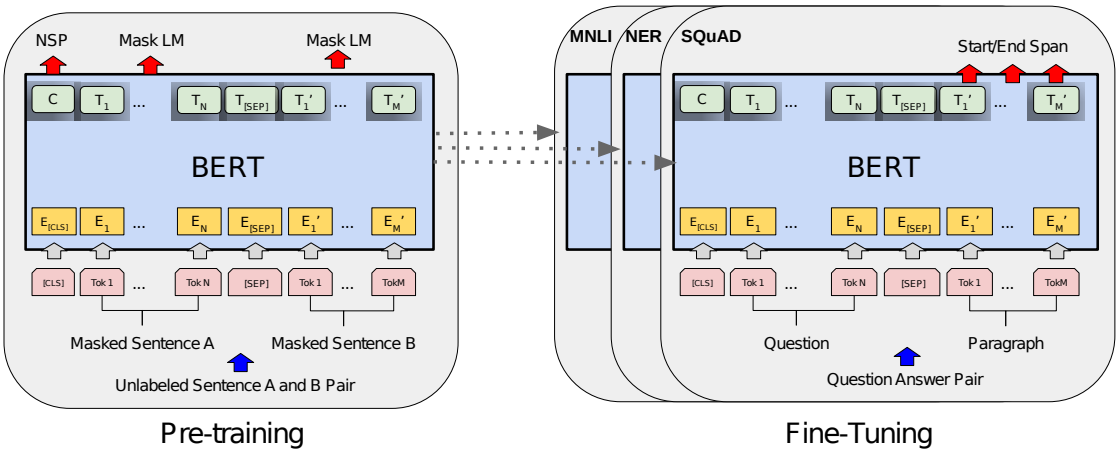


Fig. 2.7: Pre-entrenamiento y fine-tuning de BERT para distintas tareas. Fuente: Devlin et al. [45]

BERT es pre-entrenado conjuntamente sobre dos tareas: una, la ya mencionada tarea de modelado de lenguaje enmascarado; la otra, la tarea de *predicción de próxima oración* (Next Sentence Prediction, NSP). Esta tarea consiste en predecir si, dado un par de oraciones, la segunda es la que sigue a la primera. El 50 % de las ocasiones, las dos oraciones de entrada son contiguas en el texto de origen, y el 50 % restante son dos oraciones aleatorias concatenadas. Esta tarea debiera guardar cierta relación con la semántica de las oraciones y su interrelación, necesaria en tareas como NLI y Question Answering.

Dos caracteres especiales son utilizados en *BERT*: `[CLS]` y `[SEP]`. Estos caracteres se utilizan para delimitar las oraciones en la entrada de la red, y también para separar las dos oraciones de entrada. Durante el pre-entrenamiento, la representación generada por `[CLS]` es utilizada como la predicción de la tarea NSP y la representación de cada token enmascarado es utilizado como entrada de una capa softmax para predecir el token faltante. Para la mayoría de las tareas de clasificación de texto, `[CLS]` es usado como la representación de la (o las) oraciones de entrada. La Figura 2.7 ilustra esta metodología para la tarea de QA, ligeramente más compleja que la clasificación de texto.

BERT utiliza WordPiece [169], un algoritmo de tokenización muy similar a BPE [151], con un vocabulario de largo 30,000 para separar su entrada en subpalabras de manera eficiente. A su vez, para cada posición de su entrada entrena embeddings posicionales (a diferencia de los embeddings posicionales fijos originales de Vaswani et al. [160]) con un límite de 512, y dos embeddings especiales para la primer oración y la segunda oración de la entrada. Los vectores que ingresan a la primer capa de transformers son la suma de los embeddings de cada token, los embeddings posicionales, y los embeddings de oración.

El proceso de pre-entrenamiento es realizado sobre la concatenación de dos corpus: BooksCorpus [179] y la versión en inglés de Wikipedia. Estas dos fuentes son utilizadas ya que permiten extraer pares de palabras contiguas, algo necesario para la tarea de NSP.

Liu et al. [95] proponen dos modificaciones al pre-entrenamiento de *BERT*: en primer lugar, remover la tarea de NSP; y en segundo lugar, realizar un pre-entrenamiento más extenso y con batch sizes más grandes, pasando de lotes de 512 a 8,192 oraciones. Este modelo pre-entrenado (al cual se denomina *RoBERTa*) obtuvo mejor desempeño que *BERT* en el dataset de GLUE y otras tareas.

Luego de estos modelos de lenguaje, una suerte de guerra armamentística tuvo lugar para entrenar arquitecturas más grandes y con más parámetros al observar que aumen-

tando la cantidad de estos mejoraba la performance en distintas tareas – sin observarse aún un techo más que los recursos computacionales y energéticos disponibles en el planeta. Sólo para ilustrar el punto, la versión *base* de *BERT* tiene 110M parámetros, su versión large 330M, GPT-2 1,500M, Turing NLG de Microsoft 17,000M y finalmente GPT-3 [28] tiene la asombrosa cantidad de 175,000 M parámetros.

Parte II

EXTRACCIÓN DE OPINIONES DE REDES SOCIALES Y DETECCIÓN DE DISCURSO DE ODIO

3. EXTRACCIÓN DE OPINIONES DE TEXTOS SOCIALES

La extracción de opiniones en distintos espacios virtuales ha atraído mucho interés desde los comienzos de la Web 2.0. Inicialmente motivados por fines puramente comerciales, diferentes oportunidades han surgido mediante al desarrollo de la técnica y la mayor interacción de usuarios en Internet: desde fines sociológicos –como el análisis de discurso de odio o las reacciones a la pandemia– hasta políticos – como observar cuál es la opinión general sobre tal o cual candidato o sobre un tema candente. A partir de los años 2000, y debido a la combinación del desarrollo de métodos de aprendizaje estadístico y la cantidad creciente de datos disponibles generados por usuarios en Internet, numerosos trabajos han analizado este tipo de textos para poder extraer conocimiento **subjetivo** de quienes vuelcan sus pensamientos en las redes sociales y otros espacios virtuales.

Debido a la inmensa cantidad de contenido generado en diversos sitios y redes sociales (se estima que en el mundo se generan 500 millones tweets por día para 2021[1]), esta tarea es difícil de realizar sin algún tipo de automatización. Para ello, muchísimo esfuerzo se ha volcado en utilizar técnicas de aprendizaje automático para poder extraer este tipo de información de los textos creados por usuarios. El avance de las técnicas de NLP –como hemos descrito en el capítulo anterior– han permitido avanzar sobre este terreno; sin embargo, muchas de las limitaciones actuales del área junto a las dificultades particulares de las interacciones en medios sociales hacen esta tarea difícil.

En este capítulo hacemos una breve introducción al análisis de sentimientos o extracción de opiniones sobre textos de redes sociales. Esto es, dado un texto generado por un usuario (un post en Facebook, Instagram, un tweet, etc) predecir alguna característica discreta de éste, como por ejemplo si es un texto positivo o negativo, si tiene algún tipo de emoción de ira, alegría, u otra; si contiene discurso de odio contra algún grupo o no; si es irónico; entre otras. En base a datasets en español para distintas tareas, presentamos modelos de clasificación basados en técnicas del estado del arte.

Analizamos también algunas cuestiones relacionadas a la adaptación de dominio y representaciones generadas sobre dominios de textos generados por usuarios, planteando algunas líneas que retomaremos en capítulos posteriores.

3.1. Motivación

Las motivaciones para extraer opiniones subjetivas de usuarios en Internet son múltiples, aunque intentaremos categorizarlas en algunos grupos de notable interés. Dado el aumento considerable de contenido generado por usuarios desde la popularización de la WWW –y subsiguientemente con la explosión de las Redes Sociales– una de las cuestiones que motoriza este área es netamente comercial: ¿qué opinan los usuarios sobre este nuevo producto? ¿cuáles creen que son sus falencias? ¿qué tal es el servicio en tal o cual Restaurant? Desde ya más de 20 años, numerosos sitios y aplicaciones brindan la posibilidad de que los clientes vuelquen sus opiniones al respecto de los productos que consumen en sus plataformas. Para citar unos ejemplos, IMDb permite agregar comentarios sobre películas, Google Maps sobre distintos sitios –tanto turísticos como locales comerciales–, o los distintos sitios de venta minorista como MercadoLibre, eBay, o Amazon sobre los productos

[1] Fuente: `https://www.internetlivestats.com/twitter-statistics/`

que compran sus usuarios. Sobre esta información disponible, una encuesta de 2008 [75] reportó que cerca del 81 % de los usuarios de Internet de entonces (60 % de los habitantes de Estados Unidos) realizaron investigación online antes de sus compras, aunque también reportaban problemas a la hora de encontrar información valiosa para sus fines.

Con la explosión de las redes sociales, nuevas oportunidades y posibilidades de preguntas a contestar se abrieron mediante la extracción de opiniones en este medio [2]. Uno de estos horizontes, que es de interés particular para esta tesis, es el de las preguntas de carácter sociológico y político. Preguntas que pueden suscitar interés dentro de este punto pueden ser:

- ¿cuál es la opinión de los usuarios acerca de la legalización del aborto en cierto país que tiene en tratamiento este tema? ¿es representativo de la población en general? [64]
- ¿cómo se ha modificado el humor social de acuerdo a crisis económicas o pandemias como la del COVID-19? [102]
- ¿qué características tiene el discurso de odio contra los inmigrantes en España en el medio del auge de la ultraderecha en dicho país? [30]
- ¿qué artículos periodísticos suscitan la mayor cantidad de discurso discriminatorio en las redes sociales?
- ¿cuáles son las principales vulnerabilidades e intereses de ciertos sectores de la población? [180] [3]

entre otras. Estos tópicos son de gran interés para investigadores y políticos. Usualmente, la forma más estandarizada de acceder a la opinión de distintos actores sociales ha sido la de encuestas; sin embargo, la recolección y extracción automática de opiniones de medios virtuales brinda una alternativa (a veces) más económica y masiva aunque con un sesgo poblacional distinto al de otras metodologías.

3.2. Clasificación de textos sociales

Muchos de estos problemas de extracción de opiniones se pueden plantear como tareas de clasificación de texto [117]. El análisis de polaridad se puede plantear como una tarea de predecir si un texto tiene un sentimiento positivo, negativo, o neutro. El análisis de emociones se puede plantear (entre otras formas) como la de predecir la emoción predominante en el texto sobre un conjunto de seis posibles emociones.

Algunas variantes de estos problemas se pueden dar en el contenido analizado. El *Análisis de Sentimiento basado en aspectos* (usualmente denominada *ABSA* en la literatura por sus siglas en inglés) es una variante de la clasificación de polaridad en la que queremos predecir el sentimiento de un texto para cierto aspecto [120]; por ejemplo, en la oración "lindo lugar, la comida está muy bien pero la cerveza es horrible" (en una posible reseña de

[2] Si bien algunas preguntas de carácter sociológico tuvieron lugar con anterioridad, podemos marcar el uso intensivo de Facebook y Twitter como el comienzo de un estudio más sistemático de ellas dado el enorme volumen de datos accesibles para los investigadores

[3] Esto, según parece, fue utilizado en el affaire de Cambridge Analytica en las elecciones de 2016 que consagraron a Donald Trump en EEUU

un restaurant) podemos identificar dos sentimientos distintos: uno positivo para la comida y otro negativo para la cerveza. Dentro de estos problemas que complejizan la entrada, podemos contar algunos de carácter multimodal: en Sharma et al. [152] se plantea un problema de análisis de emociones para memes donde la entrada (el contenido social) consta de imágenes y texto, y se intenta predecir la emoción predominante.

Análogamente, se puede agregar cierta complejidad en la salida. El *Stanford Sentiment Treebank (SST)* [154] plantea una tarea de análisis de polaridad asignando una escala de Likert [93] donde cada comentario está etiquetado como muy negativo, algo negativo, neutral, algo positivo o muy positivo. Así mismo, otra posibilidad es la de predecir conjuntamente varias variables: por ejemplo, predecir si un comentario es discriminatorio, si es dirigido a un grupo o una persona, y si es agresivo, como el dataset de hatEval [10]; o bien, dado un comentario de una nota periodística, predecir las características que discrimina si es que hay alguna (como ser a las mujeres, al colectivo LGBTI, por motivos raciales, etc). De este último ejemplo hablaremos en los capítulos 5 y 6.

3.3. Trabajo previo

El análisis de sentimientos, opinion mining u opinion extraction suscitó interés casi desde el comienzo de la generación masiva de contenido de parte de los usuarios en la WWW. Particularmente desde la eclosión de las redes sociales, la inmensa cantidad de contenido generado por usuarios ha sido una fuente de información sin antecedentes para la extracción de todo tipo de opiniones. La bibliografía que comprende este tema es demasiado extensa y escapa los objetivos de esta tesis centrada en la detección de discurso de odio. Mencionamos algunos estudios exhaustivos relevantes y una pequeña selección de trabajos a continuación.

Pang and Lee [117] ofrecen un amplio repaso sobre los usos, aplicaciones, técnicas y dificultades de la extracción de opiniones en la era pre-deep learning y pre-redes sociales, mencionando cuestiones como las dificultades que los diferentes dominios presentan a las técnicas del entonces estado del arte. El trabajo de Pak and Paroubek [116] es uno de los pioneros en plantear a Twitter como una fuente de mensajes para la extracción de opiniones – en particular, para analizar polaridad de mensajes – proponiendo una metodología para recolectar y etiquetar datasets sobre esta red social. Yue et al. [173] presentan un racconto de los diversas tareas de análisis de sentimientos aplicados a textos de redes sociales y las técnicas para atacarlo.

Dentro de los recursos para tareas de opinion mining, Maas et al. [99] presenta un dataset de análisis de polaridad con dos etiquetas (positivo y negativo) sobre películas en la plataforma de IMDb, extensamente utilizado para la tarea de análisis de sentimientos. El Stanford Sentiment Treebank (SST) [154] es un dataset que contiene información granular en una escala símil Likert de polaridad sobre las distintas subpartes de cada oración. Esta tarea forma parte del benchmark de General Language Understanding Evaluation (GLUE) [161]. El workshop SemEval [4] ha generado numerosos recursos para tareas de opinion mining en redes sociales, como Análisis de Polaridad, Análisis de Polaridad basado en aspectos, análisis de emociones, entre otras.

Centrándonos en los recursos y tareas en español, uno de los principales polos de esto es el Taller de Análisis de Sentimientos (TASS) [40, 57, 101] organizado por la Sociedad Española de Procesamiento Natural (SEPLN) y a partir de 2020 en el marco del evento

[4] `https://semeval.github.io/`

Tarea	Dataset	#Mensajes	Clases	
Análisis de Sentimientos	TASS 2020	14,509	Neg	39.8 %
			Neu	29.5 %
			Pos	30.7 %
Análisis de Emociones	EmoEvent	8,409	Otra	49.0 %
			Alegría	21.6 %
			Tristeza	12.0 %
			Ira	10.2 %
			Sorpresa	4.1 %
			Disgusto	1.9 %
			Miedo	1.1 %
Detección de Ironía	IroSVa 2019	9,000	No irónico	66.7 %
			Irónico	33.3 %

Tab. 3.1: Tareas evaluadas en este capítulo, junto a datos estadísticos de los datasets utilizados

Iberian Languages Evaluation Forum (IberLEF). En este foro se presentaron tareas y datasets de análisis de sentimiento [57], de emociones [131], de toxicidad [171], entre otras.

3.4. Tareas analizadas

Tres tareas de extracción de opiniones sobre redes sociales fueron utilizadas como benchmark para las diferentes técnicas de clasificación. La Tabla 3.1 contiene información sobre las tareas analizadas y los datasets utilizados para ellas. Una de las tareas es la de **análisis de polaridad**: dado un tweet, detectar si tiene una polaridad general positiva, negativa, o neutra. Utilizamos el dataset de TASS 2020 [57], anotado con estas 3 clases y con información de las diferentes variedades dialectales del español a la que pertenece cada tweet. Para nuestro análisis, ignoramos estas distinciones y fusionamos todos los datos en un solo conjunto de datos (con las tres particiones correspondientes de entrenamiento, validación, y test).

Para el **análisis de emociones**, también usamos el conjunto de datos de TASS 2020 *EmoEvent* [131]. Este dataset multilingual (español e inglés) contiene tweets etiquetados con las seis emociones básicas de Ekman :*ira*, *disgusto*, *miedo*, *alegría*, *tristeza*, *sorpresa* y también una emoción *neutral* [47]. El dataset fue recolectado en base a a ocho eventos globales de diferentes dominios (políticos, entretenimiento, catástrofes o incidentes, conmemoraciones globales, etc.) por lo que las emociones siempre están relacionadas con un fenómeno en particular. Solo conservamos la parte en español, que contiene 8,409 tweets.

La **detección de ironía** es una tarea que ha ganado popularidad recientemente. Algunos trabajos han mostrado que tiene importantes implicaciones en otras tareas de NLP de caracter semántico: para citar uno, Gupta and Yang [67] muestran que el uso de funciones derivadas de la detección de sarcasmo mejora el rendimiento de ciertos modelos en la tarea de análisis de sentimientos. Además de esto, el contenido generado por los usuarios es una rica y vasta fuente de ironía, por lo que esta tarea es de particular importancia para el dominio de las redes sociales. IroSVa

Tarea	Clase	Ejemplos
Emociones	Neutral	Espectantes para ver el tercer capitulo de HASHTAG. El principio empieza bien.
	Alegría	Lo de Messi ha sido increíble! HASHTAG
	Tristeza	Un día lamentable. Se perdieron años de historia, de cultura, de arquitectura... Me siento devastada al ver las imágenes del incendio de la catedral de HASHTAG en HASHTAG URL
	Ira	URL que discurso para ponerlo a llorar, Putos humanos, para cuando la extinción??
	Sorpresa	Santa Maria Madre de Dios HASHTAG URL
	Miedo	Joder, izquierda venció ¿y qué? A mi me preocupa y mucho que Vox haya pasado de 0 a 24!! ¿a nadie le parece un montón? No sé, debo ser idiota. HASHTAG
	Disgusto	Como se nota que HASHTAG ya no tiene el apoyo de los libros. Vaya mierda de temporada se están sacando.
Sentimientos	Negativo	que triste es la realidad
	Positivo	Hola a todos corazones, buen día Y FELIZ NAVIDAD A TODOS USTEDES! DIOS ME LOS BENDIGA Y ME LOS LLENE DE TODO SU AMOR
	Neutral	@AlfonsoEmilioL La próxima vez que no vea lo entrevisto y pregunto que escucha.
Ironía	Irónico	Pues, noticia de última hora, eres un anormal que no sabe escribir. Es LESIONARAN. El juez paraliza la exhumación de Franco alegando peligro para los operarios. Pues me parece muy bien, porque con las ratas hay que ser muy precavido.
	No irónico	No me sirvió para nada todo fue en vano. Espero mejore la calidad.

Tab. 3.2: Ejemplos de instancias para las distintas tareas y clases.

[114] es un dataset en Español (publicado en el contexto de TASS 2019) que tiene la particularidad de considerar los mensajes no como textos aislados sino con un contexto dado (un titular o un tema). Consta de 7,200 instancias y 1,800 ejemplos de prueba divididos en tres variantes geográficas de Cuba, España y México, cada una con una etiqueta binaria que indica si el comentario contiene ironía o no. A diferencia de las dos tareas anteriores mencionadas aquí, este conjunto de datos contiene no solo mensajes de Twitter, sino también de comentarios de noticias y foros de debate como 4forums.com y Reddit.

La Tabla 3.2 ilustra algunos ejemplos seleccionados para las distintas tareas y sus clases. En el caso de la tarea de detección de emociones, podemos ver que algunos ejemplos han sido preprocesados por sus autores para ocultar los hashtags y urls.

3.5. Normalización y preprocesamiento

Una de los pasos más importantes para la manipulación de texto proveniente de redes sociales es el preprocesamiento. Con esto nos referimos al conjunto de técnicas dedicadas a disminuir la variabilidad del texto y aproximarlo a una forma lo más normal posible, aún cuando ciertos autores discuten la existencia de tal forma [46]. El texto generado por usuarios en medios informales suele ser más irregular que el texto proveniente de otras fuentes, con errores ortográficos, usos coloquiales y otros usos que hacen difícil el tratamiento por algoritmos de NLP. Para poner un ejemplo, la frase "¡qué lindo día, loco!" puede ser representada de las siguientes maneras:

- q lindo día loco
- k lindo diaaaaaaaa loco
- ke lendo diaa lk

entre otras formas posibles. Uno de los primeros trabajos que aborda este problema para el dominio de redes sociales es el de Han and Baldwin [70]. En base a un dataset de tweets, observaron que las palabras fuera de vocabulario (OOV en inglés) [5] en dicha red social tienen una alta frecuencia de incidencia. Ejemplos de estas palabras son neologismos, errores ortográficos, typos, contracciones típicas de esta red social (lk), sustituciones fonéticas (wacho en vez de guacho), entre otras. Este problema de la desnormalización del texto generado por usuarios planteaba un serio inconveniente para los métodos del estado del arte de ese entonces basados en bolsas de palabras o representaciones sobre palabras aisladas [6]. Para mitigar la alta dimensionalidad que generan estas palabras fuera de vocabulario, los autores propusieron diversas estrategias para normalizar las palabras y testean sus métodos sobre datasets de Twitter y SMS.

El trabajo de Eisenstein [46] trató desde una perspectiva más amplia los enfoques utilizados hasta el momento para tareas en medios sociales, planteando dos posibilidades: la **normalización** sería una forma de adaptar el texto a las herramientas, mientras que la **adaptación de dominio** sería adaptar las herramientas al texto. Con lo primero, el trabajo menciona al conjunto de técnicas que podemos utilizar para acercar la distribución del texto lo más posible a un dominio formal, mientras que por adaptación de dominio a la construcción de conjuntos de datos y algoritmos particulares para distintas tareas de NLP en redes sociales, como por ejemplo POS tagging [60], NER [139], entre otros.

La alta desnormalización siguió siendo un escollo para los modelos basados en redes neuronales y embeddings (como *GloVe* o *word2vec*) ya que cada representación se calcula sobre las palabras o tokens de la oración. En el caso de un elemento fuera de vocabulario, un mecanismo habitual es asignarles un token especial "<unk>", y por lo tanto, una única representación para todas esas palabras OOV. Sin embargo, esto puede ser problemático ya que elimina muchas palabras similares a otras que

[5] fuera del vocabulario de un diccionario estándar de GNU en inglés

[6] Recordemos que para el momento de la publicación de este trabajo aún no se usaban redes neuronales, word embeddings, ni mucho menos métodos más avanzados como *fasttext*, que ayuda mucho en las palabras OOV

sí tenemos en el vocabulario. Bojanowski et al. [24] propuso una solución a esto al permitir formar la representación de cada palabra mediante una combinación lineal de las representaciones de las "subpalabras" de cada una (ver Sección 2.3).

Con el advenimiento de los modelos basados en transformers, otros tipos de tokenización fueron propuestos que permiten reducir las palabras OOV. Word Piece [150] y Sentence Piece [88] son algoritmos de tokenización que, en lugar de partir la palabra en tokens o n-gramas de caracteres fijos como *fasttext*, convierten cada palabra en una tira de subpalabras provenientes de un vocabulario. Estos vocabularios no están prefijados sino que son entrenados sobre conjuntos de datos con variantes de Byte-Pair Encoding (BPE) [151] que intentan minimizar la cantidad de tokens utilizados para representar el texto. Esta técnica permite reducir la incidencia OOV notablemente, ya que muchas de estas subpalabras entrenadas guardan relación morfológica con el idioma (y el dominio) de los datos de entrenamiento.

Nguyen et al. [113] plantearon experimentos en tareas sobre texto proveniente de redes sociales usando dos formas de normalización: una **débil**, donde sólo convirtieron nombres de usuario en un token especial `@USER` y a las URLs en otro token especial `HTTPURL`, y otra estrategia **fuerte** donde utilizaron diccionarios de normalización y otras técnicas para reducir la variabilidad del texto basadas en Han and Baldwin [70]. Para un conjunto de tareas de clasificación sobre Twitter y distintos modelos pre-entrenados, los resultados de los experimentos arrojaron que la normalización fuerte empeora levemente la performance.

Teniendo estas consideraciones en cuenta, adoptamos una estrategia similar a la normalización **débil** mencionada en el trabajo recién mencionado:

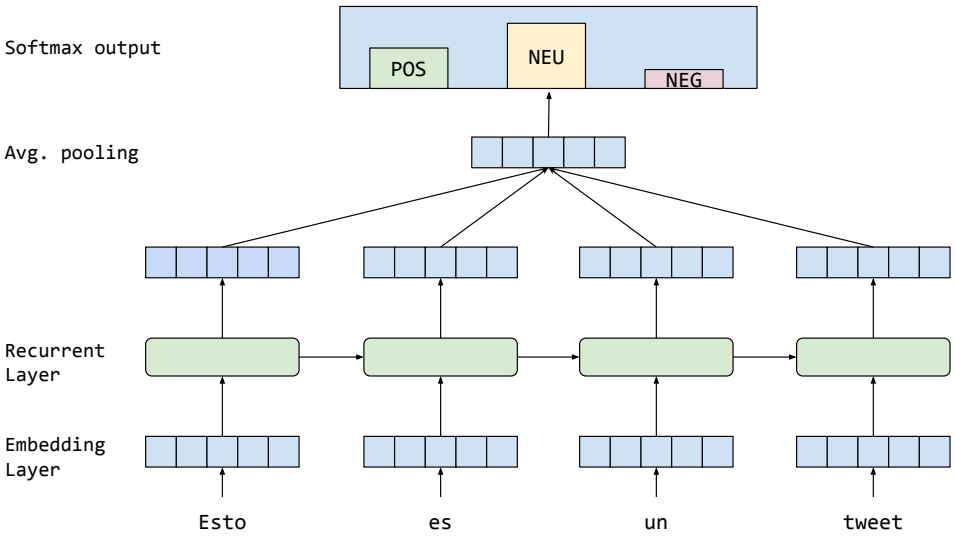
- Convertimos los handles a un token especial `@usuario`
- Convertimos las URLs a un token especial `URL`
- Convertimos los emojis a representaciones textuales usando la librería *emoji* [7]
- Normalizamos risas ("jajajajjjjajaja" lo convertimos a "jaja")
- Procesamos hashtags: #EsteHayQueNormalizar lo convertimos a *hashtag esto hay que normalizar*, utilizando las mayúsculas dentro del hashtag
- Limitamos repeticiones de caracteres a 3 ocurrencias

Si bien en algunos fragmentos de esta tesis usamos variaciones de estos métodos (como por ejemplo en la Sección 4.5.1), en general seguiremos esta estrategia de normalización.

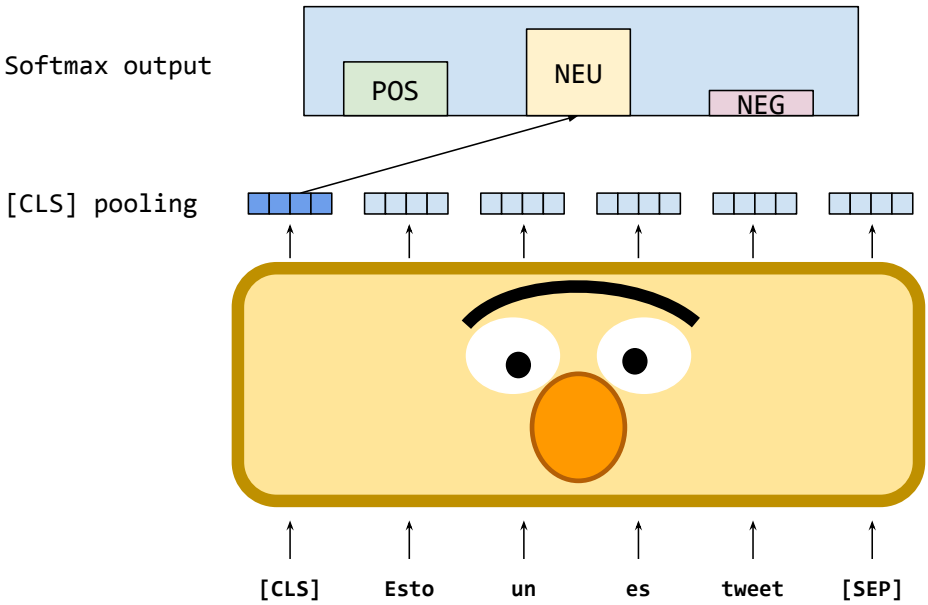
3.6. Modelos de clasificación

Describimos a continuación los clasificadores utilizados para las tareas. Podemos, a grandes rasgos, describir a todos nuestros clasificadores como compuestos de dos partes: un *codificador* (o *encoder* en inglés) que genera una representación continua

[7] `https://pypi.org/project/emoji/`



(a) Clasificador basado en redes recurrentes



(b) Modelos de clasificación basado en BERT y símiles

Fig. 3.1: Clasificadores propuestos para las tareas de Análisis de Polaridad, Análisis de Emociones y Detección de Ironía. La subfigura 3.1a muestra la arquitectura del modelo recurrente, que usa una capa de embeddings basados en *fasttext* y codifica el tweet como el promedio de las salidas de la capa recurrente. La subfigura 3.1b muestra un clasificador basado en BERT, donde tomamos la salida del token `[CLS]` como la codificación del tweet. Ambos usan un decodificador softmax

de longitud fija del texto de entrada, y un *decodificador* que toma esa codificación y la convierte a la salida deseada.

Todos nuestros problemas son de clasificación múltiple: elegir exactamente una clase entre varias. Para ello, nuestro decodificador será de la forma $\text{softmax}(Wx+b)$, donde $W \in \mathbb{R}^{c \times h}$ es una matriz de pesos y $b \in \mathbb{R}^c$ un vector de sesgo, siendo c la cantidad de clases y h el tamaño de la codificación de entrada. Los clasificadores planteados difieren entonces en los codificadores. Propusimos las siguientes variantes:

- **FFN**: un perceptrón multicapa (feed-forward network) con una función de activación intermedia *ReLU*
- **GRU/biGRU**: una red neuronal recurrente donde la capa oculta es una Gated Recurrent Unit (GRU) unidireccional o bidireccional.
- **Transformers**: un modelo pre-entrenado de lenguaje basado en transformers. Consideramos los modelos *BETO* [31], *RoBERTa* en su versión en español [69], *BERTin* [8], como así también el modelo multilingual *mBERT*

La capa de entrada de los modelos **FFN** y **GRU/biGRU** fueron entrenados con embeddings no contextualizados basados en *fastText*. Utilizamos dos versiones de estas representaciones: las canónicas generadas por los autores, entrenadas sobre Common Crawl en español; y también una versión generada por nosotros, entrenada sobre tweets en español, cuya recolección describimos en la Sección 7.2.1.

La codificación final que utilizamos para la red **GRU/biGRU** se da como el promedio de los vectores salida de cada paso. Para los modelos basados en transformers, la codificación se da tomando la salida del caracter de inicio (`[CLS]`). Si bien podría también tomarse el promedio como en las redes recurrentes, los modelos basados en transformers no sufren el cuello de botella que se genera tomando la última representación de la red recurrente. La Figura 3.1 ilustra la arquitectura de los clasificadores recurrentes y basados en transformers.

Usamos un tamaño oculto de 512 para los modelos exceptuando los basados en Transformers –que suelen tener 768 como valor estándar– y para todos los casos una regularización del tipo dropout [155] de 0,1 sobre la codificación del tweet. Como algoritmo de optimización utilizamos *Adam* [84] con un learning rate de 0,001 y un decay de 0,01. Para los clasificadores basados en transformers, usamos Adam con un learning rate triangular de 10^{-5} y un warmup del 10 % de los pasos. Entrenamos todos los modelos por 5 epochs y nos quedamos con los modelos que mejor performance tengan sobre el split de validación en términos de la métrica correspondiente a la tarea.

3.7. Resultados

La Tabla 3.3 muestra los resultados obtenidos por los distintos modelos, expresados como la media de diez corridas de los experimentos de clasificación junto a sus desviaciones estándar. A su vez, reportamos un puntaje promedio entre todas las tareas. Puede observarse que los modelos basados de transformers obtienen un

[8] `https://huggingface.co/bertin-project/bertin-roberta-base-spanish`

Modelo	Polaridad	Emociones	Ironía	Puntaje
RoBERTa	$67{,}0 \pm 0{,}6$	$52{,}7 \pm 1{,}5$	$72{,}1 \pm 0{,}8$	66,46
BERTin	$66{,}6 \pm 0{,}5$	$52{,}4 \pm 0{,}7$	$71{,}3 \pm 1{,}2$	66,01
BETO_U	$65{,}1 \pm 0{,}6$	$53{,}2 \pm 1{,}2$	$70{,}1 \pm 0{,}7$	65,26
BETO_C	$66{,}2 \pm 0{,}5$	$51{,}6 \pm 1{,}2$	$70{,}5 \pm 0{,}9$	65,17
mBERT	$61{,}7 \pm 0{,}3$	$49{,}3 \pm 1{,}0$	$68{,}1 \pm 1{,}0$	62,73
biGRU_{TW}	$58{,}5 \pm 1{,}1$	$26{,}4 \pm 0{,}7$	$63{,}1 \pm 1{,}1$	51,80
biGRU_{CC}	$60{,}2 \pm 0{,}4$	$26{,}9 \pm 0{,}3$	$62{,}8 \pm 1{,}4$	50,94
GRU_{TW}	$55{,}3 \pm 0{,}8$	$23{,}1 \pm 0{,}6$	$62{,}5 \pm 0{,}9$	48,57
GRU_{CC}	$56{,}4 \pm 0{,}4$	$23{,}7 \pm 0{,}5$	$58{,}1 \pm 1{,}6$	47,44
FFN_{TW}	$53{,}8 \pm 0{,}5$	$20{,}2 \pm 0{,}2$	$57{,}9 \pm 1{,}2$	44,05
FFN_{CC}	$51{,}1 \pm 0{,}3$	$17{,}3 \pm 0{,}2$	$51{,}1 \pm 0{,}5$	39,67

Tab. 3.3: Resultados de la evaluación de los distintos modelos para las tareas analizadas (Análisis de Emociones, Análisis de Emociones, Detección de Ironía). Los resultados están dados en porcentajes de Macro F1, y expresados como la media de diez corridas junto a su desviación estándar. El puntaje de cada modelo es el promedio de las métricas para las tres tareas

rendimiento sustancialmente mejor que los basados en redes neuronales recurrentes. Entre los primeros, la versión española de *RoBERTa* obtiene marginalmente los mejores resultados.

Dentro de los clasificadores de redes recurrentes y feed-forward, aquellos que consumen embeddings entrenados en textos sociales (marcadas como $_{TW}$) tienen un rendimiento superior que aquellos que consumen embeddings entrenados en Common Crawl (marcados como $_{CC}$). Esta diferencia, en todos los casos, es estadísticamente significativa luego de realizar un test de Mann-Whitney U ($p \leq 0{,}05$ para el caso de ironía y *biGRU*, para todas las demás comparaciones $p \leq 0{,}001$).

3.8. Discusión

Para las 3 tareas planteadas, los clasificadores basados en modelos pre-entrenados de transformers obtuvieron mejores resultados que los basados en redes recurrentes y feed-forward. Como es esperable (y se observa en la literatura) los modelos monolinguales (*RoBERTa*, BERTin y *BETO*) tienen un rendimiento sensiblemente mejor el modelo multilingual *mBERT*. Dentro de los modelos de mejor performance se posiciona primero *RoBERTa*, aunque su mejora es pequeña respecto de *BETO*.

Algo que observamos es que, entre los modelos recurrentes y feed-forward que consumen word-embeddings, la utilización de representaciones entrenadas directamente sobre textos generados por usuarios resultan en una mejor performance de los clasificadores sobre estas tareas que tienen datos de ese mismo dominio. Si bien puede pensarse que el entorno pequeño o el texto ruidoso de los textos pueden ser un problema a la hora de construir representaciones, los experimentos realizados indican lo contrario. Retomaremos esta idea en el Capítulo 7, donde por un lado generamos un modelo basado en *RoBERTa* entrenado sobre tweets, y por otro lado realizamos un estudio comparativo de su performance contra un modelo *BETO*

adaptado a este nuevo dominio.

3.9. pysentimiento: un paquete de python para Análisis de Sentimiento

Algo que suele obstaculizar la utilización de herramientas de extracción de opiniones en redes sociales con fines de investigación es la dificultad a su acceso. O bien estos servicios están detrás de sistemas privados detrás de APIs con precios demasiado altos para los presupuestos académicos o están disponibles pero no en español u otros idiomas de bajos recursos [9]. En otros casos, estos recursos están disponibles pero no para ser usados de manera sencilla, lo cual es un escollo para investigadores que no sean expertos en NLP.

Como una pequeña contribución de esta tesis y con el objetivo de facilitar el acceso de estos recursos para la investigación, creamos el paquete **pysentimiento** [10]. Esta biblioteca provee modelos pre-entrenados y herramientas de preprocesado para textos sociales en español e inglés. Si bien tiene soporte multilingual, su eje es el de proveer recursos para el español que tiene una disparidad importante en recursos.

pysentimiento utiliza el model hub de *huggingface* [11], un repositorio de modelos de libre acceso. Allí es donde alojamos todos los modelos entrenados, tanto de sentimientos, emociones, y algunos más que serán discutidos a lo largo de esta tesis. Cada tweet que es analizado por la librería pasa primero por una etapa de preprocesamiento (siguiendo el proceso explicado en la Sección 3.5), y luego analizado por el modelo que nos brinda la salida correspondiente.

Al momento de escribir estas líneas, los modelos de **pysentimiento** se encuentran entre los más descargados de *huggingface* para el idioma español, dando cuenta de la necesidad de estas herramientas de libre acceso.

3.10. Conclusiones

En este capítulo hemos hecho una introducción a la extracción de opiniones usando técnicas de clasificación basadas en redes neuronales. Analizamos tres problemas de extracción de opiniones en Español: análisis de polaridad, análisis de emociones y detección de ironía. Presentamos el andamiaje básico para tareas de clasificación que utilizaremos en el resto de la presente tesis, puntualmente sobre el preprocesamiento de textos provenientes de redes sociales y arquitecturas básicas de clasificadores basados en redes neuronales.

Algo que observamos en nuestros experimentos es que las representaciones generadas sobre textos de este dominio mejoran el rendimiento de nuestros algoritmos de clasificación. Retomaremos estas ideas en el Capítulo 7 utilizando técnicas del estado del arte, concretamente modelos pre-entrenados de lenguaje.

En los siguientes capítulos centraremos nuestra atención en una tarea particular: la detección de discurso de odio.

[9] La definición de bajos recursos es subjetiva, pero tomando en cuenta la cantidad de hablantes nativos de español hay una desproporción abismal con otros idiomas

[10] `https://github.com/pysentimiento/pysentimiento`

[11] `https://huggingface.co/models`

3.11. Notas

Gran parte de este trabajo está basado en nuestra participación en TASS 2018 [101] resumida en Luque and Pérez [98]. Los resultados en esta sección no son comparables con los de ese trabajo ya que decidimos utilizar la versión del dataset de TASS 2020 [57] que unifica las dos posibles clases neutrales (*neutral* y *nula*) del dataset del mencionado trabajo.

Respecto a aquella publicación, omitimos el análisis de data augmentation mediante traducción bidireccional y nos centramos en dar una breve introducción al tema y en analizar el impacto de los embeddings generados en textos provenientes de redes sociales.

4. DETECCIÓN DE DISCURSO DE ODIO

El discurso de odio contra mujeres, inmigrantes y otros grupos protegidos es un fenómeno generalizado en la Internet y que resulta importante monitorear dada su potencial relación con actos violentos, como hemos comentado en la introducción de esta tesis. En los primeros días de la World Wide Web, algunos académicos se aventuraron a decir a que los prejuicios y el odio serían removidos en este espacio mediante la disolución de identidades en el ámbito virtual [92, 138]. Veinte años después de esta hipótesis, podemos decir que no ha sido el caso. La prevalencia del racismo en la "World White Web" y en las redes sociales ha sido estudiada en numerosos trabajos [1, 83], como así también la misoginia en el mundo virtual [52, 100], entre otros ataques discriminatorios.

Si bien el discurso racista y sexista es una constante en las redes sociales, muchos picos se documentan luego de eventos detonantes, como pueden ser asesinatos con motivos religiosos o políticos [29]. Debido a esto, algunos estados y organizaciones supranacionales han tomado cartas en el asunto instando a las empresas de redes sociales a que tomen medidas para bajar la incidencia del discurso de odio. Debido a la enorme cantidad de contenido generado por usuarios en estos medios, es necesario desarrollar herramientas que faciliten la labor humana en la detección y prevención de este fenómeno, con particular foco de aquel que incita a la violencia física.

En este capítulo hacemos una introducción a este problema desde varias ópticas. Analizamos las diversas definiciones de discurso de odio, realizando una breve reseña desde un marco legal y de tratados internacionales para luego centrarnos en este problema desde una perspectiva del procesamiento de lenguaje natural. En base al dataset de la competencia *HatEval* [10], analizamos de técnicas de detección de discurso de odio, algunas de ellas presentadas en Pérez and Luque [126]. Finalmente, marcamos algunos problemas en los enfoques actuales de la detección de discurso discriminatorio y algunas oportunidades de mejora que abordaremos en capítulos subsiguientes.

4.1. ¿Qué es el discurso de odio?

No existe una definición universalmente aceptada de lo que configura discurso de odio. Para intentar acercarnos lo más posible a este concepto, en esta sección haremos un repaso muy general de algunos tratados internacionales sobre la materia. Antes de continuar, hacemos **una aclaración:** en la normativa sobre derechos humanos muchas veces se encuentra delimitado el discurso **discriminatorio** del discurso de **odio**, siendo este último una subcategoría del primero de mayor intensidad y con incitaciones a la violencia contra grupos protegidos o individuos miembros de estos grupos. En la literatura de NLP sobre el tema se utiliza la expresión discurso de odio (*hate speech*) para referirse indistintamente a ambos fenómenos.

Aún cuando entendemos que la acepción general del discurso de odio puede entenderse como incorrecta desde la perspectiva de tratados internacionales, teniendo en cuenta que esta tesis está centrada en técnicas para su detección automática usa-

mos esta terminología para plegarnos a los usos y costumbres de la comunidad de NLP.

4.1.1. Abordaje desde una perspectiva legal y de los Derechos Humanos

Un principio general que hace a los derechos más elementales del hombre y a la vida en sociedad es la posibilidad de expresarse libremente, el **derecho a la libre expresión**. Este derecho está protegido por constituciones nacionales y numerosos tratados internacionales. Uno de estos tratados, el Pacto Internacional de Derechos Civiles y Políticos (*ICCPR* por sus siglas en inglés)[1], sancionado en 1966 en la Asamblea de las Naciones Unidas y ratificado por 166 países, incluye en su artículo 19:

> 1. Nadie podrá ser molestado a causa de sus opiniones.
>
> 2. Toda persona tiene derecho a la libertad de expresión; este derecho comprende la libertad de buscar, recibir y difundir informaciones e ideas de toda índole, sin consideración de fronteras, ya sea oralmente, por escrito o en forma impresa o artística, o por cualquier otro procedimiento de su elección.
>
> 3. El ejercicio del derecho previsto en el párrafo 2 de este artículo entraña deberes y responsabilidades especiales. Por consiguiente, puede estar sujeto a ciertas restricciones, que deberán, sin embargo, estar expresamente fijadas por la ley y ser necesarias para:
>
> a) Asegurar el respeto a los derechos o a la reputación de los demás;
>
> b) La protección de la seguridad nacional, el orden público o la salud o la moral públicas. (Artículo 19 de la ICCPR)

Este artículo que garantiza el derecho a la libertad de expresión da cuenta también de que esta libertad no es completamente irrestricta. El ejercicio de los derechos e igualdad ante la ley de otros marca este límite, no pudiéndose invocarse este derecho para avasallar los de terceros. La Convención Internacional sobre toda forma de Discriminación Racial (ICERD) [2] dice en su artículo 4 al respecto:

> Los Estados partes condenan toda la propaganda y todas las organizaciones que se inspiren en ideas o teorías basadas en la superioridad de una raza o de un grupo de personas de un determinado color u origen étnico, o que pretendan justificar o promover el odio racial y la discriminación racial, cualquiera que sea su forma, y se comprometen a tomar medidas inmediatas y positivas destinadas a eliminar toda incitación a tal discriminación o actos de tal discriminación y, con ese fin, teniendo debidamente en cuenta los principios incorporados en la Declaración Universal de Derechos Humanos, así como los derechos expresamente enunciados en el

[1] Este pacto desarrolla los derechos civiles y políticos establecidos por la Declaración Universal de los Derechos Humanos de la ONU

[2] `http://servicios.infoleg.gob.ar/infolegInternet/anexos/120000-124999/122553/norma.htm`

> artículo 5 de la presente Convención, tomarán, entre otras, las siguientes medidas:
>
> a) Declararán como acto punible conforme a la ley, toda difusión de ideas basadas en la superioridad o en el odio racial, toda incitación a la discriminación racial así como todo acto de violencia o toda incitación a cometer tal efecto, contra cualquier raza o grupo de personas de otro color u origen étnico, y toda asistencia a las actividades racistas, incluida su financiación;
>
> b) Declararán ilegales y prohibirán las organizaciones, así como las actividades organizadas de propaganda y toda otra actividad de propaganda, que promuevan la discriminación racial e inciten a ella y reconocerán que la participación en tales organizaciones o en tales actividades constituye un delito penado por la ley;
>
> c) No permitirán que las autoridades ni las instituciones públicas nacionales o locales, promuevan la discriminación racial o inciten a ella. (Artículo 4, ICERD)

Los Estados y otros organismos deben entonces tomar medidas para poder asegurar el libre ejercicio de los derechos y la igualdad de todos sus miembros, aún cuando esto pueda significar una restricción en la libertad de expresión [7]. Entendiendo entonces que este derecho tiene sus límites, podemos pensar que el discurso de odio es una de esas fronteras. Si bien este fenómeno es algo que no está completamente delimitado, repasamos algunas definiciones de este fenómeno hechas en tratados para acercarnos un poco más a las características comunes que comparten las diferentes definiciones. La Observación General 35 del Comité por la Eliminación de la Discriminación Racial de la ONU (CERD) considera que será discurso de odio, y debe ser tipificado penalmente:

> a) Toda difusión de ideas basada en la superioridad o en el odio racial o étnico, por cualquier medio;
>
> b) La incitación al odio, el desprecio o la discriminación contra los miembros de un grupo por motivos de su raza, color, linaje, u origen nacional o étnico;
>
> c) Las amenazas o la incitación a la violencia contra personas o grupos por los motivos señalados en el apartado anterior;
>
> d) La expresión de insultos, burlas o calumnias contra personas o grupos, o la justificación del odio, el desprecio o la discriminación por los motivos señalados en el apartado b) anterior, cuando constituyan claramente incitación al odio o a la discriminación;
>
> e) La participación en organizaciones y actividades que promuevan e inciten a la discriminación racial. (Recomendación 35 del Comité por la Eliminación de la Discriminación Racial, CERD)

En líneas generales, como se menciona en el reporte de la CIDH sobre discurso de odio contra lesbianas, gay, trans e intersex en Latinoamérica [37], el concepto

Fig. 4.1: Definición de discurso de odio de acuerdo al Toolkit de Article 19

usualmente es referido a expresiones que incitan a tomar algún tipo de medida hostil contra una víctima o un grupo de personas, siendo esta perteneciente a un determinado grupo social definido por alguna característica particular como ser la etnia, lenguaje, género, entre otras. Dicho esto, podría delimitarse el discurso discriminatorio del discurso de odio por la componente de la promoción e instigación de la violencia; sin embargo, para los fines de este trabajo utilizamos los términos indistintamente. Aún cuando el discurso no contenga arengas ni incitaciones a cometer actos violentos, puede entenderse ese discurso como generador de un ambiente hostil y de intolerancia que termine promoviendo estos ataques físicos [37].

Article 19 [7] condensa muchas de estas definiciones de una manera sucinta, desglosando esto en **odio** y **discurso**:

> 1. Odio: emoción intensa e irracional de oprobio, enemistad y aborrecimiento hacia una persona o grupo de personas, por tener determinadas características protegidas (reconocidas en el derecho internacional), reales o percibidas. El "odio" es más que un mero prejuicio y debe ser discriminatorio. El odio es una muestra de un estado emocional u opinión y, por lo tanto, se diferencia de cualquier acto o acción que se haya llevado a cabo. 2. Discurso: cualquier expresión que vierta opiniones o ideas, que comparte una opinión o una idea interna con un público externo. Puede adoptar muchas formas: escrita, no-verbal, visual o artística y puede ser difundida en los medios, incluyendo Internet, material impreso, radio o televisión. (Article 19: Hate Speech Toolkit)

En base a esta definición, puede entenderse al discurso de odio como un dis-

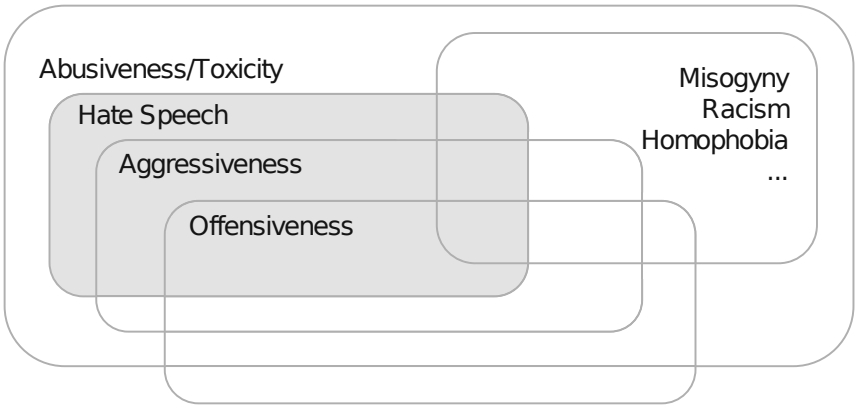


Fig. 4.2: Discurso de odio y conceptos relacionados. Fuente: Poletto et al. [132]

curso de cierta intensidad e irracionalidad que ataca a una persona o un grupo de personas por alguna característica históricamente vulnerada: por ser mujer, por su género, por su etnia, nacionalidad, religión, idioma, etc. La clave está en la combinación: un discurso irracional e intenso contra alguien que no posea una característica protegida no configura discurso de odio; por ejemplo, ataques a ciertas personas por ser periodistas. La Figura 4.1 ilustra esta definición.

No todo ataque a un individuo o una persona de algún colectivo discriminado es discurso de odio. En particular, la CIDH [37] menciona en base al informe de la UNESCO sobre discurso de odio [55] que:

> (...) el discurso de odio no puede abarcar ideas amplias y abstractas, tales como las visiones e ideologías políticas, la fe o las creencias personales. Tampoco se refiere simplemente a un insulto, expresión injuriosa o provocadora respecto de una persona. Así definido, el discurso de odio puede ser manipulado fácilmente para abarcar expresiones que puedan ser consideradas ofensivas por otras personas, particularmente por quienes están en el poder, lo que conduce a la indebida aplicación de la ley para restringir las expresiones críticas y disidentes. Asimismo, el discurso de odio tiene que distinguirse de aquellos "crímenes de odio" que se basan en conductas expresivas, como las amenazas y la violencia sexual, y que se encuentran fuera de cualquier protección del derecho a la libertad de expresión

Como vemos, no solamente es difusa la frontera fijada sobre qué es discurso de odio o insultos, sino que incluso también es difícil definir qué característica es protegida o no. En el siguiente capítulo hablaremos más de esto al describir los criterios utilizados a la hora de anotar un conjunto de datos sobre comentarios en Twitter.

4.1.2. Definiciones utilizadas desde NLP

Debido a las razones comentadas en el Capítulo 1, la comunidad de NLP ha observado en los últimos años un creciente interés en la investigación de técnicas automáticas para combatir el discurso de odio en redes sociales. Una de las primeras

Definición	Fuente
Cualquier comunicación que menosprecia a una persona o grupo en base a su raza, etnia, género, orientación sexual, nacionalidad, religión u otra característica	Warner and Hirschberg [162], Basile et al. [10]
Uso de insultos racistas o sexistas; ataques a minorías; promoción de discurso de odio o crímenes violentos; distorsiones o mentiras acerca de minorías; apoyo a hashtags racistas o sexistas; defender xenofobia o sexismo; contener nombre de usuario ofensivo	Waseem and Hovy [165]
Lenguaje que es usado para expresar odio hacia un grupo objetivo o que pretende ser despectivo, o humillar o insultar sus miembros	Davidson et al. [42]
Acto de ofender, insultar o amenazar a una persona o grupo de acuerdo a su religión, raza, casta, orientación sexual, género, o pertenencia a alguna comunidad estereotipada	Schmidt and Wiegand [149]
Contenido definido por: primero, su intención de diseminar odio, incitar a la violencia, o amenazar la libertad, dignidad o integridad de las personas; segundo, su objetivo, que debe ser un grupo protegido o un miembro de tal grupo	Sanguinetti et al. [146]

Tab. 4.1: Definiciones de discurso de odio para diferentes trabajos del área. Fuente: Poletto et al. [132]

dificultades de esta tarea es la mencionada dificultad para definir unívocamente este discurso, algo que ha provocado que los investigadores del área aún no tengan un marco teórico común acerca de su definición. Esta falta de una teoría unificada, a su vez, se debe en parte a que la investigación de técnicas automáticas de reconocimiento se encuentra aún en una etapa relativamente prematura [3], y a la relación de este discurso con otros fenómenos de las redes, como ser el lenguaje tóxico, grosero, entre otras distinciones, que también es de interés estudiar. La Figura 4.2 ilustra el marco teórico utilizado por Poletto et al. [132] en relación al discurso de odio: mientras éste es un subconjunto del lenguaje abusivo y tóxico, algunos discursos racistas o misóginos [4] no configurarían discurso de odio.

La Tabla 4.1 muestra las definiciones de discurso de odio utilizadas en algunos trabajos del área. Podemos ver que las definiciones tienen diferencias sustanciales, algunas siendo directamente operacionales, y algunas otras con matices respecto a la intención o su intensidad.

4.2. Trabajo previo

Hacemos a continuación una reseña de la literatura de la detección de discurso de odio y otros fenómenos similares. Un análisis exhaustivo de esta subdisciplina sería inviable debido a la enorme cantidad de trabajo del área, con un ritmo meteórico en los últimos años. Referimos para revisiones más extensivas a Schmidt and Wiegand [149] y Fortuna and Nunes [53]. Más recientemente, Poletto et al. [132] hacen un análisis pormenorizado y actualizado de los recursos existentes para esta tarea.

[3] la gran mayoría de los trabajos y recursos son de los últimos 5 años

[4] Los autores utilizan el término *microagresiones*

La detección del discurso del odio es una tarea de clasificación de textos relacionada con el análisis de sentimientos y ha sido estudiada para varias redes sociales [116, 145, 159]. Uno de los primeros trabajos al respecto es el de Greevy and Smeaton [65], quienes utilizan bolsas de palabras y Support Vector Machines para detectar contenido racista en páginas web, utilizando un dataset construido de manera semi-supervisada buscando sitios mediante keywords y sus links en motores de búsqueda. Siguiendo un enfoque similar, Warner and Hirschberg [162] usaron unigramas y Brown clusters [27] con SVMs para detectar mensajes antisemitas en Twitter.

Waseem and Hovy [165] anotaron un corpus y usaron técnicas basadas en n-gramas de caracteres para detectar discurso de odio en comentarios de Twitter. Badjatiya et al. [8] usaron el mismo conjunto de datos para entrenar modelos de aprendizaje profundo con embeddings ajustados a los datos, obteniendo mejoras sustanciales en el rendimiento para la tarea en cuestión aunque sujeto a algunos problemas de entrenamiento observado por otros trabajos [5]. Zhang et al. [178] entrenaron una red neuronal profunda que combina CNNs con Gated Recurrent Units [35], superando a los sistemas anteriores en varios conjuntos de datos de detección de discurso de odio. Anzovino et al. [4] recopilaron un corpus de tweets misóginos y propusieron una taxonomía para distinguirlos en diferentes categorías. A su vez, los autores mostraron que enfoques simples (como el uso de modelos lineales junto con n-gramas) logran un rendimiento competitivo en conjuntos de datos de pequeño tamaño.

En cuanto a las tareas compartidas, Fersini et al. [50] presentaron un dataset para la detección de misoginia en Twitter, tanto en español como en inglés, mientras que Fersini et al. [51] planteó un desafío similar pero en italiano e inglés. Bosco et al. [26] propuso un concurso de detección automática sobre publicaciones de Twitter y comentarios de Facebook, que incluía discursos de odio en general.

Una de las herramientas más utilizadas, no sólo para la detección de discurso de odio sino para la detección de contenido tóxico en general es Perspective API de Google, desarrollada originalmente por Jigsaw [5]. Esta API de acceso libre brinda un analizador muy potente para la detección de lenguaje tóxico, con información granular sobre los tipos de ataques. Algunos trabajos lo utilizan como algoritmo de detección en modalidad zero-shot, obteniendo mejores resultados que modelos entrenados sobre los propios datos [121]. Sin embargo, algunas de sus debilidades han sido marcadas mediante ejemplos adversariales, algo que obviamente es propio de las actuales limitaciones de las técnicas de NLP [76, 80]. Más aún, la información de grano fino –e.g. si es un ataque a un grupo protegido y a cuál se ataca– sólo está disponible para el inglés.

Dentro de los trabajos en español, del Arco et al. [43] evalúan distintos modelos pre-entrenados de lenguaje sobre la tarea de detección de discriminación usando dos datasets: el primero, Pereira-Kohatsu et al. [125] que consta de 6000 tweets, recolectado por el Estado Español para monitorear el discurso de odio en redes sociales; y el segundo, el dataset de SemEval 2019 Task 5 (*HatEval*) [10], presentado en contexto de una shared-task y que comprende ataques contra inmigrantes y mujeres.

[5] `https://developers.perspectiveapi.com/s/`

Categoría	Español			Inglés		
	Train	Dev	Test	Train	Dev	Test
No HS	2643	278	940	5217	573	1740
HS	1857	222	660	3783	427	1260
TR	1129	137	423	1341	219	529
AG	1502	176	474	1559	204	594
Total	4500	500	1600	9000	1000	3000

Tab. 4.2: Números del dataset de Basile et al. [10], por idioma y por partición. No HS representa los tweets que no tienen contenido odioso, HS aquellos que sí, TR aquellos que son individualizados, y AG aquellos que son agresivos. Entre paréntesis encontramos los porcentajes de incidencia, considerando TR y AG dentro de aquellos que son discriminatorios

4.3. Descripción del dataset utilizado

Utilizamos en este capítulo el dataset provisto por Basile et al. [10], presentado en SemEval 2019 y orientado a la detección de discurso de odio contra mujeres e inmigrantes en Twitter. Los autores recopilaron comentarios en inglés y en español de dicha red social mediante tres estrategias combinadas: monitoreando a las posibles víctimas de cuentas de odio; chequeando el historial de usuarios creadores de contenido discriminatorio; y filtrando contenido mediante palabras clave. A su vez, este trabajo distingue entre el discurso de odio dirigido a individuos y el discurso de odio genérico, y entre mensajes agresivos y no agresivos. En el Capítulo 5 construiremos un conjunto de datos contextualizado de discurso de odio en base a algunas de las limitaciones observadas sobre en este capítulo.

Las instancias del dataset poseen las siguientes etiquetas:

- **HS**: una etiqueta binaria que marca si el tweet tiene contenido discriminatorio contra mujeres o inmigrantes (0 si no lo tiene, 1 si hay discurso de odio)
- **TR**: Si hay HS, una etiqueta binaria que marca si el objetivo del discurso de odio es un objetivo genérico (0) o si se refiere a un individuo específico (1)
- **AG**: Si hay HS, una etiqueta binaria que marca si el tweet es agresivo

La Tabla 4.2 muestra los números para cada partición, cada idioma, y cada una de las etiquetas. Podemos observar entre los dos idiomas que, si bien la proporción de discurso de odio se mantiene muy similar (58 % vs 42 % aproximadamente), la proporción de discurso de odio individualizado (TR) y agresivo (AG) es notoriamente más alto para el español que para el inglés. Esto puede deberse, entre otras cosas, a distintas estrategias de recolección de los tweets. La Tabla 4.3 posee algunos ejemplos para cada una de las características en cuestión para la porción en español, que es la de nuestro interés.

4.4. Tareas de clasificación

Sobre los datos mencionados en la anterior sección, los autores propusieron dos tareas de clasificación:

Texto	HS	TR	AG
Los tomas asi puro como si fuera jugo y cuando te querés rescatar estas hablando en árabe URL	0	0	0
Como son españoles nada... sin fueran refugiados...GLORIA #migrates #refugiados #EspañaLoPrimero URL	1	0	0
@OmarPrietoGob "Extranjero sin identificación será puesto en la frontera" ENVÍA AL EJERCITO A TOMAR CONTROL DE LAS PULGAS PLATANEROS Y CURVA DE AHÍ PARA QUE VEAS COMO HAY COLOMBIANOS INDOCUMENTADOS COMO MONTE AHÍ DE BUHONERS PORQUE LA POLICÍA	1	0	1
Inmigrante da una brutal paliza a una joven por no dejarse besar en Ciudad Real.#stopinvasion #YoSiTeCreo #NoesNo lo peor que no han salido a la calle las feminas del Twitter que tanto se indignaron con la salida de La Manda a la calle.	1	1	0
@elisacarrio Callate hija de puta gorda falopera	1	1	1

Tab. 4.3: Ejemplos del dataset de SemEval 2019 Task 5: *HatEval*. HS indice la presencia de discurso de odio, TR la presencia de discriminación individualizada, y AG la presencia de discriminación agresiva

- **Tarea A**: Dado un tweet predecir si contiene discurso de odio contra mujeres o inmigrantes (HS)
- **Tarea B**: Dado un tweet, predecir si contiene discurso de odio (HS), si está dirigido contra un individuo o un grupo (TR), y si es agresivo o no (AG)

La primer tarea es la versión más básica de la detección de discurso de odio, donde predecimos una etiqueta binaria que marca la presencia de contenido de esta índole. La segunda es una versión más rica, de grano fino, donde predecimos varias características de particular interés para distinguir algunas formas potencialmente más peligrosas de este fenómeno: por ejemplo, si es agresivo y si es individualizado, lo que puede indicar alguna incitación a un ataque de un individuo o miembros de algún grupo protegido.

Basile et al. [10] propusieron para medir el desempeño en la **Tarea A** la Macro F1 de las clases positiva y negativa. Para el caso de la **Tarea B**, utilizaron dos métricas: Macro F1 de las 3 clases (HS, TR, AG) y también la medida Exact Match Ratio:

$$EMR = \frac{1}{n}\sum_{i=1}^{n} I(Y_i, Y_i^*)$$

siendo Y_i las etiquetas respectivas (HS, TR, AG), Y_i^* las etiquetas que predice nuestro sistema, e I la función indicadora ($I(x, x) = 1$; 0 en cualquier otro caso). Observado más de cerca, esto puede entenderse la accuracy sobre la 3-upla de la salida de los clasificadores, pero para evitar confusiones usamos el nombre de Exact Match Ratio (EMR). [176]

4.5. Método

En esta sección describimos los distintos modelos planteados para abordar las dos tareas de clasificación de discurso de odio, así como detalles de preprocesamiento introducidos en Pérez and Luque [126] para optimizar las representaciones de modelos lineales de clasificación.

4.5.1. Preprocesamiento

Definimos dos niveles de preprocesamiento: básico y orientado a sentimientos, dependiendo del modelo a utilizar. El preprocesamiento básico de tweets es el mismo que describimos en la Sección 3.5, y es el usado con los modelos pre-entrenados o modelos neuronales.

El preprocesamiento orientado a sentimientos incluye además lematización (usando TreeTagger [148]) y manejo de negación. Para el manejo de la negación, seguimos un enfoque simple: Buscamos palabras de negación y agregamos el prefijo 'NOT _' a los siguientes tokens. Se niegan hasta tres tokens, o menos si se encuentra un token que no sea una palabra.

4.5.2. Modelos de clasificación

Para las tareas propuestas, analizamos el desempeño de diversos modelos de clasificación. Algunos de ellos son los presentados para la shared-task *HatEval* en Pérez and Luque [126], a las cuales agregamos modelos basados en transformers. [6] Para la tarea de detección binaria (**Tarea A**) planteamos 3 tipos de clasificadores:

1. Modelos lineales: regresiones logísticas y SVM con kernel lineales, consumiendo como entrada bolsas de palabras, bolsas de caracteres, y tweet embeddings

2. Redes neuronales recurrentes: usando como entrada representaciones no contextualizadas (*fastText*) y contextualizadas (ELMo)

3. Modelos pre-entrenados de lenguaje.

Para los modelos lineales, utilizamos representaciones de cada tweet calculadas con Smooth Inverse Frequency (ver Sección 2.3.1 para más detalles), usando como base los vectores de *fastText* entrenados sobre tweets con preprocesamiento orientado a sentimientos. Los modelos recurrentes consumieron como entrada la concatenación de representaciones de *fastText* que describimos en la Sección 3.6 a las cuales adosamos vectores contextualizados basados en ELMo. [129] Para esta última técnica, usamos la versión en español entrenada por Che et al. [33]. Finalmente, consideramos los siguientes modelos pre-entrenados: para el español *BETO* [31], y para el inglés *BERT* [45], *RoBERTa* [95] y *BERTweet*. [113]

La tarea de multidetección de discurso de odio (**Tarea B**) podemos pensarla de dos maneras:

[6] Estos modelos no estaban disponibles al momento de presentar dicho trabajo. El trabajo de *BERT* [45] es de finales de 2018, y hasta finales de 2019 no fue publicada una versión entrenada en español, *BETO*

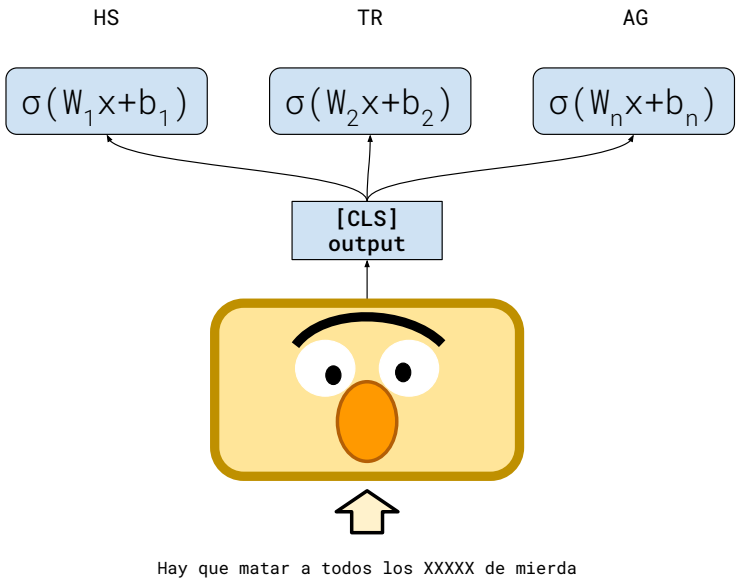


Fig. 4.3: Modelo basado en BERT para la tarea de clasificación múltiple. Cada variable (HS, TR, AG) representa un problema de clasificación en sí mismo

1. Un problema de clasificación múltiple

2. Un problema de clasificación de 5 clases

En el primer caso, el enfoque es el de predecir por separado cada una de las variables HS, AG, y TR. La segunda formulación se basa en observar que no tenemos 8 combinaciones permitidas sino sólo cinco: si no hay HS no nos interesa observar las otras dos variables. Con esta observación, convertimos cada combinación en una clase de un problema de clasificación estándar. En Pérez and Luque [126] propusimos un modelo basado en Support Vector Machines que consume la misma entrada que detallamos anteriormente y con una salida de cinco clases. No evaluamos en dicho trabajo un modelo recurrente con este esquema de clasificación ni tampoco aquí, considerando que evaluamos opciones que han demostrado tener mejor desempeño para numerosas tareas de clasificación de texto.

Asimismo, considerando la opción de multiclasificación, proponemos para la **Tarea B** modelos de lenguaje pre-entrenados con tres salidas. Recordemos que la arquitectura usual de clasificación basada en *BERT* consta de poner como última capa una softmax que consume como entrada la salida del token `[CLS]`. Esto, además, agregando como parámetro una matriz de proyección $W \in \mathbb{R}^{m \times 768}$ donde m es la cantidad de clases de nuestro problema y 768 corresponde a la dimensión de cada vector del modelo de transformers.

Para construir un modelo de multiclasificación, mantenemos la misma arquitectura pero, en lugar de usar como activación la función softmax, utilizamos la función sigmoidea elemento a elemento. En el caso de clasificación de n clases, interpretamos

a softmax$(Wx+b)_i$ como la probabilidad[7] instancia pertenezca a la clase i. Por otro lado, en el caso de multiclasificación de n variables, $\sigma(Wx+b)_i$ [8] nos da la probabilidad de predecir la etiqueta positiva para la variable i-ésima: en nuestro caso, $\sigma(Wx+b)_1$ nos da $P(HS=1 \mid x)$, $\sigma(Wx+b)_2$ nos da $P(TR=1 \mid x)$ y finalmente el tercer subíndice nos da $P(AG=1 \mid x)$. La Figura 4.3 ilustra el modelo utilizado.

Para entrenar el modelo de clasificación, evaluamos dos tipos de funciones de costo. En primer lugar, utilizamos la suma de las entropías cruzadas binarias. Concretamente, si $y = (y_{HS}, y_{TR}, y_{AG})$ son las etiquetas de una instancia e $\widehat{y}$ la predicción del modelo, la función de costo es:

$$L(y, \widehat{y}) = \sum_{k \in \{HS,TR,AG\}} J(y_k, \widehat{y_k}) \tag{4.1}$$

donde J es la entropía cruzada binaria. Esta función de costo, sin embargo, ignora cualquier tipo de jerarquía entre las variables; por ejemplo, si para una instancia tenemos $HS = 0$, calcula el costo también de las variables TR y AG. Contemplamos entonces una variante de la función descripta en la ecuación 4.1 para tener en cuenta esto:

$$L(y, \widehat{y}) = J(y_{HS}, \widehat{y_{HS}}) + \beta(y_{HS}) \sum_{k \in \{TR,AG\}} J(y_k, \widehat{y_k}) \tag{4.2}$$

donde $\beta(y_{HS})$ pondera la pérdida de las variables del segundo nivel de nuestra jerarquía. Una opción puede ser considerar $\beta(1) = 1, \beta(0) = 0$, donde ignoramos las pérdidas de las variables TR y AG cuando no hay discurso discriminatorio. Análogamente, $\beta(y) = 1$ sería el caso descripto en la ecuación 4.1. Una forma de generalizar esto es agregando un hiperparámetro $\gamma \in [0, 1]$ para escribir $\beta(y) = (1-y)\gamma + y$. Optimizamos este hiperparámetro realizando una búsqueda lineal entre 0 y 1 utilizando pasos de 0,1.

Al momento de realizar inferencia, realizamos las evaluaciones de los modelos realizamos poniendo una máscara por encima de estos modelos de clasificación múltiple de manera de evitar salidas incoherentes (por ejemplo, $HS = 0, TR = 1, AG = 0$).

4.6. Resultados

La Tabla 4.4 muestra los resultados de la evaluación para la detección de discurso de odio binaria (**Tarea A**), marcando con un asterisco aquellos modelos presentados en Pérez and Luque [126]. Respecto a los resultados en español, el clasificador basado en SVMs obtiene una buena performance, aún comparado con aquel basado en embeddings contextualizados. Este algoritmo basado en SVMs obtuvo el mejor desempeño en la competencia con 0,730 de Macro F1 [10]. El pobre desempeño de ELMo contra un algoritmo mucho más simple puede deberse a un mal pre-entrenamiento del modelo base en español[9] y también debido al cambio de dominio, a los cuales los modelos pre-entrenados previos a BERT son sumamente sensibles [73].

[7] Estrictamente hablando, más bien sería un puntaje entre 0 y 1

[8] Esta expresión es elemento a elemento

[9] No queda claro que en entrenar este modelo sobre 20M palabras sea suficiente, ni que sea un dataset suficientemente general

Modelo	Idioma	Precision	Recall	F1	Macro F1
SVM*		63,9	80,0	71,1	73,0
ELMO-RNN	es	66,1	75,3	70,4	73,5
BETO		**67,4**	**83,9**	**74,7**	**76,4**
BERT		47,4	**96,8**	63,7	49,6
RoBERTa	en	47,0	96,7	63,2	48,6
BERTweet		**49,5**	95,9	**65,3**	**54,6**

Tab. 4.4: Resultados de la evaluación para la detección de discurso de odio en el dataset de test, medidas por % de precisión, sensitividad y F1 sobre la clase positiva (discurso de odio) y por la métrica Macro F1. Con * están marcados los resultados presentados en Pérez and Luque [126]. En negrita, el mejor resultado.

Modelo		Idioma	HS F1	TR F1	AG F1	Macro F1	EMR
BETO	multi	es	74,1	**76,5**	**68,8**	**73,1**	68,5
	hier		73,5	75,8	67,4	72,2	**70,3**
	combi		**74,2**	76,3	66,8	72,4	69,8
BERT	multi	en	63,8	60,0	44,3	56,0	38,0
	hier		64,2	59,2	45,1	56,2	38,8
	combi		64,4	59,3	44,2	56,0	39,8
RoBERTa	multi	en	63,4	57,8	45,4	55,5	36,5
	hier		63,7	57,2	45,6	55,5	37,0
	combi		63,6	57,6	44,2	55,1	37,7
BERTweet	multi	en	65,8	62,9	**46,2**	**58,3**	42,6
	hier		65,6	61,7	45,0	57,4	42,3
	combi		**66,6**	**63,7**	44,4	58,2	**44,9**

Tab. 4.5: Resultados de la evaluación para para **Tarea B** en términos de las F1 de las clases HS (Hate Speech), TR (Targeted), AG (Aggressive), el Exact Match Ratio (EMR), las Macro F1 de las clases en cuestión, y la Macro F1 de la clase HS. Las 3 variaciones de los modelos son: *multi* es la salida de multiclasificación estándar, *hier* es la salida de multiclasificación con una jerarquía de clasificación, y *combi* es la salida de multiclasificación con una combinación de clasificaciones. Los resultados están expresados como las medias de 10 corridas independientes.

Para ambos idiomas, los modelos basados en Transformers [160] obtienen la mejor performance, con considerables mejoras respecto a los modelos basados en ELMo y a los SVMs [10]. Particularmente, en el caso del inglés, *BERTweet* [113] obtiene la mejor Macro F1, algo esperable considerando que está particularmente diseñado para Twitter.

La Tabla 4.5 muestra los resultados de la **Tarea B**, reportado por las F1 de cada variable predicha (HS, TR, AG), así como por la Macro F1 de las 3 variables mencionadas y el Exact Match Ratio. Los resultados están expresados como la media de 10 corridas independientes del experimento para cada configuración distinta. Consideramos las 3 versiones: *multi* refiere a clasificación múltiple, *hier* a clasificación

[10] Un modelo que no evaluamos en el presente trabajo es la versión en español de RoBERTa, recientemente entrenada. En el Capítulo 7 evaluaremos su rendimiento en esta tarea

múltiple con la función de costo jerárquica, y *combi* a la conversión del problema en una clasificación de cinco clases.

Podemos observar que para español, la mejor performance en términos de EMR (la métrica más estricta) es el clasificador entrenado con la función de costo definida en 4.2 (con el hiperparámetro $\gamma = 0{,}1$); sin embargo, la diferencia entre las performances no es significativa al correr un test de Kruskal-Wallis ($H(9) = 3{,}492, p = 0{,}174$). En términos de Macro F1, la mejor performance es de *BETO* con la salida múltiple y sin la función de costo jerárquica (*multi*) pero de nuevo esta diferencia no es significativa ($H(9) = 3{,}656, p = 0{,}16$).

Respecto al inglés, los mejores resultados pueden observarse en el modelo entrenado con *BERTweet* con la salida de cinco clases en el caso del EMR, y con la salida múltiple (sin pérdida jerárquica) para la Macro-F1. Este resultado, sin embargo, queda en términos de EMR por debajo del baseline propuesto por los autores del dataset [10], aunque cercano en términos de Macro F1 a los mejores resultados de la competencia. En Gertner et al. [59], se basaron en un ensemble de modelos entrenados con BERT y usando también un ajuste de dominio sobre tweets. Esta baja performance de nuestros modelos (y de los modelos en general sobre ese dataset) puede deberse a problemas de anotación y a que las particiones de train y test no son idénticamente distribuidas. [11]

La Tabla 4.6 muestra la comparativa para la detección de discurso de odio (HS) para aquellos clasificadores que obtuvieron mejores resultados para **Tarea A** y **Tarea B** (*BETO* y *BERTweet*). Consideramos para la **Tarea B** al clasificador *multi* de cada modelo de lenguaje. Lejos de dañarse la performance de la detección de lenguaje discriminatorio (lo que analizamos en la **Tarea A**), predecir más de una variable pareciera mantener el desempeño general; más aún, podemos observar que en términos de Macro F1, incluso parecieran tener una ligera mejora al ser entrenados sobre una tarea más compleja.

4.6.1. Análisis de Error

Para tener una mejor idea de lo ocurrido con nuestros clasificadores, realizamos un análisis de error sobre los datasets en español. Tomamos las salidas de diez clasificadores *BETO* entrenados cada uno con distinta semilla y analizamos el error tomando el ensamble por voto mayoritario para disminuir los efectos de la varianza

[11] En Gertner et al. [59] dan evidencia de esto, algo que perjudica el desempeño de estos modelos

Modelo	Idioma	Tarea	Precision	Recall	F1	Macro F1
BERTweet	en	A	$49{,}5 \pm 1{,}2$	$95{,}9 \pm 1{,}2$	$65{,}3 \pm 0{,}9$	$54{,}6 \pm 2{,}7$
		B	$50{,}5 \pm 1{,}1$	$94{,}8 \pm 1{,}8$	$65{,}8 \pm 0{,}5$	$56{,}7 \pm 2{,}2$
BETO	es	A	$67{,}4 \pm 2{,}1$	$83{,}9 \pm 2{,}6$	$74{,}7 \pm 0{,}7$	$76{,}4 \pm 1{,}1$
		B	$71{,}3 \pm 4{,}2$	$77{,}8 \pm 5{,}4$	$74{,}1 \pm 1{,}3$	$77{,}1 \pm 1{,}5$

Tab. 4.6: Comparación de la performance sobre la detección de discurso de odio para los clasificadores entrenados sobre **Tarea A** y **Tarea B**. Resultados expresados como la media de 10 corridas independientes del experimento junto a sus desviaciones estándar. Ambos clasificadores de la **Tarea B** están entrenados sobre el problema de multi-clasificación

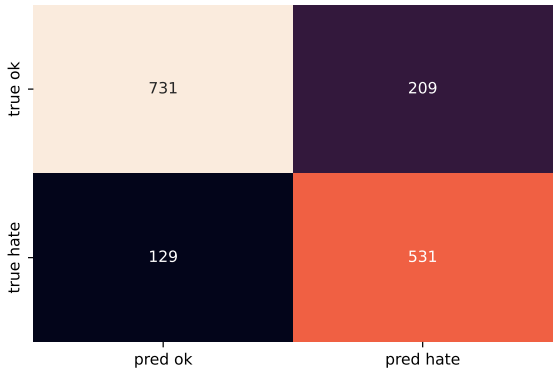


Fig. 4.4: Matriz de confusión para la detección de discurso de odio sobre un ensemble de voto mayoritario de 10 clasificadores entrenados para la **Tarea B**

de los modelos. De esta manera, tratamos de buscar aquellos errores frecuentes, aquellos que la mayoría de los clasificadores erran. La Figura 4.4 muestra la matriz de confusión de nuestros clasificadores sobre el dataset de test. Podemos observar que la mayor fuente de error son falsos positivos, así que centraremos nuestro análisis en estas instancias.

En la Tabla 4.7 podemos observar una selección de los falsos positivos que cometen los clasificadores. Por un lado, algunos errores se deben a overfitting a ciertas palabras "clave" (como las nro. 1, 2 y 3) muchas de las cuales son producto del proceso de recolección que está fuertemente basada en keywords (inmigrante, sudaca, por ejemplo). Probando los clasificadores sobre entradas manualmente editadas, vemos que algunas de esas palabras automáticamente disparan una salida positiva de discurso de odio[12]. Otros de los casos incorrectamente marcados como discurso de odio son los reportes de discriminación (como el caso 5, parcialmente también el 1), casos de apropiación cultural (caso 3 o 5) o bien contenido irónico (casos 3, 4, 7 y 8). Estos ejemplos dan cuenta de la dificultad de la detección de este fenómeno.

Un problema que hemos observado es que hay –a nuestra consideración– una fuerte cantidad de errores de etiquetado en el dataset utilizado. La Tabla 4.8 muestra algunos de estos ejemplos, todos estos etiquetados incorrectamente como discurso de odio. Si bien es difícil establecer una causa específica para cada uno de estos errores, es posible que sean causados por una combinación de:

- Confusión entre lenguaje ofensivo y discurso de odio (ejemplos 1, 2, 3, 4 y quizás 5)
- Desconocimiento cultural (ejemplo 7, quizás 10 también)
- Opinión pero no necesariamente discurso de odio (ejemplo 9)
- Tweets truncados (el ejemplo 6 fue mal recolectado y está truncado)

[12] Puede probarse esto en el cuadro interactivo de `https://huggingface.co/pysentimiento/robertuito-hate-speech?text=migrante`. Si bien este modelo es el que presentaremos en el capítulo 7, el problema se mantiene

Índice	Tweet
1	Una cosa que le pasó a una amiga mexicana en Barcelona. Una chica repartía flyers en castellano y catalán, le dio uno a mí amiga en catalán y al segundo se lo quitó y se lo dio en castellano. -No, déjalo que entiendo catalán -Ah, es que como te he visto cara de sudaca Racismo 0 URL
2	Los fascistas salen de la ratonera para atacar a los 100 inmigrantes que han logrado saltar la valla de Ceuta. Asco de prensa y partidos políticos que se pelean por lograr los votos de los fascistas. Ningún ser humano es ilegal.
3	@saludhernandezm Esas personas que extrañan dictaduras, cualquier dictadura, aún más si son las ibéricas que ponen como eje a Madrid; su cultura, política e idioma sobre la idiosincrasia de los demás. Esos mismos que añoran a Franco y viven en un país 'Sudaca' creyéndose dueños de la verdad.
4	@caciena después sigo siendo un sudaca pero bueno
5	@ArturoMonteduro Pues tienes toda la razón del mundo. Y lo peor es que seguro que tenía papeles y DNI español, pero queda mejor decir .ªrgelino.º üno d'estos del top manta", "puto inmigrante", o "moro mierdaz ya pues matas dos pájaros de un tiro.
6	Obvio me ofendo cuando algn Sudaca hace algn comentario presuponiendo que los mexicanos somos feos, o que el pas est de la verga. Entre mexicanos podemos hablar mierda de Mxico, pero que a ningn pinche extranjero se le ocurra, porque va a haber pedo!
7	TODOS LOS INMIGRANTES Y GITANOS FUERA!!! Menos: el colombiano que me vende coca, el negro que me consigue putas, el moro que me pasa costo y el gitano que me vende maría.
8	Ayer nos fuimos a tomar algo con los cumpas: Dos españoles, un ponja, un africano y un sudaca. Estamos para campaña de United Colours of Benneton.

Tab. 4.7: Falsos positivos del modelo de clasificación para la detección de discurso de odio (HS).

Dentro de estas asignaciones posiblemente erróneas, en el análisis de error separamos un subconjunto especial sobre el cual entendemos que falta contexto para asignar una etiqueta. Si bien este contexto a veces puede ser inferido por un humano (por ejemplo, entendiendo que un comentario habla de tal o cual suceso), un algoritmo de detección podría beneficiarse de esta información e incorporarla como entrada de las diferentes instancias a clasificar. Para el dataset en concreto, muchos de estos comentarios tienen un tópico implícito: hablan de incidentes en la frontera de Ceuta (España) entre inmigrantes y la policía, esto debido al proceso de recolección realizado. La Tabla 4.9 muestra algunos de estos ejemplos que parecieran carecer de información adicional para ser comprendidos. En algunos, podemos observar que falta contexto conversacional: aquellos donde tenemos un @username al principio del tweet indican que son una respuesta. En otros casos, falta información de un suceso o noticia sobre la que se está comentando, como aquellos que terminan con URL que suelen ser citas de otros tweets o links a noticias.

4.7. Discusión

Respecto a la performance de los modelos presentados, los modelos basados en Transformers son notoriamente superiores a los demás modelos, en ambas tareas e

Índice	Tweet
1	Callate la puta boca pesada @Lauu_tb
2	me cansada tía cállate ya la puta boca lo que te queda de vid
3	FLOR SALTANDO AHÍ LA RE PUTA MADRE TE LO MERECES PEDAZO DE FORRA TE LO MERECES, EN LA CARA DE TODOS HIJA ASÍ SE HACE
4	Callate! cerra el orto un ratooo! es mas, no te queres morir mejor? la puta que te pario
5	Esta mujer es lo más la puta madre URL
6	@csdb530 @carvasar Claro, porque la culpa la tienen las niñas embarazadas y las víctimas de acoso sexual, violación... URL
7	#Pendejos Don't call me gringo You fuckin beaner Stay on your side Of that goddamn river Don't call me gringo You beaner No me digas beaner Mr. puñetero Te sacaré un susto Por racista y culero No me llames frijolero Pinche gringo puñetero
8	Mónica que te calles la puta boca #OTGala7
9	@pablocasado_ @imbrodamelilla @TeoGarciaEgea pablo no quieren concertinas , un muro de 12 metros o mas de altura de dos plantas de pisos ,se acabo los problemas @guardiacivil @policia @EMADmde @MonarquiaEspana nadie se quejaría de nada
10	#OTGala7 Noemí JAJAJAJAJAJA ESTA MUJER ES LA PUTA AMA

Tab. 4.8: Ejemplos mal etiquetados como discurso de odio. En etiqueta marcamos cómo están etiquetados (erróneamente). El índice es meramente para referencia.

idiomas. Particularmente, en inglés podemos observar que aquellos pre-entrenados sobre tweets como *BERTweet* tienen mejor performance que aquellos que son entrenados sobre Wikipedia como *BERT* o *RoBERTa*. El discurso de odio está muchas veces basado en la utilización de jergas e insultos raciales o misóginos, con lo cual es esperable el mejor desempeño de un modelo que tiene en sus datos de entrenamiento este tipo de expresiones.

Sobre la tarea más difícil de detección múltiple de discurso de odio (**Tarea B**), propusimos varios enfoques: uno basado en predecir cada variable por separado y otro en predecir una variable que indique la combinación en cuestión. El modelo de predicción múltiple entrenado con la función de costo jerárquica obtuvo la mejor performance en términos de EMR, y la de multi-clasificación obtuvo la mejor en términos de Macro F1. En el caso de inglés, el modelo entrenado sobre cinco clases obtuvo la mejor performance en EMR y de nuevo el de multi-clasificación sobre Macro F1; sin embargo, esta queda por debajo de la mejor performance de la competencia (obtenida por el equipo MITRE [59]) que usa una compleja combinación de técnicas, algunas de las cuales veremos en el Capítulo 7. De estos dos casos, el modelo de multi-clasificación corre con la ventaja de calcular cada variable de manera independiente y tener un hiperparámetro menos.

Etiqueta	Tweet
No HS	- @ArturoMonteduro Pues tienes toda la razón del mundo. Y lo peor es que seguro que tenía papeles y DNI español, pero queda mejor decir “argelino” o “uno d’estos del top manta”, “puto inmigrante”, o “moro mierda” y ya pues matas dos pájaros de un tiro.
	@UbedaJoseA @WalkingDead_AMC Perro cállate ala puta
	@iancandado Es sudaca como.tu
	@Stopfalsasdenun Pero que tienes en la cabesaaaa?(con voz de sudaca de mierda)
	@_Lassana_ Me la has puesto como un tanque apuntando a un campanario negrata
	@odiomistweets Es una analogía pendejo sudaca
	@ElQuemaMarias te parece que un sudaca tatuado excelentemente bien dotado y carismático no es buen ejemplo?
	@HQGatica @AnglicaVillarN1 Este sueco con cara de bolita .xq eso es este aspiracional desclasado . x dignidad debería irse . no lo hará xq este sudaca no la tiene !!!
	@SleeplessXenn @envdiaN Este finde te recojo y vamos al barrio negrata de envdian
	@elmundoes El q quería quitar concertinas y luego quitarse a los inmigrantes !!
	@AlexTor1975 @sanchezcastejon Disfràcese de moro o negro, y al momento tiene resuelto el problema de espera, de verdad, mano de santo!. Nada como ser invasor, refugiado, o inmigrante en este paraíso terrenal. Ahora, si eres nativo del paraíso!!!!, Se te puede complicar la existencia!!!!.
HS	CALLATE YA Y HAZME LA PUTA CENA JODERRR URL
	Basta ya de buenismo, si te no respetas a la polica ni a la guardia civil en la frontera nada de concertinas, simplemente 4 tiros. Fin del problema
	@benemeritosGC La solución es fácil, muro de hormigón de 3 metros de alto, garitas cerradas para vigilancia y por delante la valla actual con concertinas. No es caro, es cuestión de quererlo hacer.
	Por desgracia, no queda otra, aportan poco y nos cuestan mucho. Incluido nuestra seguridad. #Inmigración #Immigration URL

Tab. 4.9: Algunos ejemplos observados en el dataset que son difíciles de entender sin contexto conversacional. Etiqueta es la asignada en el dataset

Algo que merece cierta atención es que, lejos de empeorar el desempeño de nuestros modelos, agregar nuevas variables a predecir (además de la existencia de discurso de odio) pareciera mejorar levemente la performance de la detección de este fenómeno, a la vez que obteniendo salidas más ricas e interpretables. Más aún, observamos que otros trabajos [59] utilizando el mismo conjunto de datos mejoraron la performance con una capa adaptadora que modela las dos variables latentes codificadas conjuntamente en la salida binaria: la misoginia y el racismo. Teniendo esto en cuenta, una pregunta a explorar es si contar con esta información (las características agredidas) puede mejorar la performance de los clasificadores o tener salidas más interpretables que sólo una etiqueta binaria.

Hacemos a continuación una disquisición no sólo sobre este trabajo y el dataset en el que se basa sino en líneas generales sobre los recursos y enfoques actuales en el área de detección de discurso de odio. Continuando con la idea del párrafo anterior, una limitación que puede verse es que la mayoría de los trabajos atacan una, dos, o a como mucho tres características protegidas. Por ejemplo, los trabajos de Waseem and Hovy [165] y Basile et al. [10] sólo consideran racismo y sexismo, mientras que el de Davidson et al. [42] agrega homofobia a esta consideración. Sería deseable poder contar con un dataset que como mínimo cuente con estas tres características en conjunto a otras quizás menos utilizadas: odio de clase (a veces conocida como *aporofobia*), discriminación por aspecto físico, por discapacidad, entre otras. A su vez, contar con la información de la característica atacada (algo que no ocurre en los datos utilizados en este capítulo) puede ser interesante para tener una mejor interpretabilidad de las salidas de nuestros algoritmos, y posiblemente para mejorar su rendimiento.

Un problema particular que se puede observar en los datos de Basile et al. [10] (pero que atraviesa a muchos otros) es el proceso de recolección de los datos: los tweets son recolectados mayormente a través de keywords. Como mencionamos en la Sección 4.3, el proceso de recolección consta de varias estrategias combinadas; sin embargo (ver Apéndice A), hay una altísima incidencia de algunas palabras (como *sudaca* o *inmigrante*) que sesgan fuertemente el dataset. Esto (entre otras cuestiones) puede ser un problema para los modelos que se entrenan sobre estos datos, haciendo que aprendan correlaciones espurias generadas por estas distribuciones. De todas formas, esto es una limitación general para estos tipos de aprendizaje sobre datos crudos y etiquetas, donde es difícil establecer e interpretar cómo un clasificador termina encontrando patrones para detectar el fenómeno medido.

La **anotación**, la etapa subsiguiente a la recolección de datos, pareciera presentar en este dataset algunos problemas. Hemos visto en la anterior sección una lista no extensiva de varios errores de etiquetado, aún cuando este dataset fue realizado con un complejo sistema combinando crowdsourcing, etiquetado con expertos y desempate. Si bien es difícil trazar las razones detrás de estos problemas, observando las instancias incorrectamente etiquetadas puede hipotetizarse que esto es producto de un no entendimiento de las expresiones en los distintos dialectos del español y diferentes realidades socioculturales. Waseem [164] mostró que las anotaciones "amateurs" (producto del uso de crowdsourcing) tienden a tener mayores instancias de Hate Speech (algo que daría la impresión de ocurrir aquí) y que datasets anotados por expertos mejoran la performance de los modelos. Este problema podría profun-

dizarse dado que no queda claro si los anotadores son hablantes nativos de español, al no tener información detallada de quienes realizaron la tarea de etiquetado.

Un problema del dataset estudiado en este capítulo (pero que aplica a muchos otros también) es la **falta de contexto**: los mensajes carecen de información adicional sobre la noticia o el tema del que se está hablando. Cuando leemos un mensaje de un tweet, casi siempre lo leemos en el contexto de una noticia, o un trending topic. Muy rara vez leemos un mensaje en total aislamiento. De hecho, gran parte de los comentarios de este dataset tiene un contexto implícito: la noticia de conflicto migratorio en Ceuta. Otros comentarios, por otro lado, no se entienden bien ya que son respuestas a un tweet y que según el hilo de conversación pueden entenderse o no como discriminatorios.

Sobre esta falta de contexto, hay muchos mensajes aislados que pueden requerir información adicional para entender su significado. Por ejemplo, un comentario que dice "hay que matarlos" puede o no entenderse como discurso de odio. Si el objeto del mensaje se refiere a mosquitos, ese mensaje no es odioso; si, por otro lado, está hablando sobre migrantes chinos en el contexto del COVID-19, entonces ese mensaje es discriminatorio (y además llama a tomar una medida violenta). Podemos preguntarnos sobre este punto si el acceso a información contextual nos puede auxiliar en la detección de discurso de odio, siendo este contexto un hilo de conversación, una noticia a la que se refiere, o alguna otra forma de de información adicional.

Finalmente, otro problema que suele ocurrir relacionado al anterior es que no tenemos **información granular** de los datos anotados. Si bien algunos trabajos agregan información de la característica vulnerada, la mayoría simplemente agrega una etiqueta binaria sobre la existencia o no de discurso de odio (o bien algún nivel intermedio como si hay o no discurso ofensivo, como el caso de Davidson et al. [42]). Teniendo en cuenta lo observado en este capítulo, agregar información más detallada sobre cada caso puede ayudar a mejorar la detección del discurso de odio mediante una señal más rica a nuestros clasificadores sobre las diferentes fronteras de cada característica ofendida.

4.8. Conclusiones

En este capítulo hemos hecho un primer acercamiento a la tarea de detección de lenguaje discriminatorio, repasando de su definición desde un marco legal y desde el usado en la literatura de procesamiento de lenguaje natural. Analizamos técnicas de clasificación del estado del arte sobre el dataset presentado en la shared task multilingual [10]. En base a este dataset, analizamos dos tareas: detección binaria de discurso de odio, y detección de múltiples variables (si es discurso de odio, si es dirigido, si es agresivo).

Para estas tareas, presentamos técnicas de clasificación basadas en modelos lineales que consumen distintos tipos de entrada como ser tweet embeddings y bolsas de caracteres; modelos basados en redes recurrentes que consumen embeddings contextualizados; y finalmente, utilizamos modelos de lenguaje pre-entrenados usando la arquitectura de Transformers. Para ambas, los modelos de Transformers obtuvieron el mejor desempeño, superando ampliamente a las demás técnicas.

En el caso de la tarea de detección múltiple, propusimos dos formas de ata-

car el problema: como clasificación múltiple (prediciendo simultáneamente las tres variables), y convirtiendo a un problema de clasificación simple sobre cinco clases posibles. Observamos, a su vez, que lejos de dañar la performance de la detección de discurso de odio, predecir más de una variable mejora la performance de nuestros clasificadores.

Analizando este dataset y algunos otros de la bibliografía, marcamos algunas oportunidades de mejora y observaciones en la detección de discurso de odio. En primer lugar, la posibilidad de agregar información contextual a los mensajes a analizar, sea sobre el tópico del que se está hablando o contexto conversacional previo. En segundo lugar, agregar **información granular** sobre las características ofendidas. Y finalmente, un punto no menor a la hora de la creación de recursos para un fenómeno tan complejo y social es indispensable tener muchos recaudos a la hora de la anotación –algo que ya ha sido observado en otros trabajos– teniendo particular cuidado sobre el trasfondo sociocultural de quienes tomen esa tarea.

En los siguientes capítulos, exploraremos algunas de estas oportunidades de mejora. Particularmente, nos centraremos en la incorporación de contexto en la detección de discurso discriminatorio, construyendo un conjunto de datos que incorpore esta información a los mensajes anotados, y explorando cómo mejorar los algoritmos del estado del arte que aprovechen esa información.

Parte III

DETECCIÓN CONTEXTUALIZADA DE DISCURSO DE ODIO

5. CONSTRUCCIÓN DE UN DATASET DE DISCURSO DE ODIO CONTEXTUALIZADO

Por lo marcado en la discusión de la anterior sección, consideramos interesante el problema de analizar el impacto del contexto en la detección de lenguaje discriminatorio. Antes de proseguir, podemos preguntarnos: ¿a qué nos referimos con el término "contexto"? La contextualización, según John Cook-Gumperz, es:

> (el) uso que hacen hablantes y oyentes de señales verbales y no verbales para poder conectar lo que se dice en un momento con el conocimiento adquirido a través de la experiencia para poder mantener la participación en la conversación y entender lo que se pretende decir. (Gumperz [66])

En este sentido, cualquier señal que pueda ayudar a entender las intenciones del interlocutor en una red social es información que ayuda a situar los mensajes: desde el hilo de una conversación, la noticia a la que hace referencia, el historial de conversaciones previas entre los interactores, información sociocultural de los interlocutores, entre otras [153]. Para poner un ejemplo de por qué es necesario disponer de información adicional al comentario analizado, el mensaje "sos un hombre" en solitario puede parecer inofensivo; ahora, si ese mismo mensaje está dirigido hacia una mujer trans, su sentido es claramente discriminatorio. El comentario –con claro tono agresivo– "hay que tirar una bomba ahí" puede tener carácter discriminatorio si lo consideramos en el contexto de una nota que habla sobre China y el COVID-19; sin embargo, es distinto si estamos hablando de un partido de fútbol, donde el remitente de un club manifiesta su enemistad contra otro equipo.

Vimos en el anterior capítulo que muchos mensajes analizados no se entendían bien al carecer de información contextual, tanto conversacional o del tópico al que hace cuestión. En líneas generales, la mayoría de los problemas de NLP sobre textos sociales suelen plantearse sobre comentarios sin ningún otro tipo de dato de quien lo emite, a quien se lo dirige, ni sobre qué tema está hablando. Para analizar esto desde el problema de la detección de discurso de odio, nos abocamos en primer lugar a la tarea de crear un conjunto de datos que no sólo contenga un mensaje/comentario, sino que provea un contexto para éste. Un ámbito natural para esta tarea son las notas periodísticas, donde disponemos de un artículo y comentarios realizados sobre la nota. En este escenario, el comentario es el texto a analizar, mientras que el contexto está dado por la nota.

Muchos sitios de noticias disponen de sistemas embebidos de comentarios, pero vista la dificultad para la recolección y los limitados datos provistos por estos sitios acerca de sus usuarios nos llevaron a buscar otro medio: Twitter. Esta red social provee una sencilla API para descargar datos, a la vez que tiene términos y condiciones amigables para poder publicarlos. Así mismo, podemos pensar que algunas secciones de Twitter operan de una manera similar a un foro de comentarios de un sitio de noticias. Este dominio (comentarios sobre artículos periodísticos) tiene una naturaleza particular ya que las agresiones discriminatorias son usualmente a

personajes públicos o colectivos de personas, y se dan de manera indirecta (a través del comentario en la noticia) y no directa (es decir, como respuesta al usuario de Twitter ofendido).

El trabajo realizado en este capítulo tuvo lugar en el contexto de un Proyecto Interdisciplinario de la UBA[1] junto a sociólogos, abogados, lingüistas, y computólogos. Particularmente, el trabajo de la construcción del manual de etiquetado fue discutido en conjunto, contemplando varias perspectivas a la hora de armar una definición propia (algunas de estas ya fueron vertidas en la discusión en la Sección 4.1). Teniendo en cuenta que muchos trabajos del área de detección de discurso de odio mediante técnicas de NLP no se realizan desde una mirada interdisciplinaria, es un aspecto a remarcar de la construcción de este recurso.

5.1. Trabajo previo

Pocos trabajos del área de detección de lenguaje abusivo o discurso de odio incorporan algún tipo de contexto a los comentarios recolectados para estas tareas. En esta sección haremos una revisión de los trabajos que han abordado la construcción de recursos que contengan algún tipo de información contextual. Gao and Huang [56] construyeron un conjunto de datos de lenguaje discriminatorio sobre 1518 comentarios del sitio de Fox News, siendo estos anotados por etiquetadores que observaron tanto el comentario como el titular de la noticia en conjunto. Sobre estos datos, los autores efectuaron experimentos de clasificación usando regresiones logísticas y redes neuronales. En estos experimentos, observaron que un clasificador (tanto lineal como neuronal) mejora su performance al consumir el título de la noticia, dando indicios de que se puede aprovechar el contexto para mejorar la detección de este fenómeno. Sin embargo, como marcan Pavlopoulos et al. [121], este trabajo cuenta con algunos problemas: en primer lugar, el tamaño del dataset es pequeño, y está extraído de sólo 10 noticias, lo cual limita fuertemente los posibles contextos de los comentarios. A su vez, la anotación fue realizada mayormente por una única persona, lo cual hace poco confiables las etiquetas obtenidas. Finalmente, algunos detalles menores debieran ser analizados con mayor detalle, como por ejemplo la utilización de los nombres de usuarios como variables predictivas.

Mubarak et al. [112] recolectaron comentarios en árabe con contenido abusivo del portal Al Jazeera para distintos artículos periodísticos. Sin embargo, este dataset tiene un problema: los comentarios son presentados a los anotadores sobre noticias, ignorando todo el thread de la conversación. Esto hace que el contexto sea presentado de manera parcial.

Paralelamente a nuestro trabajo, Pavlopoulos et al. [121] analizaron el impacto de agregar contexto a la tarea de detección de toxicidad. En particular, plantearon dos preguntas:

- ¿Qué tanto afecta el contexto a la toxicidad percibida por humanos en conversaciones online?

[1] `https://cyt.rec.uba.ar/vinculacion-transferencia/piuba/`

- ¿Puede el contexto ayudar a mejorar la performance de clasificadores de toxicidad en comentarios?

Para responder estos dos puntos, los autores construyeron dos datasets en base a Wikipedia Talk Pages [78], un conjunto de datos de discusiones del sitio de Wikipedia. En primer lugar, armaron un pequeño conjunto de 250 comentarios anotados por dos grupos disjuntos de anotadores: uno de los grupos anotó los comentarios de manera contextualizada, viendo tanto el comentario en cuestión como el título de la discusión; el otro grupo sólo vio el comentario a anotar sin contexto alguno. En dicho experimento observaron que los anotadores que observaron el contexto percibieron 6.4 % de comentarios tóxicos versus un 4.4 % de quienes anotaron sin contexto, una diferencia significativa aplicando un test Mann-Whitney U. Desagregando estos resultados, observaron que 13 de los 250 comentarios (5.2 %) tuvieron diferencias de anotación entre los dos grupos, con 9 (3.6 %) comentarios donde aumentó la toxicidad percibida y 4 comentarios donde bajó la toxicidad al ser agregado el contexto.

Para responder la segunda pregunta, anotaron 20 mil comentarios del mencionado foro, la mitad anotados por un grupo que etiquetó viendo el contexto y la otra que no lo vio. Entre todos los comentarios recolectados, eligieron aquellos con profundidad entre dos (respuestas directas) a cinco, y que fuesen entre 10 y 400 caracteres de largo. Luego, entrenaron varios clasificadores con técnicas del estado del arte, sobre los cuales pudieron observar que el contexto no pareciera mejorar significativamente la performance en la detección de toxicidad en comentarios. En el próximo capítulo nos extenderemos sobre las técnicas utilizadas por este trabajo.

Xenos et al. [170] continuaron el trabajo de Pavlopoulos et al. [121] desagregando el resultado de la segunda pregunta. Puntualmente, y observando que sólo un porcentaje pequeño de los comentarios parecen ser incididos por el contexto en el trabajo anterior, construyeron una nueva tarea: estimación de sensibilidad al contexto. Para ello, y usando como base el conjunto de datos de Civil Comments [25], reanotan un subconjunto de sus comentarios usando información de contexto a través de crowdsourcing, y usando etiquetas de toxicidad en un estilo similar a una regresión ordinal: no tóxico, incierto, tóxico, y muy tóxico. Sobre las anotaciones originales (que fueron hechas sin contexto) y las nuevas anotaciones, definieron para cada comentario una sensibilidad al contexto, dada por:

$$\delta(p) = s^{oc}(p) - s^{ic}(p) \tag{5.1}$$

donde s^{oc} es la fracción de anotadores sin contexto que marcaron toxicidad, y s^{ic} los que no tienen contexto. En el siguiente capítulo haremos un repaso de los experimentos de clasificación obtenidos en este trabajo.

Sheth et al. [153], en un trabajo muy reciente, señalaron algunas oportunidades y desafíos para incorporar fuentes de información más ricas a la tarea de detección de toxicidad. Por ejemplo, incorporar información como el background socio-cultural de los interactores puede ayudar a distinguir algunos tipos de reapropiación de términos potencialmente catalogados como tóxicos – por ejemplo, personas afroamericanas interpelándose con términos racistas entre sí. Así mismo, el historial de interacción entre los usuarios puede ayudar a distinguir interacciones abusivas de charlas amistosas

Fig. 5.1: Boceto del conjunto de datos: artículos periodísticos y sus respectivos comentarios en Twitter

entre amigos que usan vocabulario potencialmente tóxico. Finalmente, se promueve el uso de contenido externo para acercarse lo más posible al conocimiento humano a través de conocimiento del contenido, el individuo (atacado) y la comunidad. Para ello, se promueve el uso de bases de conocimiento y knowledge-infusion learning [58] para combinar cómputo neuronal sobre datos no estructurados y estructurados.

Wiegand et al. [166] mencionan formas implícitas de abuso, mucho más complejas que las basadas solamente en palabras ofensivas. Por ejemplo, deshumanizaciones ("los judíos son una plaga que merece ser eliminada"), llamadas a la acción ("hay que tirar una bomba en ese país"), acusaciones ("los chinos inventaron el coronavirus"), entre otros tipos sutiles de comportamiento tóxico. También menciona que la mayoría de los datasets no consiguen capturar estos fenómenos debido a la forma de recolección usualmente basada en keywords.

Sap et al. [147] plantean un esquema bastante más complejo dentro de la detección de toxicidad o lenguaje abusivo. El conjunto de datos presentado en ese trabajo consta de comentarios recolectados de diversas redes sociales que son analizados con un formalismo al que denominan *Social Bias Frames*. Cada instancia está etiquetada jerárquicamente de acuerdo a: toxicidad, intencionalidad, obscenidad, si está dirigido a un grupo, a qué grupo, qué implicancia tiene (*"los XXX son todos YYY"*), y si el emisor es perteneciente al mismo grupo social que está siendo en teoría atacado.

5.2. Esquema del conjunto de datos

Para construir un conjunto de datos contextualizado analizamos algunas alternativas. Como vimos en otros trabajos, se puede entender el contexto de un mensaje de varias maneras: un contexto temático, donde sabemos que cierto comentario ha-

bla sobre un tema en particular; y un contexto conversacional, donde tenemos una secuencia de comentarios (un hilo o thread) y podemos extraer un comentario padre para cada uno salvo el raíz. La primera opción es la explorada por Gao and Huang [56], Mubarak et al. [112], donde recolectan comentarios de Fox News y Al-Jazeera respectivamente. El contexto conversacional, como hemos relatado anteriormente, es explorado en Pavlopoulos et al. [121] y Xenos et al. [170]; sin embargo, como es marcado en el primer trabajo, la recolección de datos es no trivial, aún en un caso más amplio como el lenguaje abusivo, ya que la incidencia de comentarios de esta índole es relativamente baja. Es esperable que la tasa de ocurrencia de contenido discriminatorio sea aún menor, dificultando la recolección de datos interesantes para nuestro estudio.

Para analizar el impacto del contexto en la tarea de detección de discurso de odio, decidimos entonces ir por la primera opción: comentarios sobre notas periodísticas. No vamos a considerar un hilo de respuestas (contexto conversacional) sino simplemente aquellos comentarios que sean directos sobre el artículo periodístico. En ese punto, la idea sería similar a la de Gao and Huang [56], aunque una diferencia respecto a este conjunto de datos es la de incorporar dos modos de contexto: uno corto, donde sólo tengamos el título de la noticia; y uno largo, donde tengamos el texto completo del artículo.

Algo no menor a la hora de considerar la construcción del dataset es la posibilidad de publicar los datos. Por citar un ejemplo, el conjunto de datos recolectado por Gao and Huang [56] es de libre acceso [2] pero no queda claro que los términos y condiciones de la fuente permita esto. Más aún, si hubiésemos querido extraerlo de múltiples fuentes (por ejemplo, varios diarios), deberíamos chequear y/o acceder a permisos para cada sitio, a la vez que tendríamos el problema de tener fuentes diversas de los datos: diferentes longitudes, formatos, metadatos, entre otros.

Para evitar muchos de estos inconvenientes y poder reutilizar parte del trabajo desarrollado en esta tesis, decidimos recolectar comentarios en Twitter. Concretamente, decidimos recolectar respuestas de usuarios a posteos hechos por cuentas de medios. De alguna manera, esto emula un foro de comentarios de medios, teniendo la ventaja de un formato único para comentarios y un acceso a una audiencia de usuarios mucho más amplia que la de los microforos de cada sitio de noticias. La Figura 5.1 ilustra un boceto de lo que queremos recolectar. A su vez, una ventaja de Twitter es que posee términos y condiciones de Twitter amigables para publicar los datos con fines de investigación. Las notas periodísticas fueron también descargadas pero debido a restricciones no tenemos aún en claro si podrán ser publicadas.

Finalmente, la elección del idioma. El conjunto de datos construido consta de comentarios realizados en idioma español, más precisamente en la variedad dialectal del Río de la Plata (español rioplatense). Una primera consideración al respecto de esto es la de generar recursos por fuera del inglés, un eje planteado para esta tesis. Por otro lado, también es importante señalar que el discurso de odio es un fenómeno cultural, y es importante que quienes estén a cargo de la construcción de este recurso sean conscientes del trasfondo sociolingüístico donde están situados los discursos discriminatorios. Es por eso que a lo largo de este capítulo tuvimos

[2] `https://github.com/sjtuprog/fox-news-comments`

Nombre	username	#Followers
La Nación	@LANACION	3,6M
Clarín	@clarincom	3,2M
Infobae	@infobae	3,0M
Perfil	@perfilcom	0,8M
Crónica	@cronica	0,8M

Tab. 5.1: Cuentas de medios utilizadas para la recolección de datos, junto a sus nombres de usuarios y la cantidad de seguidores en Twitter (al momento de la recolección)

particular cuidado en esta dimensión, tanto desde el proceso de recolección hasta la selección de los anotadores que estén inmersos en la realidad cultural local.

5.3. Proceso de construcción

Dividiremos la construcción del dataset en tres etapas:

1. Recolección: Proceso de recolección de datos de Twitter y de los artículos periodísticos

2. Selección: Proceso de selección del conjunto de artículos y comentarios recolectados a etiquetar

3. Anotación: Proceso de etiquetado de los artículos seleccionados

Si bien en muchos casos las dos primeras etapas suelen ser la misma o bien la selección se limita a una muestra aleatoria de la recolección, este procedimiento sería muy ineficiente para nuestro estudio. Esto se debe a que en el dominio de comentarios periodísticos y discurso de odio, encontramos este tipo de discurso distribuido de manera muy poco uniforme, usualmente concentrado alrededor de ciertos tópicos disparadores. Para poder recolectar datos con una proporción razonable del fenómeno estudiado, evaluamos algunas posibilidades de selección de los artículos y sus respectivos comentarios.

En algunos trabajos previos, la recolección y selección constan conjuntamente de la búsqueda en base a ciertas palabras clave, que son utilizadas para recolectar tweets o bien para preseleccionar usuarios productores de discurso de odio [10, 165]. En nuestro caso, la selección de artículos y comentarios presenta cierta novedad y complejidad, con lo cual separamos este procedimiento para explicarlo detalladamente en las siguientes secciones.

5.4. Recolección de datos

En esta sección detallamos el proceso de recolección de datos, cuya salida es un conjunto de artículos mencionados en Twitter y sus comentarios respectivos realizados por usuarios. Describimos a continuación las decisiones realizadas respecto a las fuentes y a otros detalles técnicos.

En primer lugar, limitamos nuestra recolección de datos a cuentas de medios de la República Argentina y, puntualmente, nos centramos en diarios con comunidad mayormente rioplatense. Esto lo realizamos teniendo en mente que los anotadores serían nativos de esta variedad dialectal ya que, como mencionamos anteriormente, el discurso de odio contra mujeres, grupos nacionales y otros depende fuertemente de la jerga y de las variaciones dialectales de cada lugar. Esta elección, se debe además a que, habiendo buscando en otros medios de Argentina (como por ejemplo "La voz del Interior", diario dirigido mayormente a un público fuera de la Metrópolis de Buenos Aires) observamos que la interacción en Twitter de estos medios es muy baja, con muy pocos usuarios comentando sus notas. Centrándonos en diarios que generen interacción, seleccionamos medios periodísticos de gran llegada y tradicionales, los cuales listamos en la Tabla 5.1.

Si bien recolectamos notas de otros medios, no los consideramos a partir de ahora, y los dejamos para análisis posteriores. De los cinco medios elegidos, todos son medios formales y con varios años en el medio, siendo cuatro de ellos con soporte escrito y uno sólo (Infobae) enteramente digital. Consideramos la posibilidad de elegir medios no tradicionales y más orientados a grupos de la "derecha alternativa", dada su alta incidencia de contenido de odio. Sin embargo, finalmente tomamos la decisión de descartarlos de la etapa de anotación.

5.4.1. Método de recolección

La API de Twitter, en su versión gratuita, nos brinda dos modos de recolectar tweets de su plataforma[3]:

1. *Search API*: permite buscar tweets en base a términos, de hasta 15 días atrás sobre una pequeña muestra, recreando lo que vemos en la UI de Twitter

2. *Stream API*: permite buscar tweets en tiempo real sobre una muestra de cerca del 1 % de todos los tweets de la red social

La *Stream API* (también conocida como *Spritzer*), mientras por un lado limita temporalmente la recolección de datos, por el otro nos brinda la posibilidad de recolectar una mayor cantidad de información en tiempo real. Más aún, dada la naturaleza de nuestro problema (discurso de odio), se corre el riesgo de que ciertos tweets con el tiempo sean moderados e inaccesibles para cualquier búsqueda con la *Search API*.

Por lo explicado, usamos la *Stream API* de Twitter, buscando tweets que mencionen a cualquiera de las cuentas de medios periodísticos listadas en la Tabla 5.1. Si estamos entonces recolectando tweets sobre `@medio`, la *Search API* nos da:

1. Tweets de `@medio`

2. Respuestas a los tweets de `@medio`

[3] Usamos la versión 1.1 de la API. La versión 2.0 parece facilitar la recopilación de conversaciones. Recomendamos investigar mejor esta versión actualizada para esquivar muchas de las dificultades técnicas que incursionamos para lo descripto en esta sección

3. Tweets de terceros que mencionan a `@medio`
4. Retweets (RT) de tweets de `@medio`
5. Citas de tweets de `@medio`

Los RTs y tweets que simplemente arroben a `@medio` carecen de interés para nuestro estudio, con lo cual los descartamos. Por otro lado, también descartamos las citas, aunque podrían entenderse en algún punto como respuestas a los tweets originales. Nos quedamos finalmente con tweets de `@medio` y las respuestas a estos. Si bien la API nos da estos tweets de manera desestructurada, reconstruimos el árbol de la discusión mediante el campo `in_reply_to_status_id`[4], que marca el tweet al que responde.

Algo importante a remarcar es que, para el propósito de este trabajo, sólo estamos interesados en el primer nivel de respuestas al tweet original, y no incorporamos hilos de respuestas. Trabajo futuro debería explorar este nivel adicional de complejidad incorporando contexto conversacional adicional.

Accidentalmente, la recolección de datos se dio al mismo tiempo del estallido de la pandemia del COVID-19. Dadas las implicancias de la pandemia sobre el discurso discriminatorio en las redes sociales, se volcó el foco de la recolección hacia artículos relacionados con el coronavirus. Para ello seleccionamos artículos buscando la ocurrencia de ciertas palabras en el cuerpo del artículo, específicamente relacionadas al COVID-19: *coronavirus, encierro, síntomas, covid, fase, fiebre, cuarentena, infectados, distanciamiento, normalidad, Wuhan, aislamiento.*

Por último, nos quedamos con aquellos tweets de los medios periodísticos que tuvieran un link a un artículo. Para ello, utilizamos el módulo de Python *newspaper3k*[5], que permite acceder a la información relacionada a los artículos en cuestión, en particular siendo lo que más nos interesa el cuerpo del artículo. Aquellos tweets de medios periodísticos que no contengan un link a un artículo fueron descartados por considerar que no representaban una noticia en sí.

5.4.2. Datos recolectados

La Tabla 5.2 contiene la cantidad de artículos recolectados por cada medio, luego de ser aplicado el filtro de palabras mencionado en la anterior sección, y cantidad de comentarios por medio. Si bien recolectamos más artículos de otros medios, no son enumerados, aunque fueron conservados para otros experimentos. Entre los medios seleccionados, Infobae fue el más prolífico en artículos y también será finalmente sobre el que más comentarios etiquetemos. En el apéndice B.1 puede encontrarse la distribución temporal de los datos. Si bien tenemos un pequeño bache en los datos por un problema técnico en la recolección, los artículos de este conjunto datan de Marzo del 2020 hasta Febrero del 2021.

En la siguiente sección realizamos un filtrado de la mayoría de estos artículos previamente a la anotación. En conjunto a los datos de otros medios, estos artículos

[4] Ver la documentación y la referencia al campo en `https://developer.twitter.com/en/docs/twitter-api/v1/data-dictionary/object-model/tweet`

[5] `https://newspaper.readthedocs.io/en/latest/`

Medio	#Artículos	#Comentarios
@infobae	45,652	822,462
@clarincom	29,050	672,401
@perfilcom	8,764	61,203
@LANACION	16,040	506,091
@cronica	17,250	70,872
Total	116,756	2,133,029

Tab. 5.2: Elementos recolectados por cada medio, medidos en cantidad de artículos (tweets de los medios + artículos correspondientes) y cantidad de comentarios (respuestas en Twitter a los primeros)

y comentarios no etiquetados son preservados para efectuar ajustes de dominio, y serán liberados como se recomienda en Gururangan et al. [68]. Hablaremos más sobre esto en los Capítulos 6 y 7.

5.5. Selección de datos a anotar

Un problema que se presenta antes de comenzar el etiquetado es el de seleccionar los artículos que vamos a etiquetar, teniendo en consideración la gran cantidad de datos recolectados y los recursos disponibles. Una primera posibilidad para esto es realizar una selección aleatoria de artículos y comentarios. Sin embargo, los comentarios discriminatorios no se distribuyen de manera uniforme entre los artículos sino que se concentran sobre algunos temas que generan este tipo de contenido. Es mucho más probable encontrar comentarios de índole discriminatoria en notas que tengan temas cercanos a alguna de las características protegidas: por ejemplo, es esperable encontrar contenido discriminatorio en notas sobre China y el Coronavirus o sobre una chica transgénero antes que en un artículo de fútbol o economía. Si bien una selección aleatoria preservaría una tasa de incidencia mucho más cercana a la observada en el universo de comentarios, es más importante poder obtener una mayor cantidad de observaciones que reflejen el fenómeno estudiado.

Teniendo esto en cuenta, evaluamos varias alternativas para realizar la selección de artículos. La primera fue intentar seleccionar aquellos artículos que consideramos como candidatos a fomentar contenido discriminatorio. Una posibilidad para esto sería usar algunas palabras semilla para seleccionar artículos interesantes en base a ciertos temas que consideramos relevantes.

Otra posibilidad evaluada fue la de buscar directamente comentarios que marquen que ese artículo suscita contenido discriminatorio. Para ello, podemos listar algunos insultos comunes o expresiones peyorativas hacia los grupos protegidos considerados. Es necesario remarcar que esto lo hacemos para seleccionar **artículos** y no los comentarios que contengan esos insultos; hacer esto último nos genera una muestra muy distorsionada y tendiente a encontrar el fenómeno más explícito de la discriminación (el insulto racista, homofóbico, etc.). Esta estrategia guarda relación con la descripta por Basile et al. [10] para seleccionar usuarios generadores de contenido discriminatorio.

Describimos a continuación las alternativas analizadas para seleccionar los ar-

tículos y sus respectivos comentarios.

5.5.1. Selección en base a artículos

En primer lugar, consideramos la posibilidad de hacer una selección en base al contenido de los artículos. Luego de hacer un análisis exploratorio de los datos usando LDA [21] para buscar tópicos posibles de las notas, decidimos realizar una selección controlada y determinística en base a la utilización de palabras y expresiones clave. Estas expresiones las recolectamos de manera subjetiva y en base a la observación de los tópicos y de nuestra percepción de la generación de discurso discriminatorio en los comentarios de los usuarios.

La Tabla 5.3 muestra el conjunto de expresiones utilizado para recolectar artículos. Como vemos, hay diversas palabras que recogen temáticas de posibles tópicos generadores de contenido discriminatorio, algunos muy locales respecto a eventos concretos durante la pandemia. Si algún artículo contiene una de las expresiones mencionadas, es seleccionado para ser etiquetado.

Para realizar esta búsqueda de términos en el cuerpo de los artículos, indexamos los textos en *MongoDB*[6], una base de datos no relacional. Este motor de bases de datos permite la utilización de índices en base a texto, permitiendo realizar búsquedas en base a expresiones, palabras, e inflexiones.

5.5.2. Selección en base a comentarios

Otra posibilidad para seleccionar artículos candidatos a ser etiquetados es la de observar los comentarios de usuarios en lugar del texto completo de éste. En base a los comentarios, podemos tener alguna medida de si el artículo suscita reacciones potencialmente discriminatorias. Por ejemplo, si observamos que en un artículo hay comentaristas que usan expresiones discriminatorias contra la comunidad LGBTI, podemos pensar que el contenido de la noticia es interesante para nuestro estudio.

El procedimiento para este tipo de selección es similar al mencionado anteriormente con artículos, sólo que aplicado a comentarios: buscamos respuestas de usuarios que contengan alguna de las expresiones semilla listadas en la Tabla 5.4. Es-

[6] `https://www.mongodb.com/`

China	piqueteros	mamá	domésticas
Cuba	villas	de género	la modelo
cubano	la villa	aborto	la periodista
bolivia	movimientos sociales	actriz	la cantante
paraguayo	organizaciones sociales	actrices	travesti
judío	tomas de tierras	feminista	trans
camionero	toma de tierras	femicidio	gay
ladrón	sindicatos	enfermera	homosexual
represión	Guernica	madre	de la V
criminal	mapuches	Ofelia	

Tab. 5.3: Palabras semilla utilizadas para la selección de artículos. Cada palabra se busca sobre el cuerpo del artículo candidato a ser etiquetado

bija	urraca	viejo puto	trolo	peruano	matarlos	negra
prostituta	tucán	trabuco	sodomita	peruca	una bomba	negro de
feministas	putita	travesti	chinos de	judío	vayan a laburar	negros
feminazis	reventada	trava	bolita	sionista	vayan a trabajar	bala
aborteras	marica	degenerado	paraguayo	villeros	gorda	uno menos

Tab. 5.4: Palabras utilizadas para recolectar comentarios. Cada palabra se busca sobre el texto de un comentario para marcarlo como potencialmente discriminatorio.

tas palabras fueron recolectadas de manera subjetiva en base a la observación y a la experimentación sobre los datos, tratando de contener diversas expresiones de contenido mayormente discriminatorio. La lista contiene expresiones ofensivas para diversas características de interés: insultos racistas, homofóbicos, misóginos; insultos dirigidos dirigidos a algún personaje particularmente atacado en las redes sociales; expresiones de odio de clase; etc.

Dado un artículo, marcamos los comentarios que contengan una o más de las expresiones listadas. Si el artículo tiene tres o más comentarios marcados, entonces el artículo es seleccionado; caso contrario, es descartado. Vale remarcar que este proceso de selección es para los *artículos*, no para los comentarios. De lo contrario, sólo buscaríamos respuestas que contengan alguna de estas expresiones.

Luego de algunos análisis experimentales y observacionales de las dos posibles metodologías, decidimos utilizar el muestreo de artículos en base a comentarios. En base a un análisis subjetivo, los artículos seleccionados parecían tener mayor incidencia de mensajes discriminatorios y eso nos decantó hacia esa opción.

Una posibilidad adicional analizado fue utilizar un clasificador que nos señale posibles comentarios discriminatorios, usando esta información para seleccionar artículos candidatos a etiquetar. Para ello, aplicamos un clasificador basado en BETO [31] entrenado sobre el dataset de *HatEval* (ver Sección 4) sobre los comentarios de los artículos. Una evaluación subjetiva de esto nos dio pobres resultados, tanto porque no captaba algunas agresiones discriminatorias (de características no incluidas en el dataset de Basile et al. [10]) como muchos falsos positivos o errores debido al cambio de dominio (temático y también dialectal). Si bien descartamos este método, puede ser de relevancia usar algún método que no esté basado en palabras semillas o utilizar algún método semi-automático para encontrar candidatos a etiquetar.

5.5.3. Muestreo de comentarios

Una vez que seleccionamos los artículos, resta decidir qué comentarios vamos a anotar. No podemos seleccionar todos ya que muchos artículos cuentan con una cantidad importante de comentarios (en el orden de los cientos) y es deseable mantener un balance entre los comentarios anotados por artículo. Tampoco es deseable (en pos de maximizar el producto de la anotación) seleccionar comentarios de artículos escasamente discutidos. Teniendo esto en mente, conservamos sólo los comentarios de artículos que tengan al menos 20 comentarios. Luego, para cada artículo, seleccionamos aleatoriamente hasta 50 comentarios entre aquellos que no contengan URLs u otro contenido no textual.

En este punto, consideramos el muestreo aleatorio como la forma menos sesgada

para seleccionar nuestros comentarios, pero mencionamos de todas formas algunas alternativas evaluadas. Una fue la de considerar todo el universo de comentarios y seleccionar la muestra de allí. Sin embargo, esto sobrerrepresentaría a aquellos temas muy comentados, siendo muchos de ellos acerca de temas políticos que se filtraron en nuestra selección. Otra consideración posible es la de utilizar información de usuarios y sus conexiones, información que Twitter nos brinda a través de los followers de cada usuario. Muchos usuarios que generan contenido discriminatorio en redes sociales se agrupan en comunidades: subgrafos de usuarios altamente conectados entre sí. Usar algún tipo de información sobre esto (por ejemplo, con algún algoritmo como el de Louvain [22]) podría auxiliar al balance de comentarios posiblemente discriminatorios. Para un ejemplo de la utilización de esta técnica, Lai et al. [90] y Furman et al. [54] usan este tipo de algoritmos como manera semi-supervisada de detectar las posturas de los usuarios respecto a distintos temas.

5.6. Anotación

Hasta este momento describimos la recolección de los datos, a lo cual le siguió la selección de los artículos y comentarios a anotar. Pasamos ahora a detallar el último paso de la construcción del conjunto de datos: el etiquetado. En primer lugar, adoptamos nuestra propia definición de discurso de odio, con la cual confeccionamos el respectivo manual de etiquetado. Con esto en mano, definimos las variables que nos interesa anotar sobre los datos. Llamamos a esto *modelo de etiquetado* [134].

Finalmente, especificamos el proceso concreto de etiquetado. Por un lado, la selección de anotadores, sus perfiles y la herramienta de etiquetado. Por el otro, el esquema utilizado para distribuir el trabajo a los anotadores.

5.6.1. Definición de discurso de odio y manual de etiquetado

Teniendo en cuenta las consideraciones de la Sección 4.1, elaboramos nuestra propia definición de discurso de odio. Entendemos que hay discurso de odio en un comentario de una red social si éste contiene declaraciones de carácter intenso e irracional de rechazo, enemistad y aborrecimiento contra un individuo o contra un grupo de personas por poseer (o aparentar poseer) una característica protegida

Nombre	Descripción
MUJER	Misoginia, agresiones basadas en ser mujer
LGBTI	Homofobia, transfobia, y ofensas a la comunidad LGBTI
RACISMO	Racismo, Xenofobia, Judeofobia, etc
POBREZA	Basado en su condición de clase
POLITICA	En base a la filiación política del agredido
ASPECTO	Gordofobia, gerontofobia
CRIMINAL	Criminales, presos, y personas en conflicto con la ley
DISCAPACIDAD	Discapacidades y problemas de adicciones

Tab. 5.5: Características protegidas consideradas en este trabajo. Consideramos una agrupación de ciertas características bajo una misma denominación: por ejemplo, LGBTI contempla homofobia, transfobia, entre otras; análogamente racismo puede contemplar xenofobia, y otras variantes de este fenómeno.

[7]. Esta expresión puede manifestarse de manera explícita como insultos directos, celebraciones de crímenes, incitaciones a tomar medidas contra el individuo o grupo, o también expresiones más veladas, por ejemplo de contenido irónico. Consideramos en esta definición que no es suficiente un insulto o una agresión para que se configure discurso de odio: es necesario hacer una apelación explícita o implícita al menos a una característica protegida.

A diferencia de otros trabajos, nuestra definición comprende varias características, incluso algunas que están en la frontera de ser protegidas. Los estudios previos que hemos listado en la Sección 4.2 han estado centrados mayormente en racismo y misoginia, con algunos que han agregado homofobia. En nuestra definición comprendemos –además de la misoginia y el racismo– a la homofobia y transfobia; al odio de clase (a veces conocido como aporofobia); al odio por aspecto físico; a la discriminación por discapacidad o problemas de salud; entre otras. En particular, hay dos características no convencionales que tuvimos en cuenta. En primer lugar, consideramos el **discurso de odio político**, que, de acuerdo a la CIDH [37], es difícil considerar como protegido ya que puede dar lugar a censura y restricciones a la libertad de expresión. Por otro lado, consideramos el **discurso de odio contra criminales**, presos, y otras personas en situación de conflicto con la ley. Si bien este punto ni siquiera es considerado como protegido en ninguno de los tratados mencionados en la Sección 4.1, agregamos esta característica debido a la enorme cantidad de contenido alentando la violencia contra criminales en las noticias policiales. Teniendo en cuenta que nos interesa detectar particularmente incitaciones a realizar acciones violentas contra individuos o grupos, relajamos nuestra definición para incluir esta clase de agresiones.

Tenemos entonces ocho características que agrupan distintos tipos de discurso de odio: contra las mujeres; racismo y xenofobia; contra la comunidad LGBTI; odio de clase; gordofobia, gerontofobia y demás discurso de odio por aspecto; por su ideología política; contra criminales; y finalmente contra discapacitados y adictos. Las características en cuestión son listadas en la Tabla 5.5 junto a nombres de referencia que serán usados en éste y el próximo capítulo.

Con esta definición confeccionamos un manual de referencia para los anotadores. Tanto el manual como la definición fueron desarrollados iterativamente, primero realizando algunas pruebas de etiquetado entre miembros del equipo y rondas de discusión posteriores analizando los ejemplos problemáticos y casos borde. De estas iteraciones logramos ir mejorando la definición y el manual hasta llegar a una versión definitiva. Para cada característica agregamos consideraciones adicionales sobre lo que pensamos que configura discurso de odio: por ejemplo, para la característica MUJER no es suficiente con que un insulto sino que es necesario se apele a algo distintivo de la mujer (“algo que no le diría a un hombre”); para la característica LGBTI incluimos particularmente las expresiones de asco, y puntualmente a los emojis que representen esto. En el Apéndice B.3 puede encontrarse el manual de etiquetado completo entregado a los etiquetadores.

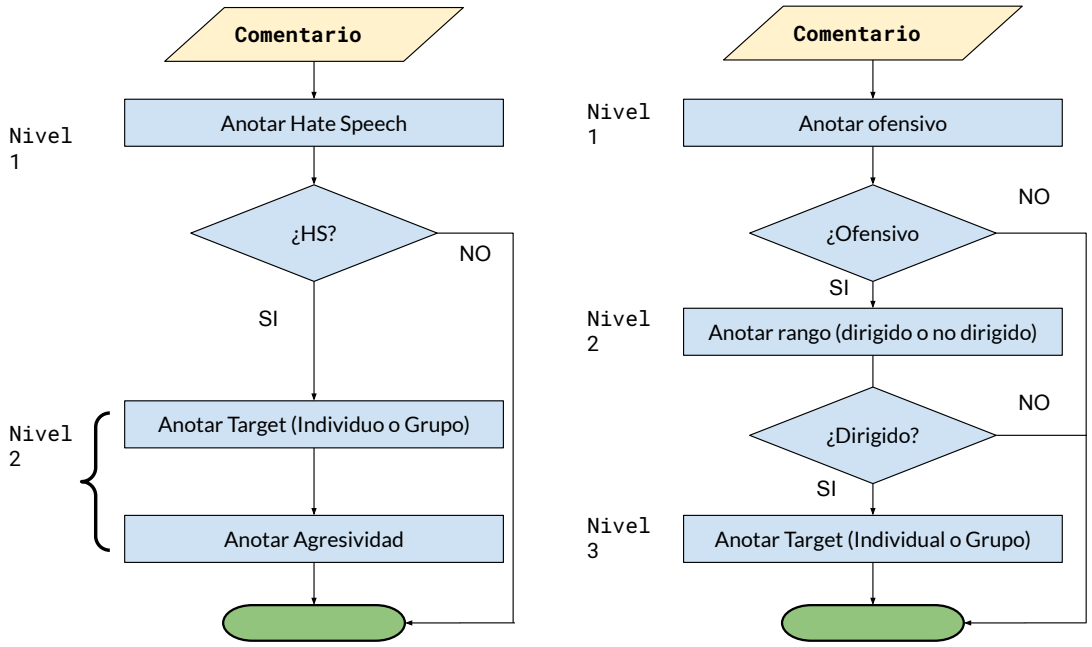


Fig. 5.2: Modelos jerárquicos de anotación. A la izquierda, tenemos el modelo jerárquico propuesto para HatEval [10], a la derecha el modelo propuesto para OffensEval [175]

5.6.2. Modelo de etiquetado

Un modelo de anotación es una representación práctica del objetivo de este proceso; es decir, del fenómeno que queremos capturar [134]. En base a la discusión del capítulo anterior, es de interés marcar comentarios discriminatorios de manera granular de modo de tener información de qué grupos y/o características se está ofendiendo. Para identificar formas más graves de discurso de odio, también es de interés identificar llamados a tomar alguna acción (violenta o no violenta) contra esa persona o grupo.

Zampieri et al. [174] introdujeron un modelo jerárquico de anotación para la tarea de lenguaje ofensivo, utilizado tanto en los datasets de OffensEval [175] y HatEval [10]. La idea de este modelo es realizar anotaciones en varios niveles, sólo marcando algunas variables de acuerdo a las respuestas del nivel anterior. Por ejemplo, en el caso de *HatEval*, tenemos un primer nivel que consta de marcar si un tweet contiene o no discurso de odio. Si el tweet tiene discurso de odio, entonces se anota en primer lugar si está dirigido a un individuo o a un grupo, y también si es agresivo o no. En el caso de *OffensEval*, primero se anota si es ofensivo, y en caso de serlo, se marca si está dirigido a un individuo o grupo o es un insulto no dirigido. Por último, si es dirigido y ofensivo, marcamos si su objetivo es un grupo o un individuo. La Figura 5.2 ilustra en modo de diagrama de flujo el modelo de anotación de ambos conjuntos de datos.

Basándonos en esta estructura jerárquica planteamos nuestro modelo, ilustrado en la Figura 5.3. Para cada comentario y su respectivo contexto (el artículo), reque-

[7] Ver Sección 4.1 para la definición de característica protegida

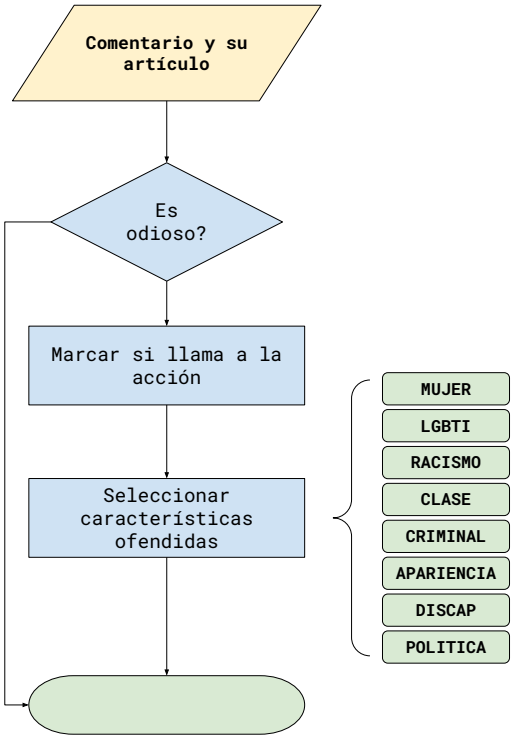


Fig. 5.3: Modelo de anotación para el dataset construido en este capítulo. El modelo jerárquico consta de dos niveles: el primero, donde se anota si es odioso. El segundo consta de anotar –en caso de haber sido marcado como odioso– si contiene un llamado a la acción y a qué características ofende.

rimos una anotación para decidir si el comentario es odioso o no. Si no es odioso, no se necesita más información. En caso de haberse marcado como odioso, el par artículo-comentario debe contener, además, una anotación por si llama o no a la acción, y al menos una categoría protegida marcada como ofendida.

5.6.3. Etiquetadores

A diferencia de otros trabajos de detección de discurso de odio, decidimos garantizar que nuestros anotadores estuvieran más cerca culturalmente al problema en cuestión. El discurso de odio tiene un fuerte componente cultural, muchas veces expresado a través de jerga o expresiones dialectales muy particulares, y está relacionado con noticias muy propias de esta región. Es por esto que decidimos buscar por nuestros propios medios perfiles alineados a estos puntos y no depender de plataformas externas de crowdsourcing para esta tarea.

Reclutamos etiquetadores hablantes nativos de español rioplatense, estudiantes o graduados/as de carreras de ciencias sociales, humanidades o afines – como ser Psicología, Sociología, Comunicación, Artes, Antropología. Para evitar sesgos en la tarea, nos interesó que no tuvieran conocimientos de inteligencia artificial o ciencia de datos. Por último, también fue de interés que sean usuarios asiduos de redes

sociales para poder captar las sutilezas del lenguaje en ese medio.

El proceso de reclutamiento constó de una breve entrevista donde corroboramos que los etiquetadores fueran hablantes nativos de español rioplatense, a la vez que les describimos la tarea que debían realizar y la herramienta correspondiente de etiquetado. Luego de la entrevista, se les solicitó hacer una prueba paga que constó de leer el manual de etiquetado y anotar 10 artículos. Esto fue realizado para corroborar la calidad de los etiquetadores, aunque no rechazamos ningún postulante en este proceso. La Tabla 5.6 brinda información desagregada sobre los seis etiquetadores contratados para la tarea. Los etiquetadores reclutados tienen un perfil altamente escolarizado, y dos de quienes contratamos con experiencia previa en la tarea de anotación. Un punto adicional a marcar es que dos de las etiquetadoras eran activistas –al momento de realizarse el estudio– en organizaciones relacionadas a alguno de los grupos vulnerados que estudiamos en este trabajo.

Luego de la entrevista, se les dio una devolución de su anotación y se les reasignaron cinco de los artículos seleccionados junto a diez más (15 en total) para su anotación a modo de entrenamiento. Este fue el único conjunto de artículos que fue anotado por la totalidad de los anotadores. Al finalizar esta etapa, se les brindó una nueva devolución para ajustar el criterio de anotación, y se procedió a la etapa de anotación del dataset.

5.6.4. Esquema de anotación

Pasamos ahora a describir la mecánica del proceso de anotación. Para lo que respecta a este trabajo, tomamos como la unidad de anotación al artículo. Cada etiquetador, al serle presentado un artículo, tuvo dos opciones: etiquetarlo o saltearlo. En caso de decidir etiquetarlo, tuvo que marcar las etiquetas correspondientes a cada uno de los comentarios del artículo de acuerdo al modelo descripto. La idea de permitir el salteado fue doble: evitar contenido poco interesante en términos de comentarios discriminatorios, y también dar la posibilidad al trabajador de evitar contenido sensible o perturbador para su persona.

Considerando la dificultad de la tarea y los límites borrosos del discurso de odio, decidimos seguir un esquema de etiquetado similar al de trabajos previos donde múltiples personas anotan una misma instancia. Una posibilidad considerada en un principio fue asignar el artículo completo a tres anotadores; sin embargo, esta moda-

Género	Edad	Estudios	Área	Identificación	¿Activista?	Experiencia
F	25-30	Doctorado*	Psicología	Mujer	No	Sí
NB	30-35	Grado*	Artes	LGBTI	No	No
F	30-35	Grado*	Antropología	Mujer, LGBTI	Feminista	Sí
M	35-40	Grado	Sociología	No	No	No
F	35-40	Doctorado	Psicología	Mujer	No	No
F	30-35	Grado	Comunicación	No	Migrantes	No

Tab. 5.6: Información sobre los anotadores. En el caso de estudios, * indica en curso. Indentificación se refiere a si se autopercibe como perteneciente de una característica protegida considerada en este trabajo. Experiencia se refiere a haber etiquetado previamente otros conjuntos de datos.

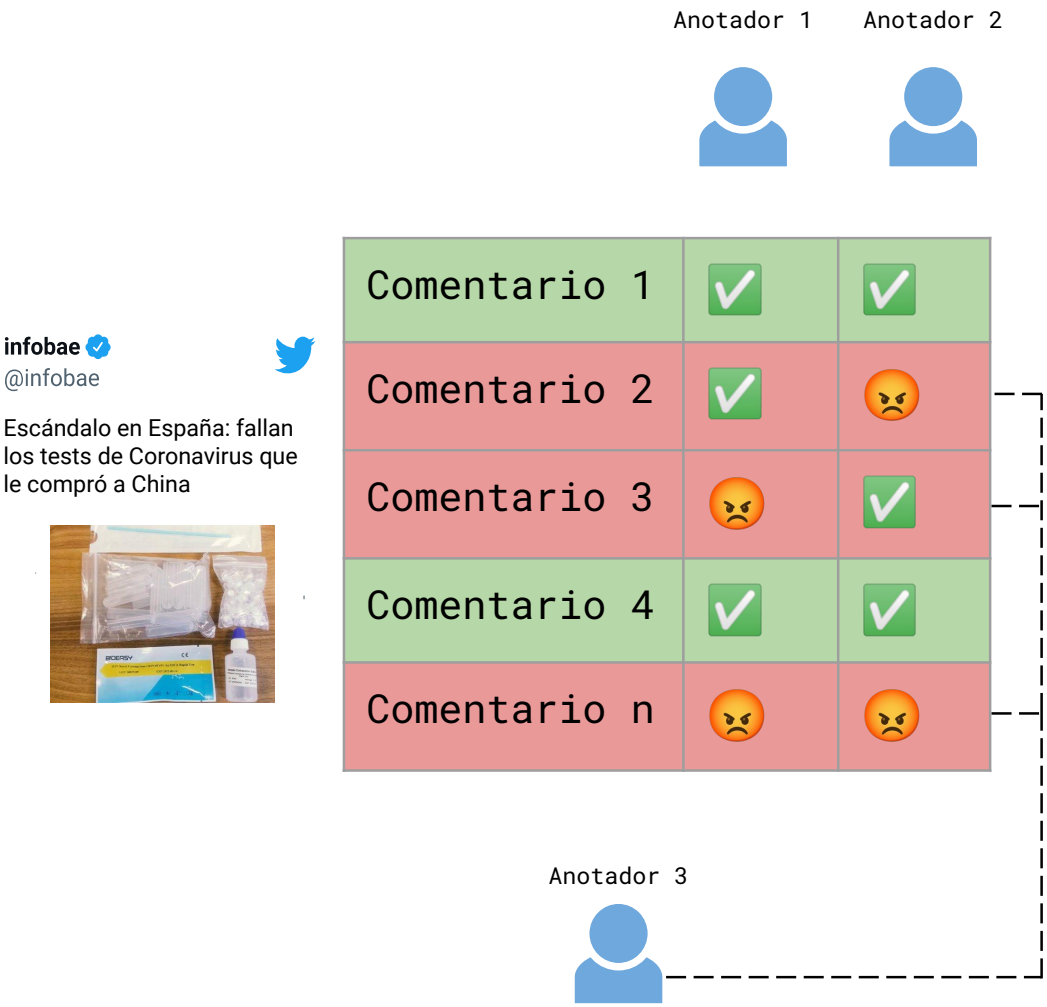


Fig. 5.4: Esquema de anotación. Caso en que ambos anotadores etiqueten los comentarios del artículo

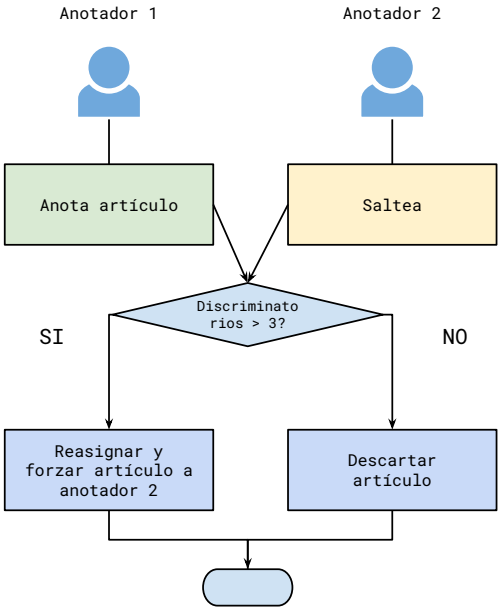


Fig. 5.5: Esquema de anotación. Caso en que un anotador decida saltear

lidad sería ineficiente dada la baja cantidad de contenido discriminatorio observada preliminarmente entre los comentarios seleccionados. Decidimos entonces ir por un esquema de desempate similar al de Basile et al. [10]: dos personas anotan un artículo, y luego un tercero anota sólo aquellos comentarios donde al menos uno marcó que existe contenido odioso. Esta modalidad permite que haya una tercera anotación incluso cuando las dos previas marcaron contenido odioso, siendo esto realizado para recolectar más información sobre los comentarios. Dada la incidencia de comentarios odiosos (que veremos luego en los resultados) la adición de esta tercera etiqueta en estos casos técnicamente innecesarios tuvo un bajo costo.

La Figura 5.4 ilustra el flujo en el caso de que ninguno de los anotadores asignados decida saltear el artículo. Ahora ¿qué pasa si alguno de los dos anotadores decide saltear el artículo?

1. Si los dos etiquetadores asignados deciden pasar por alto el artículo asignado, entonces lo descartamos del conjunto de datos

2. En el caso de que uno lo saltee y el otro lo anote y encuentre menos de 4 comentarios odiosos, entonces también descartamos el artículo

3. En el caso de que uno lo saltee y el otro lo anote encontrando cuatro o más comentarios odiosos, entonces reasignamos el artículo al etiquetador que salteó, sin darle esta vez posibilidad a que no lo anote

Tomamos la decisión de descartar los artículos en los dos primeros casos intentando maximizar la tasa de comentarios discriminatorios encontrados. En caso de

reasignar el artículo, revocamos la posibilidad de saltearlo [8]. La Figura 5.5 ilustra el esquema de anotación de artículos recién descripto, que se complementa con el de la Figura 5.4.

Como resultado de este esquema, cada comentario de nuestro dataset puede tener dos o tres anotaciones, siendo los casos posibles los siguientes:

1. Dos anotaciones negativas: es decir, que no encuentran discurso de odio en el comentario

2. Tres anotaciones, con al menos una que marque el comentario como odioso

5.6.5. Herramienta de etiquetado

Al no optar por utilizar servicios de etiquetado tercerizado, desarrollamos nuestra propia aplicación para esta tarea. Cada etiquetador tuvo acceso a un sitio web donde se le mostraron uno a uno los artículos asignados, de acuerdo al orden establecido por los administradores de la aplicación. La Figura 5.6 ilustra la interfaz de anotación presentada a los usuarios del sistema. Cada artículo es presentado junto a su tweet correspondiente, al texto replegado de la noticia –en caso de que un usuario requiera su lectura– y a los comentarios.

Ante esto, el etiquetador puede elegir saltear el artículo o etiquetarlo. Si decide etiquetarlo, debe para cada comentario marcar usando un control de tipo interruptor:

1. Si el comentario contiene discurso discriminatorio.

2. En caso de ser discriminatorio, marcar si llama a la acción.

3. En caso de ser discriminatorio, marcar al menos una característica ofendida.

Para el desarrollo de la aplicación usamos *Django* [9], un framework de Python para desarrollo web, y Javascript plano. Como base de datos utilizamos *SQLite*, ya que el sistema no tuvo grandes requerimientos de concurrencia (sólo seis o siete usuarios simultáneos).

Cada tweet fue presentado con un preprocesado básico que consistió en reemplazar handles de usuarios por un token especial `@usuario` para evitar cualquier sesgo. Por ejemplo, si un usuario A conocido como difusor de discurso discriminatorio retwittea la noticia y otro responde a ese retweet, en el tweet aparece el nombre de A, lo cual podría condicionar a quien tenga que evaluar ese comentario.

5.6.6. Asignación

Llamamos **asignación** al procedimiento de colocar **etiquetas** (también llamadas *gold labels*) a las instancias de nuestro conjunto de datos [134]. El modelo descripto en la Sección 5.6.2 consta de una etiqueta binaria que marca si el contenido es

[8] Teniendo en cuenta la posibilidad de que hayan salteado por contenido perturbador para el anotador, dimos la posibilidad de que nos avisen que no querían trabajar en ese artículo. No hubo problemas al respecto de todas formas

[9] `https://www.djangoproject.com/`

Fig. 5.6: Pantalla del etiquetador

discriminatorio o no (notamos HS) en el primer nivel, y luego 9 etiquetas binarias para el segundo: una para las llamadas a la acción (notamos LLAMA) y otras ocho para las características ofendidas listadas en la Tabla 5.5. Recordemos que una anotación negativa sólo consta de HS negativo, mientras que una positiva consta de un HS positivo, una etiqueta para LLAMA y al menos una etiqueta positiva de las características ofendidas.

Pensamos el proceso de asignación de etiquetas como una votación, donde cada anotador da un voto para cada variable a asignar. Detallamos a continuación cómo fueron asignadas las etiquetas del conjunto de datos:

1. Para la etiqueta de HS, asignamos mediante votación mayoritaria: 2 o más votos para HS positivo, caso contrario HS negativo.

2. En caso de haber marcado que hay HS: marcamos LLAMA es positivo por votación mayoritaria.

3. En caso de haber marcado que hay HS: marcamos como positivas todas aquellas características marcadas por los anotadores.

La primer decisión es la más obvia y razonable: para que un comentario sea considerado como odioso (HS positivo) tiene que ocurrir que al menos dos etiquetadores lo marquen como tal. Si se anota HS positivo, para que LLAMA sea positivo tiene que haber al menos dos votos en tal sentido. Si un anotador marcó que hay llamado a la acción y otro que no, asignamos LLAMA negativo.

En el caso de las características no realizamos votación mayoritaria, sino que la anotamos si al menos un etiquetador lo hizo. Esta decisión podría haberse tomado de otra manera; por ejemplo, sólo tomando aquellos casos donde haya cierto grado de coincidencia entre los comentarios. Sin embargo, al considerar que los límites entre las características son difusos (por ejemplo, APARIENCIA y MUJER tienen una intersección no nula y a veces CLASE y RACISMO también) preferimos anotarlas de esta manera.

5.7. Resultados

El conjunto resultante consta de 1,238 artículos etiquetados, y 56,869 comentarios respectivamente, de los cuales 8,715 contienen contenido discriminatorio según los criterios de asignación antes referidos. Aproximadamente 1 de cada 6 comentarios es discriminatorio, aunque vale aclarar que esto no es representativo del universo de notas periodísticas ya que la selección de los datos no fue aleatoria.

La Tabla 5.7 contiene los números de los comentarios anotados y desagregados por las distintas características consideradas y los llamados a la acción. La categoría con más comentarios es RACISMO, seguido por APARIENCIA y CRIMINAL. Dentro de los tweets que llaman a algún tipo de acción, se coloca en primer lugar los dirigidos hacia la categoría CRIMINAL, muchos en la forma de llamados a matar a criminales y delincuentes. La categoría RACISMO acapara también muchos llamados a la acción, mayormente contra población china a la que se culpa de la

Característica	Cantidad	Llamadas	α
RACISMO	2,469	674	0,93
APARIENCIA	1,803	34	0,87
CRIMINAL	1,642	722	0,93
POLITICA	1,428	136	0,81
MUJER	1,332	18	0,78
CLASE	823	135	0,71
LGBTI	818	11	0,92
DISCAPACIDAD	580	4	0,85
TOTAL	8,715	1,451	0,58

Tab. 5.7: Datos por característica de los comentarios discriminatorios del conjunto de datos resultantantes, junto a la tasa de acuerdo medida por α de Krippendorff. El acuerdo sobre *discurso de odio* es reportado sobre todos los etiquetadores que hayan analizado cada comentario. Para el resto de las características, el acuerdo es calculado sólo sobre aquellas anotaciones que marcaron discurso de odio.

pandemia del COVID-19 y conteniendo llamados a tomar distintas sanciones contra sus integrantes.

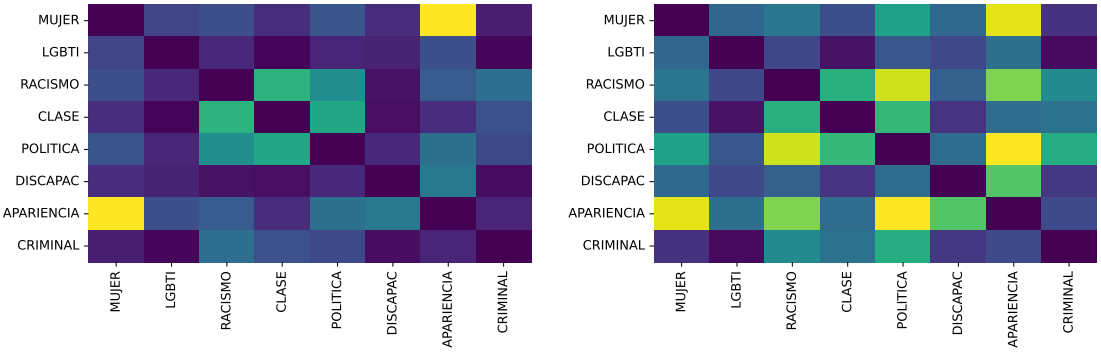
En la misma tabla se reporta el acuerdo entre anotadores usando la métrica *alfa de Krippendorff* [86]. Esta métrica mide el acuerdo entre diversos anotadores, donde 1 es acuerdo total, y 0 o valores negativos indican ningún tipo de acuerdo. Puede entenderse como una generalización de la métrica *kappa de Fleiss* para el caso en que los anotadores no etiqueten todas las instancias. Utilizamos para su cálculo la implementación en Python de la librería *krippendorff* [10]. Reportamos en primer lugar el acuerdo para HS sobre todas las etiquetas. En el caso de las etiquetas del segundo nivel del modelo jerárquico (características y llamado a la acción) calculamos el acuerdo sólo sobre aquellas que hayan marcado que el comentario contiene discurso de odio. Esto es equivalente en términos del cálculo propuesto en Krippendorff [86] a calcular el acuerdo con una etiqueta faltante en el segundo nivel para aquellos anotadores que hayan marcado que no hay HS.

Si bien el acuerdo sobre cada característica tiende a ser alto, debe leerse como el acuerdo sobre la razón detrás del discurso de odio. La mayor penalización queda reservada a la etiqueta de discurso de odio que tiene $\alpha = 0{,}58$, algo que podría marcarse como un acuerdo razonable teniendo en cuenta valores observados en la literatura [132].

5.7.1. Co-ocurrencia de características ofendidas

De los 8,715 comentarios odiosos, el 77 % de ellos (6,777) contiene una sola característica ofendida de acuerdo al proceso de asignación realizado. Cerca del 20 % tienen dos características ofendidas, y 220 comentarios tienen tres o más. La Figura 5.7 ilustra la matriz de co-ocurrencia entre las distintas características para aquellos comentarios que tengan más de una característica ofendida. En ella podemos ver que la máxima co-ocurrencia se da entre las características MUJER y APA-

[10] `https://github.com/pln-fing-udelar/fast-krippendorff`



(a) Co-ocurrencia de las características ofendidas en un comentario

(b) Co-ocurrencia de las características ofendidas en un artículo

Fig. 5.7: Matrices de co-ocurrencias de características ofendidas. La Figura 5.7a muestra la co-ocurrencia dentro de un mismo comentario, y la Figura 5.7b muestra la co-ocurrencia dentro de los comentarios de un mismo artículo. Más luminoso indica más co-ocurrencia

RIENCIA, seguidos por RACISMO y CLASE, POLITICA y CLASE, y RACISMO y POLITICA.

Otra forma de analizar la co-ocurrencia es agrupando por artículos las distintas características de sus comentarios, para así observar como un mismo contexto puede suscitar distintos tipos discriminación. La Figura 5.7b ilustra las interacciones entre las distintas características por artículo. Puede observarse en esta figura mayor dispersión en las co-ocurrencias que en la Figura 5.7a, apreciándose algunas interacciones adicionales como por ejemplo entre RACISMO y POLITICA y –quizás inesperadamente– entre APARIENCIA y POLITICA. La característica que parece tener menos interacción con las demás es LGBTI, indicando que está bien delimitada de las demás posibles causas de discurso discriminatorio.

La Tabla 5.8 muestra algunos ejemplos con más de una característica ofendida. Algunos de los comentarios se encuentran en la frontera de las características, como por ejemplo, APARIENCIA y MUJER en los ejemplos 1 y 2, o CLASE y RACISMO en el ejemplo 3. Otras instancias son conjunciones de expresiones discriminatorias, como en el caso del número 5, donde tenemos una conjunción de RACISMO, CLASE, y LGBTI; o bien en el comentario 6, de APARIENCIA y LGBTI.

5.7.2. Análisis por característica

Las tablas 5.9, 5.10 y 5.11 ilustran ejemplos seleccionados de comentarios discriminatorios para las distintas características estudiadas. Hacemos a continuación un análisis cualitativo y observaciones generales sobre cada categoría.

En primer lugar, en la Tabla 5.9 podemos apreciar que la característica MUJER revista cierta complejidad. En particular, algunos casos son de difícil interpretración, como las acusaciones de mentirosa a una mujer víctima de una violación[11], apreciaciones a su cuerpo, entre otros comentarios misóginos.

[11] `https://www.lavanguardia.com/gente/20181212/453520382646/denuncia-actor-juan-darthes-violar-thelma-fardin-argentina-patito-feo.html`

Índice	Contexto	Comentario
1	Ofelia Fernández apoyó al Gobierno en la polémica por los presos y apuntó a la Justicia que "odia a las mujeres"	Hijadept„ ojala pronto recibas la visita de alguno de esos gusanos. Te van a quedar. Ganas de apoyar al. Gobierno? Larva rastrera gorda. Decerebrada
2	"Es hora de ponerle límites al odio" | Por Victoria Donda	Justo ésta zurda mugrienta, ignorante y altanera...
3	Coronavirus en la Argentina: un video pone en evidencia la violación de la cuarentena en la Villa 1-11-14	Cierren esa nido de negros y napalm. Hasta reducís el crimen y el gasto público.
4	Fabiola Yáñez denunció a un periodista por publicaciones agraviantes	Claro si ofendel a la que se cuelga en el caño xq ahora cree ser primera dama? hay que ser peruka para dar asco y ser basuras bigote enseguida ordena como se metió en Facebook y en todo que culpa te.emos que saque la mujer del cabarute?
5	Los infectados en villas porteñas crecieron un 80 % en cuatro días	Ojalá que el virus penetre más en las villas y maten a todos esos delincuentes que viven ahi, hay paraguayos narcos, bolivianos que traen la droga de bolivia, y gente de mala vida. También hay travas que van a trabajar de noche a palermo.
6	El enojo de Moria Casán contra Rocío Oliva: "Mucha agua oxigenada, le quedó media neurona para jugar a la pelota"	Y la vieja Moria, mucha cirugía y estiramiento. de cara que parece un travesti
7	Ricky Martin: "Soy un hombre latino y homosexual viviendo en los Estados Unidos, soy una amenaza"	Ridículo perdiste tú rumbo das náuseas famosos eternos (víctimas) ándate a Puerto Rico entonces ahí no serás una amenaza

Tab. 5.8: Ejemplos con más de una característica ofendida marcada

Una categoría desafiante es la de los comentarios discriminatorios contra la comunidad LGBTI. Más allá de algunos insultos explícitos (*trolo, trabuco, maricón*, etc), hay muchas instancias que tienen un contenido difícil de descifrar. Particularmente, aquellos comentarios contra personas transgénero. Muchos de estos mensajes discriminatorios hacen alusiones a su genitalidad o a su cuerpo en general, de manera metafórica o irónica, puntos que se presentan como desafiantes para algoritmos de detección automática. A su vez, es claro que es sumamente necesaria la información contextual para poder comprender el caracter abusivo de estos comentarios, algo que muchas veces ni siquiera queda claro del artículo ya que no todos mencionan –ni tienen por qué hacerlo– el género de la persona atacada.

En el caso de la categoría CRIMINAL ilustrada en la Tabla 5.10, se puede observar por un lado comentarios muy violentos (*"bala", "mátenlos", "plomo"*) que necesitan el contexto para entenderse como ofensivos en los términos planteados en nuestro trabajo (si un artículo fuese sobre una plaga de osos o langostas no debe-

Tipo	Contexto	Texto
MUJER	Por el mundo: Florencia Peña mostró su lujosa nueva casa con bar, muelle y pileta	@usuario Cuando succionas las correctas
	El exabrupto de Rodolfo Barili con una modelo que se metió en el baúl de un auto para violar la cuarentena	@usuario Barili, ojo es mujer, en cualquier momento te caen las hordas de feminazis, pero estoy al 100 % CON VOS!!
	Video: tres mujeres forzaron la puerta de un edificio en Palermo y desvalijaron dos departamentos	@usuario 3 feministas , claramente es el perfil de las feministas
	Británica llegó a Perú por 10 días y se quedó por amor: vive sin agua potable y entre insectos	@usuario Pero empomada todo el día...
	“Me esterilicé, pero no odio a los niños”: mi vida dentro del movimiento “libre de hijos”	@usuario Es que sos incogible..quien te va a hacer un hijo.....algun ciegoooo
	Mia Khalifa: actuó en videos porno durante pocos meses, cobró fama mundial y ahora lucha por borrar su pasado	@usuario JAJAJAJAAJAJAAJA SEGUILA CHUPANDO....
	Thelma Fardin revictimizada: las víctimas de violación no pueden quedar presas del estigma mientras los violadores siguen libres	Forra anda a laburar a PornHub que la caras te salen bien. Inmunda sos una mentira
		@usuario Cada dia dudo mas si Dartes la violo a la estupida 😕
	Narda Lepes: “Mil veces me tocaron el culo en las cocinas de los restaurantes”	@usuario Tenes lindo culo?
	Un policía se defendió de un robo en Pompeya, mató a un ladrón y estuvo dos días preso	@usuario Que jueza hija de mil puta ojalá t afanen a vos y t maten un ser querido a ver qué haces LLEGUA de mierda puta
	“Cristian, mirá lo que hago por vos”: quién es la modelo que se metió en el baúl de un auto para violar la cuarentena	@usuario Jajaja modelo Hot jaja Telosu!! Mami diría la Moria !! Jaja ridícula tafaldegaver
	La aberrante arenga machista que dio el preparador físico de Güemes de Santiago del Estero	@usuario Que aliade te volviste Clarín, vos te culeas la que tiene olor a pingo (feminazi)
LGBTI	Por qué Flor de la V no continuó en Mujeres de eltrece, tras la salida de Claudia Fontán	@usuario y..porque no es mujer, más claro echale agua
	Histórico: Mara Gómez fue habilitada y será la primera jugadora trans en el fútbol argentino	@usuario Unos huevos bárbaros tiene esta mina!!!!!
	La historia de la modelo colombiana trans que besa la panza de su esposo embarazado de ocho meses	@usuario Un macho besando a otro macho
	Luis Novaresio le dedicó un romántico mensaje a Braulio Bauab por su cumpleaños	@usuario Guacale
	Eugenio Zaffaroni le contestó a Sergio Berni tras la polémica por las domiciliarias: “Es el populacherismo vindicativo que llenó las cárceles”	@usuario cuando se muere este viejo trolo enfermo
	La impactante historia de la tenista trans que hoy es la N° 3 de Argentina en la categoría senior femenino	@usuario Vergonzoso que las mujeres toleren esto.\nEse tenista debería jugar con hombres o a lo sumo, en un torneo de sujetos como él.
	Joe Biden nominó a Rachel Levine, una mujer transgénero, para que sea su subsecretaria de Salud	@usuario Este presidente es la dejeneracion total del mundo
	Así luce el actor Elliot Page tras declararse trans	@usuario Tiene Bija? No. Tiene Concha? Si. Es mujer entonces
	El abuelo que a los 90 años confesó: “Soy gay, soy libre y estoy afuera”	@usuario Como no te agarra el Coronavirus. 🤮 🤮

Tab. 5.9: Ejemplos discriminatorios del dataset contra mujeres y la comunidad LGBTI.

Tipo	Contexto	Texto
RACISMO	Coronavirus: las terribles imágenes del mercado donde se originó la pandemia	@usuario Hay que matarlos hijos de puta
	Malestar en Washington con el Gobierno argentino porque no dejó atracar al buque más moderno de la guardia costera de Estados Unidos	@usuario Amo ver los sudacas que se creen yankis enojados por esto.
	Milagro Sala: "Seguimos presos, los que nos gobiernan tienen que cambiar las cabezas"	@usuario Negra, seguis presa por chorra.
	Al menos 7 muertos en China a causa de un virus transmitido por garrapatas	@usuario Que no venga ningún chino más a la Argentina! Por favor! Ya Basta!
	En China comenzó el tradicional Festival de Carne de Perro a pesar de la pandemia de coronavirus y una ONG intenta salvarlos	@usuario No soy racista, pero hay que matar a todos los chinos
	Científicos identificaron en China otro virus respiratorio "con potencial para convertirse en pandemia"	@usuario Nos infectan a Todos!!!
	Coronavirus. Yanzhong Huang: "Es bastante probable que ya se esté gestando un Covid-21"	@usuario Bombas urgente a esta maldita raza
	Denunciaron la nueva maniobra de China para ocultar las verdaderas cifras del coronavirus	@usuario Mundialmente mantenemos china xq todo viene de ahí y hoy estamos fundidos y en emergencia... #ChinaVirus no quiero ver un #chino x mucho tiempo!
	Villa Mascardi: impresionante operativo con tanquetas blindadas para que una fiscal ingresara a una zona controlada por mapuches	@usuario Basta!!! No son mapuches son delincuentes !!! A ver si alguien pone las pelotas donde hay que ponerlas y los cagan a tiros de una vez !!!
CRIMINAL	Aberrante: un político de Misiones admitió haber esclavizado y violado a sus tres hijastras	@usuario Solo se requiere una bala.
	Rosario: un grupo de vecinos linchó y mató a golpes a un joven acusado de robar autos	@usuario esta perfecto, ejemplo a los demás
	El panadero que mató a un ladrón en La Matanza: "No soy un asesino, estoy arrepentido"	@usuario Que dice señor ! No sé arrepienta, que hizo una obra de bien.Era su vida o la del delincuente.
	Video: salió de la cárcel por el coronavirus y murió de un tiro el mismo día al festejar su libertad	@usuario Justicia divina!!
	Le sacó el arma al ladrón que lo asaltó, lo corrió y lo mató de un disparo: quedó detenido	@usuario Buenísimo vamos por el exterminio total de estos primates.
CLASE	La Justicia ordenó el desalojo de la masiva toma de terrenos en Guernica	@usuario Lanzallamas y a otra cosa
	Hubo tensión en la Quinta de Olivos entre un grupo que apoyaba a Alberto Fernández y manifestantes del banderazo contra el Gobierno	@usuario PLANEROS Y BARRABRAVAS
	Organizaciones sociales cortaron la avenida 9 de Julio: reclamaron un salario mínimo de $ 45.000	@usuario Vayan a lo laburar hdp.
	La historia de una familia de cartoneros en la toma de Guernica: "Por primera vez sentimos que tenemos un hogar"	@usuario Bala.
	El Gobierno autorizó la apertura de las escuelas porteñas para las elecciones de Bolivia	@usuario No sería mejor deportar a los bolivianos indocumentados?.además nos suman pobreza e indigencia
	Coronavirus en Argentina: un dirigente radical deseó que la pandemia "haga una limpieza étnica" con "negros de La Matanza"	@usuario Es el deseo de todo argentino de bien
	El Polo Obrero realiza un corte en la Panamericana en contra de la flexibilización de la cuarentena y en reclamo de aumentos a los planes sociales	@usuario Clarísimo que no quieren laburar y quieren vivir de nosotros!
	Coronavirus en la Argentina: movimientos sociales reclaman asistencia alimentaria en el Obelisco	@usuario Anda a laburar lpqtp

Tab. 5.10: Ejemplos discriminatorios del dataset por motivos de clase, racismo, o contra criminales.

Tipo	Contexto	Texto
POLITICA	Confirman una mutación en el coronavirus que puede hacerlo 10 veces más contagioso que la cepa original de Wuhan	@usuario ME ALEGRO MUCHÍSIMO.\nOJALÁ LLEGUE PRONTO A ARGENTINA Y ARRASE CON TODO.\nPODRÍAMOS VER AL FIN ALGO MÁS DAÑINO QUE EL CÁNCER PERONISTA Y SU METÁSTASIS KIRCHNERISTA.
	Murió un nieto recuperado por Abuelas de Plaza de Mayo: los mensajes de Alberto Fernández y Cristina Kirchner	@usuario Un planero menos.
	Última encuesta: ¿Qué mujer superó a Alberto Fernández en imagen positiva?	@usuario Les ahorro el clickbait. Es Vidal, igual perdió por 20 puntos. Gorila LTA.
	Cómo es la cerveza "peronista" que el Chacho Coudet le regaló a Alberto Fernández	@usuario Debe ser meo de gato. berreta como todo lo peroncho
	El descargo de Nicolás Wiñazki después de que Vero Lozano se burlara de él: "Quizás le afecta la cuarentena"	@usuario Yo creo que al revés, patético operador. Solo los gorilas pueden bancarte croto
APARIENC.	Axel Kicillof recomendó una "cuarentena previa" de 14 días para "llegar sanos a Navidad y Año Nuevo"	@usuario Chuoame la verga enano moishe
	Video indignante: piba violó la cuarentena y viajó en el baúl de un taxi para ver a un chico	@usuario Habría qur buscar también y meter en cana al cirujano que le hizo la nariz!! Parece Michael Jackson la loca!!!
	El senador José Mayans defendió a Gildo Insfrán: "En pandemia no hay derechos"	@usuario Volve al gancho , docer
	El video sexy de More Rial en corpiño	@usuario Asco
	El sensual paseo en moto de Florencia Peña: "Próxima parada: tu casa"	@usuario Que tiene de sensual, ésta vieja cascoteada?prostituta de cuarta,kukaracha inmunda!
DISCAPAC.	Patricia Bullrich pidió ser drásticos con los docentes: "El que no va, tendrá que ser reemplazado"	@usuario Estragos del tinto
	El abuelo que a los 90 años confesó: "Soy gay, soy libre y estoy afuera"	@usuario Alhzeimer o demencia senil!!!!!
	Elisa Carrió dijo que "ninguna pandemia es excusa para suspender la República" y advirtió que "vienen por los campos"	@usuario Si sacan a la paciente psquiatrica es porque están hasta las manos.\n#CarcelACambiemos
	Florencia Kirchner y su posteo a favor de la amistad: "Nunca entendí la desesperación por la pareja"	@usuario La enfermita está mejor que yo,y no se calienta por la hija
INCITACIÓN	Harán un listado de los presos en situación de riesgo por el coronavirus para evaluar si deben salir de prisión	@usuario @usuario Todo al reves! Si hay alguno con coronavirus PONGANLO EN EL MEDIO!
	La advertencia de Juan Grabois: "Van a haber 1, 5, 20 Guernicas"	@usuario @usuario Habra 100 paredones
	Otro caso de peste bubónica enciende las alarmas en China	@usuario Una atómica a China...
	Coronavirus: afirman que volvió la venta de carne de murciélagos en China	@usuario Boicot a todo producto chino!!!
	Villa Mascardi: impresionante operativo con tanquetas blindadas para que una fiscal ingresara a una zona controlada por mapuches	@usuario El Diálogo se Inicia con BALAS y Finaliza con La Última
	Coronavirus en China: la ciudad de Shenzhen prohíbe comer perros y gatos	@usuario habrá alguna manera de erradicar a estos tipos del mundo ?

Tab. 5.11: Ejemplos discriminatorios del dataset. INCITACIÓN refiere a los llamados a realizar algún tipo de medida contra el grupo o la persona atacada.

ríamos considerarlos como tal). Por otro lado, algunos mensajes enumerados son más difíciles de descifrar y dependientes del contexto, como las celebraciones ante el abatimiento de un preso o criminal ("bravo", "felicitaciones!") que parecen inofensivas hasta que se lee el contexto de la noticia. Algo a remarcar de este tipo de comentarios es que tienen una polaridad positiva y contenido altamente irónico, este último punto indescifrable sólo observando el texto del tweet.

En el caso de RACISMO (la característica más marcada del conjunto de datos) hay una fuerte cantidad de comentarios discriminatorios contra la comunidad china. Estos mensajes son compatibles con el brote racista que tuvo lugar durante la pandemia del COVID-19, algo que tuvo su replica en las redes sociales y que ha sido ya marcado por He et al. [71]. Muchos de estos tweets con contenido discriminatorio incitan a la acción, algunos con llamados a tomar medidas "blandas", (*"no ir a comprarles a los supermercados"*) y otros directamente alentando al exterminio de este pueblo.

Las características listadas en la Tabla 5.11 (POLITICA, DISCAPACIDAD, APARIENCIA) poseen características más elementales y menos desafiantes, basadas en agravios directos y explícitos. A priori, uno podría pensar que son las características que menos necesidad de contexto revisten, ya que –mayormente– su carga de odio es notoria y centrada en insultos. Algunos de los ejemplos de dicha tabla ilustran técnicas de camuflaje (*tafaldegaver*, falta de verga, *docer*, cerdo) que dificultan su detección.

5.8. Discusión

De las tres etapas en las que separamos la tarea de la construcción del conjunto de datos, la recolección fue la única que no presentó decisiones complejas. La posterior etapa de selección, por el contrario, nos planteó algunos obstáculos no menores teniendo en cuenta que el discurso de odio no está distribuido uniformemente entre los distintos artículos periodísticos. Exploramos distintas alternativas para poder escoger artículos y comentarios a etiquetar, tanto observando el texto de los artículos como sus comentarios. Decidimos seleccionar los artículos en base a sus respuestas potencialmente discriminatorias usando un lexicón de expresiones, luego de evaluaciones subjetivas que resultaron en una mejor calidad de artículos seleccionados en base a este método. Utilizamos el lexicón no para marcar los comentarios a etiquetar, sino los artículos: los comentarios a etiquetar fueron elegidos –ahora sí– de manera aleatoria entre los artículos ya seleccionados. Trabajo futuro podría explorar alternativas para esta selección, como por ejemplo utilizar las conexiones de amistad en Twitter entre los usuarios comentaristas.

Para realizar la tarea de etiquetado, definimos un modelo de anotación jerárquico y granular de acuerdo a lo discutido en la Sección 4.7. El hecho de anotar las características –y no sólo la etiqueta binaria de presencia de discurso de odio– es algo que pocos trabajos previos han explorado. Seis etiquetadores nativos de la variedad dialectal rioplatense realizaron la tarea bajo un esquema de dos anotaciones y desempate. Como producto, obtuvimos cerca de 57,000 comentarios repartidos en 1,238 artículos, una cantidad de tamaño considerable en términos de comentarios aunque no tengamos parámetro de comparación ya que no existen muchos conjuntos de da-

tos similares. De los comentarios, alrededor de 8,000 comentarios tienen contenido discriminatorio, obteniendo una tasa aproximada de un comentario discriminatorio cada seis.

Un análisis exploratorio de los comentarios discriminatorios muestra ejemplos complejos y ricos, algunos de ellos altamente dependientes del contexto. Finalmente, un análisis de la co-ocurrencia de las características ofendidas da muestra de que el conjunto de datos anotado posee diversidad en sus instancias, con múltiples tipos de discriminación y artículos que poseen comentarios odiosos de diversa naturaleza. Podemos especular que tanto el texto (el comentario en sí) como el contexto (el tweet del medio periodístico y su artículo periodístico) contienen información valiosa para poder distinguir entre las distintas categorías discriminatorias.

5.9. Conclusión

En este capítulo hemos desarrollado el proceso de construcción de un conjunto de datos contextualizado de discurso de odio en redes sociales. Para ello, recolectamos respuestas de usuarios a noticias periodísticas posteadas en Twitter por los principales medios de noticias de Argentina. Describimos detalladamente el proceso de su construcción –tanto en la recolección, selección y anotación de los datos– haciendo eje en las distintas dificultades que fuimos encontrando y posibilidades de mejora.

Como resultado, obtuvimos más de 8,000 comentarios discriminatorios anotados de manera granular de acuerdo a las diferentes características ofendidas. Mediante evaluaciones subjetivas y análisis de las co-ocurrencias de las características, podemos afirmar que este conjunto de datos posee comentarios con notable complejidad, discurso de odio explícito e implícito, y artículos que suscitan distintos tipos de reacciones discriminatorias, lo cual aporta a la riqueza de los datos.

Con este conjunto de datos como insumo, pasamos ahora a analizar un punto que discutimos en Sección 4.7: la contextualización de los mensajes para la detección de discurso de odio. Este tema ha sido poco abordado en la literatura y es por ello que consideramos importante estudiar el impacto de poseer esta información adicional.

5.10. Notas

En el Apéndice B se encuentra el manual de etiquetado como así información adicional sobre la construcción del conjunto de datos. La herramienta de etiquetado puede encontrarse en `https://github.com/finiteautomata/news-labelling`.

6. EXPERIMENTOS DE DETECCIÓN CONTEXTUALIZADA DE DISCURSO DE ODIO

En este capítulo analizamos el impacto de añadir contexto en la tarea de detección de discurso de odio en redes sociales. Como hemos marcado en los capítulos anteriores, la utilización del contexto ha recibido poca atención en la literatura, limitando la tarea a analizar comentarios aislados de cualquier tópico relacionado o hilo conversacional. Para este estudio, utilizamos el conjunto de datos construido en el Capítulo 5, cuyos datos constan de comentarios de artículos periodísticos en Twitter.

El formato de los datos empleados nos brinda información adicional a cada comentario tanto por el tweet del medio periodístico al que contestan como así también por el contenido del artículo. Para evaluar si la adición de contexto resulta en una mejora en la detección de discurso de odio, realizamos experimentos de clasificación con modelos que consumen tres tipos de entrada: el comentario sin contexto, el comentario junto al tweet del medio periodístico, y el comentario junto al tweet y el cuerpo del artículo asociado.

El conjunto de datos empleado nos permite analizar una posible combinación más en base al detalle de las características ofendidas por cada comentario. Esta información granular permite no sólo analizar la existencia de discurso de odio sino que permite predecir con más detalle la ofensa cometida. Proponemos en base a esto dos tareas de clasificación: una tarea de detección **binaria**, donde sólo predecimos si hay o no discurso de odio; y una tarea de detección **granular**, donde además predecimos todas las características ofendidas (potencialmente más de una). Para estas tareas, propusimos algoritmos de clasificación sobre modelos pre-entrenados de lenguaje que tienen como entrada los distintos tipos de contexto posibles. Estos modelos tienen incorporados naturalmente la posibilidad de consumir dos entradas –el contexto y el texto– con lo cual son ideales para nuestros experimentos.

Evaluamos los resultados de los experimentos de clasificación tanto en términos del rendimiento de las distintas configuraciones de nuestros clasificadores, como así también realizando análisis de error comparativos entre los modelos contextualizados y los no contextualizados. También evaluamos en este capítulo las dificultades más generales que presenta la detección de este fenómeno sobre comentarios de notas periodísticas.

6.1. Trabajo previo

Como mencionamos en la Sección 5.1, no se ha dado demasiada atención en la literatura a la utilización de información contextual en la detección de discurso de odio y otros fenómenos similares. Pasamos ahora a repasar los algoritmos de detección utilizados sobre los conjuntos de datos descriptos en dicha sección.

Gao and Huang [56] proponen utilizar dos tipos de modelos sobre el dataset que ellos mismos recolectaron sobre comentarios de Fox News: regresiones logísticas y redes neuronales recurrentes. Para las regresiones logísticas, usaron como entradas

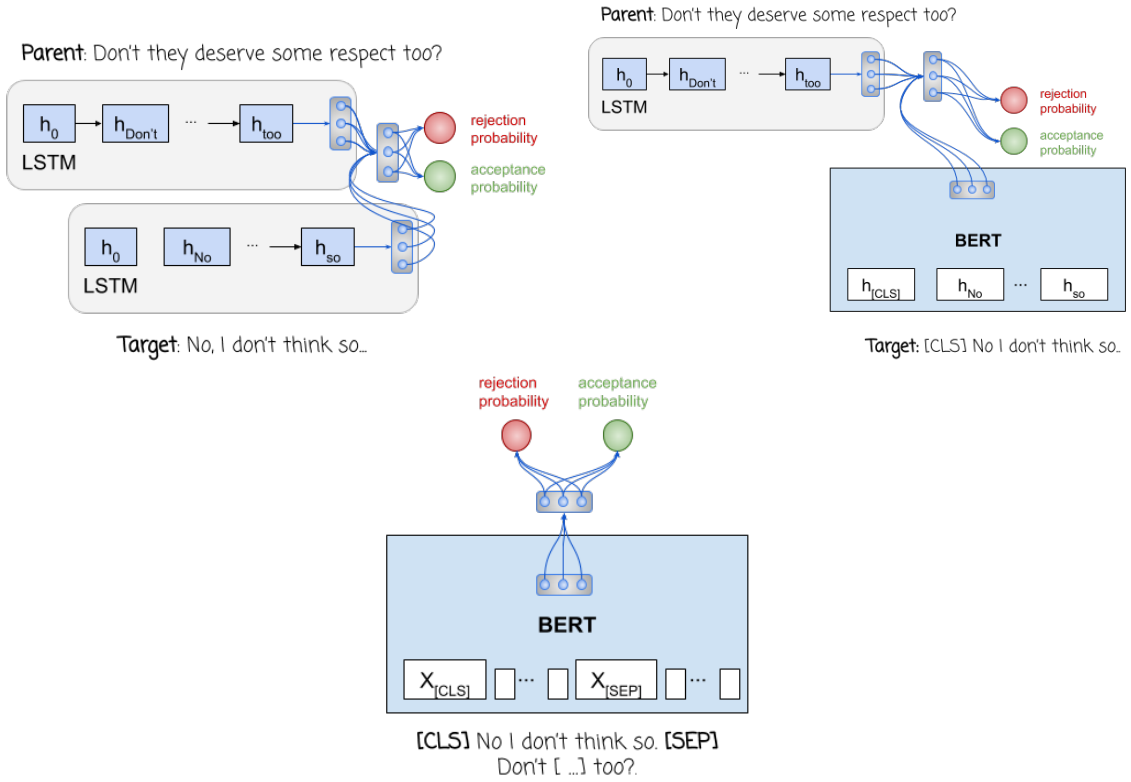


Fig. 6.1: Clasificadores que consumen contexto propuestos por Pavlopoulos et al. [121]. Los dos primeros clasificadores proponen una arquitectura de dos encoders, uno para el texto y otro para el contexto usando bi-LSTMs y BERT como posibilidades. El tercer clasificador propuesto es un BERT usando su estructura natural para codificar dos oraciones separadas por el token *SEP*

bolsas de palabras, bolsas de caracteres, vectores semánticos producidos con Linguistic Inquiry and Word Count (LIWC) [123], y otras variables de un lexicón de emociones [111]. Por otro lado, los autores también entrenan LSTM bidireccionales con mecanismo de atención de Bahdanau [9] que consumen embeddings *word2vec* de dimensión 100.

Un punto criticable de este trabajo es que utiliza el nombre de usuario como entrada, algo que a priori no suele hacerse ya que permitiría prejuzgar a un usuario antes que por el contenido de sus tweets. Si bien es cierto que la información de usuarios y sus conexiones es valiosa, introducir esta información a nuestros modelos puede dar lugar a correlaciones espurias que es preferible evitar. Otras críticas sobre el proceso de anotación de los datos fueron realizadas ya en la Sección 5.1.

En la Sección 5.1 hemos descripto el conjunto de datos construido por Pavlopoulos et al. [121], dedicado a la detección de toxicidad y que incorpora información conversacional sobre comentarios de Wikipedia Talk Pages. Nos detenemos un momento para analizar sus experimentos de clasificación ya que guardan importantes similaridades con lo hecho en este capítulo. En ese trabajo se obtuvieron dos conjuntos de entrenamiento: uno en el cual los etiquetadores tenían información del contexto, y otro conjunto en el que no. El conjunto de test, por otro lado, fue ano-

tado teniendo en cuenta el contexto bajo la asunción de que el etiquetado es de mejor calidad al tener más información contextual. Sobre la base de estos datos, los autores plantearon dos preguntas:

- ¿Mejora el rendimiento de los clasificadores que son entrenados con el conjunto de datos etiquetado con contexto?
- ¿Mejora el rendimiento de los clasificadores consumiendo información contextual?

Para responder estas preguntas, los autores consideraron las siguientes combinaciones para sus experimentos: utilizar conjunto de entrenamiento etiquetado con o sin contexto, y entrenar el clasificador con o sin contexto. Para aquellos clasificadores que no consumen contexto, los autores consideraron las mismas alternativas que hemos visto en capítulos anteriores: bi-LSTM o *BERT*. Para aquellos que sí consumen contexto, se evaluaron dos estrategias: la primera consistió en usar una única red que codifique la entrada del texto y el contexto concatenada con un token especial; la segunda consistió en usar dos codificadores distintos para el contexto y el texto. Para la segunda alternativa, y dados los recursos computacionales disponibles, no utilizaron dos modelos pre-entrenados, sino un codificador LSTM para el contexto y un *BERT* para el texto. A su vez, también utilizaron la API Perspective de Google con la misma estrategia de concatenación. Para todas las combinaciones posibles, la mejora en el rendimiento resultante de disponer de información contextual no es estadísticamente significativa.

Dos versiones de *BERT* fueron utilizadas como base para entrenar los modelos de Transformers: una, usando los pesos del modelo de *BERT* de Devlin et al. [45]; y la segunda, haciendo un ajuste de dominio de *BERT* sobre un dataset grande y no etiquetado relacionado a la tarea en cuestión. Este proceso de *ajuste de dominio* o *fine-tuning* consiste en ajustar el modelo de lenguaje sobre un conjunto de datos no etiquetados y afines a nuestra tarea final. Esta técnica ha demostrado ser efectiva para lograr mejoras sensibles en el desempeño de clasificadores sobre dominios particulares [68], y será estudiada más detenidamente en el Capítulo 7. Para este trabajo, el ajuste es realizado sobre un subconjunto de comentarios del dataset de *Civil Comments* [25] sin ningún tipo de contexto. A priori, ajustar el modelo de lenguaje sobre comentarios a secas podría inducir a pensar que puede deteriorar el rendimiento al entrenar posteriormente sobre contexto; sin embargo, en la versión no adaptada de *BERT* tampoco se observó una mejora significativa en el rendimiento.

Algunas limitaciones de este trabajo marcadas por los autores son:

- Contexto muy pequeño: sólo se consideraron como contexto el título de la discusión de Wikipedia Talk Pages y adicionalmente el comentario previo.
- El hilo completo de los comentarios es ignorado: sólo se observa el comentario previo.
- Los comentarios fueron muestreados aleatoriamente, sin tener en cuenta algunos ámbitos más propicios para la toxicidad.

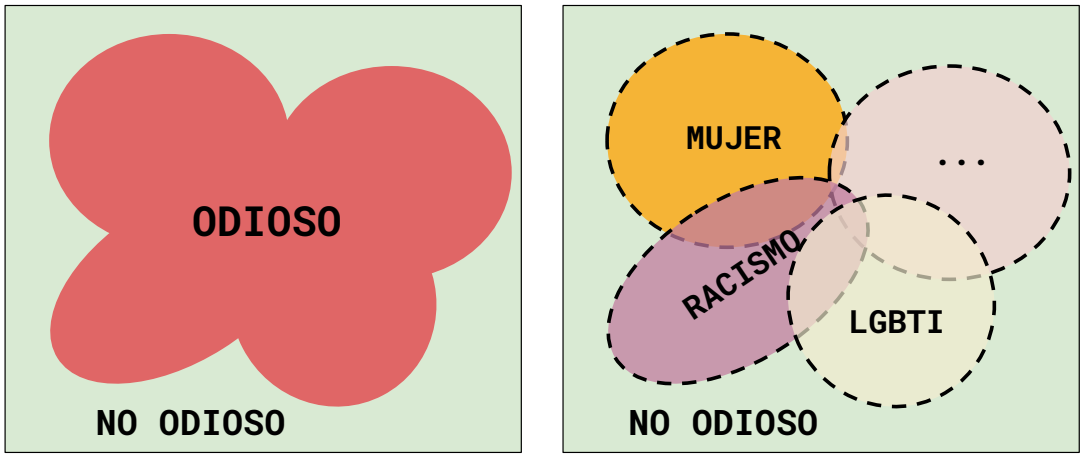


Fig. 6.2: Tareas propuestas de detección de discurso de odio. La tarea de detección binaria consta de predecir si un tweet contiene contenido discriminatorio, discriminando la frontera conjunta de todas las características. En la tarea granular, predecimos por separado cada una de las características ofendidas, pudiendo haber más de una o bien ninguna.

Xenos et al. [170] continuaron el trabajo de Pavlopoulos et al. [121] reetiquetando el conjunto de datos de Civil Comments con información contextual y –como mencionamos en la Sección 5.1– presentando una nueva tarea de detección de sensibilidad al contexto. Usando la API Perspective (y la estrategia de concatenación básica de texto y contexto) notaron que la performance del clasificador contextualizado mejora sensiblemente si restringimos nuestra atención a comentarios más sensibles a su entorno de acuerdo a la métrica definida. De esto último puede concluirse que, a diferencia del trabajo anterior, si bien el contexto no es realmente necesario para comprender la toxicidad del grueso de los comentarios, hay cierto subconjunto para los cuales esta información adicional resulta relevante.

6.2. Tareas de clasificación propuestas

Para analizar el impacto del contexto en la detección de discurso de odio, y teniendo en cuenta que contamos de un dataset con anotaciones granulares sobre las características ofendidas, propusimos dos tareas de clasificación:

1. **Detección binaria**: Dado un tweet y su contexto, predecir si es discriminatorio.
2. **Detección granular**: Dado un tweet y su contexto, predecir las características ofendidas (de haber alguna) y si contiene un llamado a la acción.

Puede pensarse la tarea de detección binaria (la que usualmente se aborda en la literatura sobre el tema) como una relajación de la tarea granular: mientras la

primera sólo nos permite detectar si hay o no contenido discriminatorio, la segunda requiere información más precisa acerca de las características ofendidas, permitiendo a su vez tener mayor información sobre la salida de los clasificadores y dando lugar a una mejor interpretación de sus errores. La Figura 6.2 ilustra las dos tareas propuestas en forma de Diagrama de Venn: mientras en la tarea binaria sólo debemos decidir de qué lado de la frontera se encuentra un comentario (si tiene o no discurso de odio) en la tarea granular se debe decidir esto mismo para cada una de las características,

Viendo las tareas propuestas como problemas de clasificación, la detección binaria consta de predecir una sola etiqueta binaria, mientras que la tarea granular consta de predecir n etiquetas binarias. Esto último puede también verse como n problemas distintos de clasificación, o una tarea de **multiclasificación**. Una observación sobre estos dos enfoques del problema es que podemos construir un clasificador para la tarea binaria a través de un clasificador entrenado para la tarea granular tomando la disyunción lógica de sus salidas: hay discurso de odio sí y sólo sí hay al menos una característica ofendida. Retomamos esta idea más adelante al hablar de cómo evaluamos nuestras técnicas de clasificación para cada tarea.

Partición	Artículos	Comentarios
Entrenamiento	990	36,420
Desarrollo		9,106
Test	248	11,343

Tab. 6.1: Particiones del conjunto de datos utilizado para las dos tareas

Para las dos tareas utilizamos el conjunto de datos construido en el Capítulo 5 separando las instancias en tres particiones: entrenamiento, desarrollo, y test. Para las dos primeras (entrenamiento y desarrollo) reservamos 990 artículos mientras que para el dataset de test tomamos 248 artículos y sus comentarios. Ambos conjuntos de noticias son disjuntos para intentar maximizar las instancias realmente diferentes para las etapas de entrenamiento y evaluación. Sobre los primeros 990 artículos, dividimos el conjunto en 36,420 instancias de entrenamiento y 9,106 instancias de desarrollo, sin garantizar que provengan de notas periodísticas distintas.

6.3. Modelos de clasificación

Entrenamos clasificadores neuronales basados en el modelo pre-entrenado *BETO* [31] tanto para la tarea binaria como para la granular. Incorporamos la información contextual en cada comentario teniendo en cuenta tres tipos de entrada por instancia: el comentario sin ningún tipo de contexto (notaremos **sin contexto**), el comentario con el tweet al que responde como contexto (**tweet**), y finalmente el comentario con el tweet al que responde y el texto del artículo periodístico (**tweet + artículo**). Para las dos versiones que consumen información contextual, separamos el texto y el contexto con el token especial `[SEP]`.

Para la tarea binaria, la salida es la estándar para un clasificador binario. En cuanto a la tarea de detección granular, la abordamos de manera similar a la **Ta-**

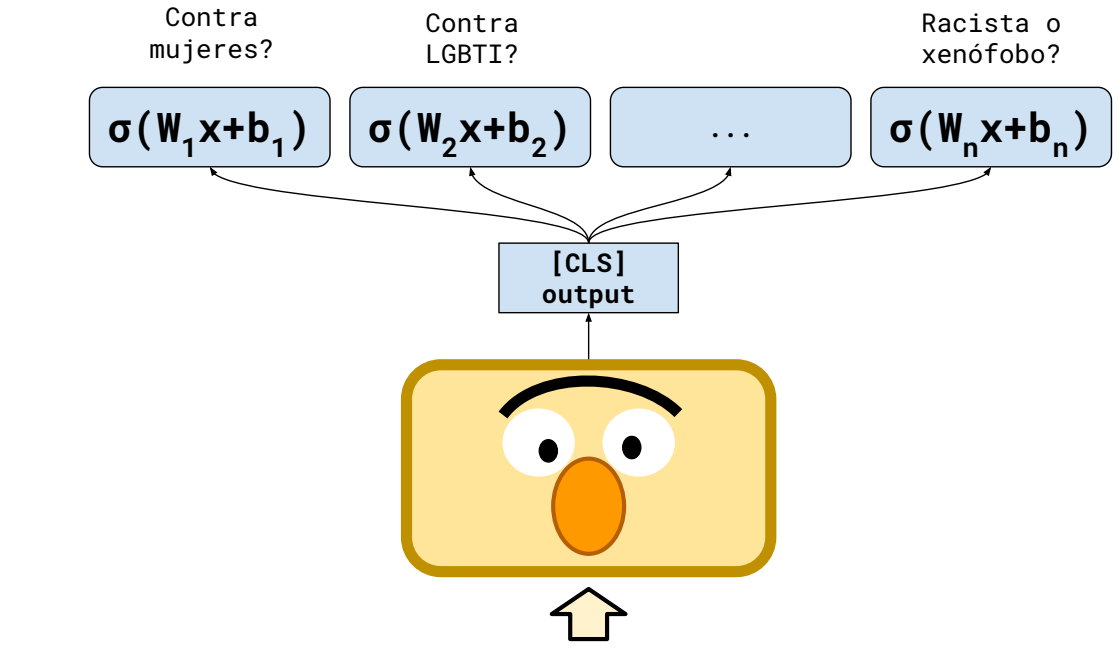


Fig. 6.3: Modelo de multiclasificación para la tarea granular basado en *BETO*. Los modelos son entrenados de tres maneras distintas: sin contexto (sólo el comentario), con el contexto del tweet, y con el contexto del tweet y el texto del artículo. La salida consta de la probabilidad de que el comentario tenga contenido odioso para alguna de las características en cuestión o bien contenga una llamada a la acción

rea B del Capítulo 4, en este caso como la predicción de nueve variables distintas: llamado a la acción (LLAMA) y las ocho características ofendidas (MUJER, RACISMO, CLASE, LGBTI, CRIMINAL, ASPECTO, DISCAPACIDAD, POLITICA). En lugar de entrenar un clasificador diferente por cada característica, entrenamos un modelo de multiclasificación BERT que comparte todos sus pesos salvo nueve capas lineales de salida distintas.

Como función de costo, empleamos la suma de las funciones de costo de cada característica. Concretamente, si y son las etiquetas de un comentario e $\widehat{y}$ las predicciones del modelo:

$$L(y, \widehat{y}) = \sum_{c \in CHAR'} J(y_c, \widehat{y}_c)$$

donde $CHAR'$ es el conjunto de todas las características protegidas junto a la variable de llamada a la acción (LLAMA), y J es la función de entropía cruzada. Compartir los pesos entre todas las salidas tiene dos objetivos: primero, poder generar un modelo más compacto (de otra forma serían nueve BERT distintos que suman alrededor de 1,000M parámetros) y segundo, compartir información común entre las distintas características atacadas, ya que guardan similaridades y muchas de ellas tienen una importante intersección, como hemos visto en la Sección 5.7.2.

Para tener costos computacionales más amigables, limitamos los largos de las secuencias a 128, 256 y 512 tokens para los modelos con entrada sin contexto, tweet, y tweet+cuerpo respectivamente. Preprocesamos ambos tweets –contexto y texto– utilizando las técnicas descriptas en la Sección 3.5: conversión de usernames a un token especial (`usuario`), tratamiento de hashtags (separación e inserción de un hashtag especial), y conversión de emojis a su representación textual. La Figura 6.3 ilustra el modelo de clasificación para la tarea granular, junto a los tres tipos de entrada considerados. La configuración de hiperparámetros utilizados en el entrenamiento es la misma que la detallada en la Sección 3.6.

6.3.1. Adaptación de dominio

Una práctica cada vez más extendida en trabajos del área de clasificación de documentos es realizar una adaptación de dominio para mejorar la performance sobre la tarea final. La técnica de adaptación de dominio para modelos de lenguaje pre-entrenados consiste en continuar el pre-entrenamiento sobre un dataset grande y no supervisado relacionado a nuestro dominio o directamente sobre el dataset de la tarea si no tenemos acceso a otros datos [68]. En la Sección 7.1 del siguiente capítulo haremos una reseña más extensa de esta técnica, pero por lo pronto podemos entender que esta técnica ajusta el modelo de lenguaje a nuestros datos, ya que estos pueden tener diferencias considerables respecto de los empleados en el pre-entrenamiento. En nuestro caso puntual, mientras *BETO* fue entrenado en Wikipedia y textos formales, nuestro dominio consta de comentarios en Twitter a notas periodísticas, con expresiones muy distintas a las encontradas en medios más formales.

Pavlopoulos et al. [121] realizaron una adaptación de dominio sobre los comentarios del corpus de *Civil Comments* sin utilizar ningún tipo de contexto. Proponemos,

Hiperparámetro	Valor
Pasos	10,000
Tamaño de batch	2,048
Tamaño de secuencia	128, 256 y 512
β_1	0,9
β_2	0,98
ϵ	10^{-6}
Decay	0,01
LR pico	0,0004
Pasos de warmup	0,1

Tab. 6.2: Hiperparámetros para la adaptación de dominio de BERT

a diferencia de este trabajo, tres tipos de adaptaciones de acuerdo al contexto considerado: una adaptación sin contexto, una adaptación con el contexto del tweet, y una adaptación con el contexto del tweet y el cuerpo de la noticia. Los datos usados para este ajuste fueron el sobrante de la recolección del anterior capítulo: alrededor de 288 mil artículos con 5 millones de comentarios que no están incluidos en el conjunto de datos etiquetado [1].

Liu et al. [95] recomienda descartar en el pre-entrenamiento de modelos de lenguajes el objetivo NSP, con lo cual realizamos nuestro ajuste de dominio exclusivamente con la tarea de modelado de lenguaje enmascarado (MLM). La Tabla 6.2 contiene los hiperparámetros utilizados al correr la adaptación de dominio correspondientes al optimizador Adam con warmup lineal para el learning rate. Estos ajustes los realizamos sobre una *TPU v2-8* y una máquina de *Google Colab Pro*, tomando alrededor de 10 hs en su largo de cadena máximo.

6.3.2. Rendimiento humano en la tarea

La tarea de detección de lenguaje discriminatorio contiene una alta cantidad de ruido y un bajo acuerdo entre humanos, como se atestigua en recientes revisiones de los conjuntos de datos generados para su clasificación [132]. En este escenario, cabe preguntarse una cota al rendimiento que puede lograr un algoritmo de detección automática para esta tarea, algo que esperamos –por la misma naturaleza del problema– que diste mucho de la perfección. Si bien en muchos trabajos se ha logrado que algoritmos basados en Transformers superen el rendimiento humano para distintas tareas de NLP [2], consideramos el desempeño de nuestros anotadores como una cota superior razonable para las técnicas automáticas que hemos propuesto.

Para obtener números indicativos del rendimiento humano en la detección de discurso de odio, utilizamos la información recolectada durante la anotación del conjunto de datos. Consideramos las anotaciones de cada uno de los seis participantes en el proceso como predicciones y las evaluamos de dos formas: la primera tomando como

[1] Utilizamos algunos datos extra recolectados a posteriori de lo mencionado en el capítulo anterior

[2] Algo discutible dado que el rendimiento humano para estos benchmarks –como GLUE– está medido a través de trabajadores contratados a través de crowdsourcing con pagas por debajo del salario mínimo

	Entre anotadores		Contra etiquetas	
	F1 media	F1 mediana	F1 media	F1 mediana
ODIO	65,3	67,5	82,9	85,1
LLAMA	43,4	49,5	70,4	84,2
MUJER	49,0	46,8	74,1	75,9
LGBTI	59,6	57,7	84,6	91,5
RACISMO	65,3	64,4	87,1	87,9
CLASE	44,3	44,4	72,2	73,2
POLITICA	46,1	43,6	79,5	81,5
DISCAPACIDAD	55,0	60,0	81,3	84,2
APARIENCIA	64,9	74,3	83,1	91,5
CRIMINAL	52,7	58,0	84,1	92,9
Macro F1	53,4	55,4	79,6	84,8

Tab. 6.3: Estadísticos del F1 (en porcentaje) entre anotadores. Las dos primeras columnas marcan las métricas medidas entre anotadores, y las dos últimas la de los anotadores contra las etiquetas asignadas en el conjunto de datos. La Macro F1 es el promedio de los F1 todas las características y de la llamada a la acción (LLAMA).

etiquetas doradas las anotaciones de otro anotador; la segunda, tomando las etiquetas generadas en la Sección 5.6.6. En el primer caso tenemos $\binom{6}{2} = 15$ combinaciones, mientras que en la segunda tenemos una para cada anotador, 6 combinaciones. Para estas dos formas de calcular rendimientos humanos, calculamos la medida *F1* entre las predicciones de cada anotador y las etiquetas doradas, y calculamos medias y medianas para obtener estimaciones puntuales de los rendimientos.

Algo a tener en cuenta en la comparación contra las etiquetas del conjunto de datos es que este *gold standard* es construido mediante votación mayoritaria de los dos o tres anotadores empleados, por lo cual estamos calculando la métrica entre dos variables correlacionadas. Mientras esta cota por un lado es muy grosera, por el otro, las métricas calculadas entre anotadores pueden ser algo bajas debido a que son predicciones muy ruidosas a diferencia de las etiquetas doradas que son más robustas al estar generadas por varios usuarios. El mejor escenario para estimar el rendimiento hubiera sido contar con una anotación extra para cada instancia y compararla contra la etiqueta dorada, aunque esta metodología es poco eficiente en términos de recursos.

La Tabla 6.3 contiene las medias y medianas del puntaje F1 tanto entre anotadores como contra el *gold-standard*. La mediana entre anotadores de la F1 es 67,5, un puntaje relativamente bajo para la detección de odio, mientras que contra el gold standard es de 85,1 puntos de F1; podemos suponer que la performance humana para la tarea se encuentra en algún lugar entre esos dos números. Respecto a las características, el rendimiento para su reconocimiento entre humanos en algunas de ellas se mantiene muy bajo, particularmente en MUJER, CLASE y POLITICA. Esto se desprende de las observaciones y dificultades descriptas en el Capítulo 5 durante el proceso de anotación, como así también del hecho que estas características tienen un solapamiento no menor entre sí.

Métrica	Sin Contexto		Tweet		Tweet + Cuerpo	
	¬FT	FT	¬FT	FT	¬FT	FT
Accuracy	88,9	89,9	90,2	**91,0**	90,4	90,5
Precisión	67,8	71,8	73,1	**74,8**	73,9	72,8
Recall	56,8	60,2	60,1	**65,3**	61,1	64,1
F1	61,8	65,5	66,0	**69,7**	66,9	68,1
Macro F1	77,6	79,8	80,1	**82,2**	80,6	81,3

Tab. 6.4: Resultados de los experimentos de clasificación para la tarea *binaria* de detección de discurso de odio, expresados como la media de las distintas métricas sobre diez corridas independientes. En negrita, los mejores resultados. Cada modelo es un BERT con tres posibles entradas: sólo el comentario (*Sin contexto*), el tweet de la noticia a la cual responde el comentario (*Tweet*), y el tweet más el cuerpo de la noticia (*Tweet + Cuerpo*). Para cada una de estas posibilidades usamos dos versiones: una sobre BETO (¬FT) y otra sobre BETO ajustado al dominio (FT).

6.4. Resultados

La Tabla 6.4 contiene los resultados para la tarea de clasificación **binaria** medidos por Accuracy, Precision, Recall, F1 de la clase positiva y Macro F1 entre las dos clases. Las métricas están expresadas como las medias de diez corridas independientes de los experimentos. Las seis columnas corresponden a la combinación de los tres posibles modelos dependiendo del contexto utilizado y de acuerdo a si ajustamos al dominio o no. Podemos observar que, en todos los casos, la adaptación de dominio (las columnas marcadas con FT) obtienen mejor rendimiento que los modelos que no están adaptados (¬FT) resultando en una mejora de alrededor de 4 puntos de F1 en los casos sin contexto y con contexto de tweet. Entre los modelos sin ajustar a dominio, el que consume el contexto completo (tweet + cuerpo de la noticia) obtiene el mejor desempeño; sin embargo, esto no se replica en el caso ajustado a dominio, donde gana el contexto simple. Viendo sólo las columnas marcadas como FT, la mejora contra el modelo que no consume contexto es de 4,2 puntos de F1. El modelo con el contexto completo, si bien mejora la performance general contra no tener contexto, pierde precisión al ser adaptado al dominio.

La Tabla 6.5 muestra los resultados de los experimentos de clasificación para la tarea de detección **granular** medidos en puntos de F1 para cada una de las características, y las métricas promediadas de precision, recall, y F1. Como era esperable, la ganancia de tener contexto disponible es más evidente en esta tarea, observándose una diferencia de aproximadamente 6 puntos de F1 entre la mejor versión sin contexto y la mejor versión con contexto (55,1 Macro F1 de la versión FT sin contexto vs 61,3 F1 de la versión FT con el contexto del tweet). Respecto a los dos tipos de contexto, de nuevo la versión simple obtiene mejor rendimiento en prácticamente todas las características, con la única excepción de POLITICA.

La Figura 6.4 muestra los resultados por característica ordenados de mayor a menor según la diferencia de rendimiento entre los clasificadores ajustados a dominio que consumen contexto y aquellos que no lo hacen. Para todas las características se observa una mejora estadísticamente significativa al correr un test Mann-Whitney U ($p \leq 0{,}005$, p valores ajustados por múltiples comparaciones con Benjamini-

Métrica	Sin Contexto		Tweet		Tweet + Cuerpo	
	¬FT	FT	¬FT	FT	¬ FT	FT
LLAMA	64,6	65,1	63,8	**68,5**	65,3	68,0
MUJER	37,3	38,9	41,1	**42,1**	38,1	**42,1**
LGBTI	35,1	36,6	45,1	**48,2**	42,7	44,5
RACISMO	63,5	65,3	68,8	**72,0**	69,1	71,1
CLASE	40,1	43,3	49,1	**51,1**	45,1	47,6
POLITICA	55,5	61,1	57,9	62,5	59,1	**64,8**
DISCAPAC	55,1	58,2	58,5	**60,9**	55,7	57,8
APARIENCIA	72,6	74,2	74,1	**76,6**	75,5	75,8
CRIMINAL	51,3	52,9	65,0	**69,9**	65,4	66,8
Macro F1	52,8	55,1	58,2	**61,3**	57,3	59,8
Macro Precision	55,8	63,0	64,2	**70,2**	67,7	67,8
Macro Recall	50,6	49,9	54,0	**55,1**	50,4	54,1

Tab. 6.5: Resultados de los experimentos de clasificación para la tarea *granular* de detección de discurso de odio, expresados como la media de las distintas métricas sobre diez corridas independientes. Cada modelo es un BERT con tres posibles entradas: sólo el comentario (*Sin contexto*), el tweet de la noticia a la cual responde el comentario (*Tweet*), y el tweet más el cuerpo de la noticia (*Tweet + Cuerpo*). Para cada una de estas posibilidades usamos dos versiones: una sobre BETO(¬FT) y otra sobre BETO ajustado al dominio (FT) de acuerdo a lo descripto en la Sección 6.3

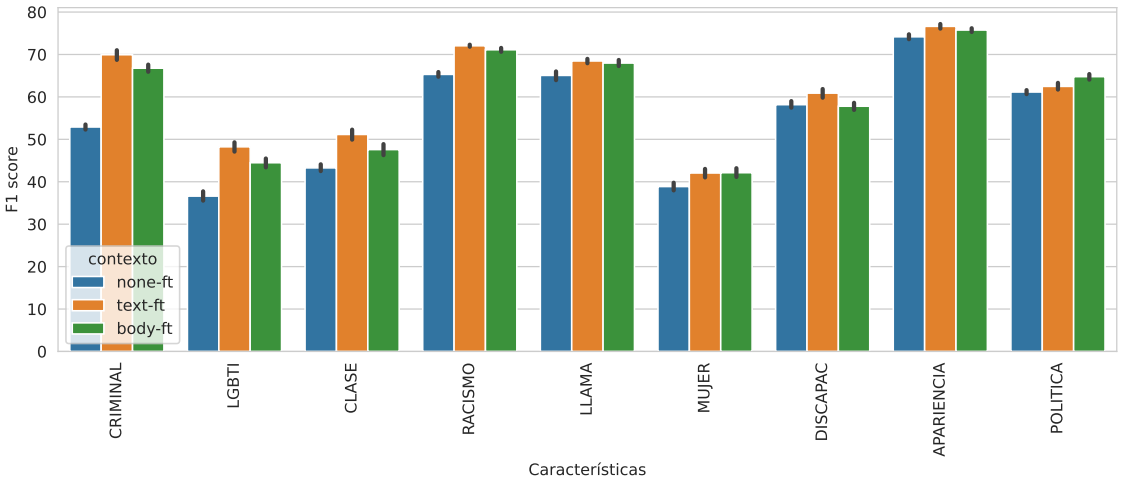
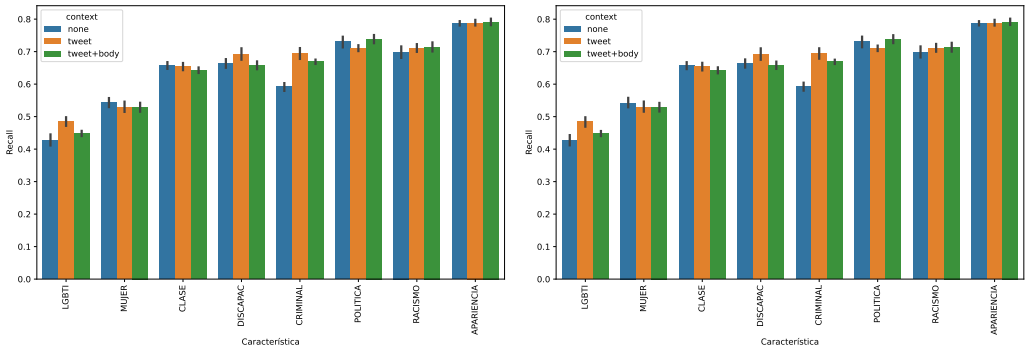


Fig. 6.4: Métrica F1 para cada característica de la tarea granular, ordenadas de mayor a menor de acuerdo a la diferencia entre el modelo sin contexto y el modelo contextualizado.



(a) Sensibilidad exacta *(b)* Sensibilidad total

Fig. 6.5: Precisión y sensibilidad para cada característica de la tarea granular. Las diferentes barras marcan el tipo de contexto que recibe el clasificador. Sensibilidad exacta (Figura 6.5a) cuenta la sensibilidad sobre la salida exacta de cada categoría, mientras que la sensibilidad total cuenta como recuperado un tweet si al menos alguna característica del clasificador lo marca como discurso de odio.

Hochberg [18]). Las diferencias más sustanciales se dan en el caso de CRIMINAL (+17 puntos F1 de diferencia), LGBTI (+12 puntos), CLASE (+8 puntos), y RACISMO (casi +7). Del otro lado, las características con menos mejora son APARIENCIA y POLITICA, algo esperable dado que el fenómeno tiene características poco dependientes del contexto como observamos en algunos ejemplos de la Sección 5.7.2 y por la misma definición y ejemplos considerados (ver Apéndice B). Finalmente, y como resumen de estas tablas, se observa que los modelos con contexto simple son los que mejor performance tienen, en general y para cada característica, con la excepción de POLITICA.

Una forma distinta de evaluar el rendimiento de los clasificadores granulares es de acuerdo a su sensibilidad o capacidad de recuperación de comentarios discriminatorios, aún cuando esto ocurra por el motivo incorrecto. La Figura 6.5 ilustra esta comparación analizando la sensibilidad de dos maneras: exacta, donde consideramos recuperado un tweet sólo si el clasificador acierta a la característica analizada (es decir, si la característica es MUJER, el clasificador debe predecir MUJER); y total, donde consideramos un tweet recuperado si alguna característica es marcada por el clasificador independientemente de si es la correcta. Podemos ver que la categoría MUJER pasa de ser la de menor sensibilidad a dejar a la categoría LGBTI como la que tiene menor tasa de comentarios ofensivos recuperados. Análogamente, la categoría CLASE obtiene una mejora sustancial en su sensibilidad, algo compatible con el hecho de su solapamiento con otras características como RACISMO, POLITICA y CRIMINAL observado en la Figura 5.7. También puede observarse que, en líneas generales, el contexto favorece una mejora de la precisión para cada característica y también un mayor recall exacto. Para el caso de la sensibilidad total, el desempeño de los clasificadores se empareja entre las versiones contextualizadas y no contextualizadas para cada característica con la excepción notoria de LGBTI y CRIMINAL.

Como se mencionó en la Sección 6.2, un clasificador sobre la tarea granular puede convertirse fácilmente en uno para la tarea binaria tomando la disyunción lógica

de sus salidas: si se detecta al menos una característica ofendida, entonces el tweet contiene discurso de odio. De esta forma, podemos evaluar el desempeño en la tarea binaria de aquellos clasificadores entrenados para la tarea granular. La Tabla 6.6 muestra esta comparativa para las distintas métricas sobre los clasificadores ajustados a dominio. Podemos observar que, en todos los casos, entrenar el modelo sobre la tarea granular produce pequeñas mejoras en la performance de los clasificadores entrenados sobre la tarea binaria (aproximadamente +0,8 puntos de Macro F1 en cada de uno).

6.5. Comparación de clasificadores y análisis de error

En esta sección realizamos un análisis de los errores de los clasificadores y comparamos sus distintas variantes. Para ello, analizamos las diferencias entre:

- las predicciones del mejor clasificador en términos de performance contra las etiquetas doradas del dataset.
- las predicciones del mejor clasificador contextualizado contra las predicciones del mejor clasificador no contextualizado.
- las predicciones entre los clasificadores entrenados sobre contra las predicciones de los entrenados sobre la tarea binaria.

Para analizar los errores de los modelos, elegimos el clasificador de mejor performance sobre la tarea granular: el clasificador que consume el contexto más el tweet de la noticia, entrenado sobre un *BETO* ajustado a dominio (ver Tabla 6.5). De una manera similar a lo que realizamos en la Sección 4.6.1, entrenamos diez clasificadores y analizamos el error sobre el ensamble de voto mayoritario. Para ver los casos más problemáticos, analizamos aquellas características donde peor desempeño muestran los clasificadores: MUJER, LGBTI, y CLASE. Por lo observado en la Figura 6.5, los clasificadores tienen una sensibilidad muy baja para estas tres características, por lo cual centramos nuestro análisis en los casos falsos negativos.

La Tabla 6.7 muestra una selección de comentarios discriminatorios para la característica LGBTI que no son detectados por los clasificadores. De estas instancias, y observando también aquellos casos donde sí puede detectar el discurso discriminatorio de esta categoría, podemos esbozar algunas posibles razones detrás de estos

	Sin Contexto		Tweet		Tweet + Cuerpo	
	Bin	Gran	Bin	Gran	Bin	Gran
Precision	71,8	71,1	74,8	**75,9**	72,8	74,0
Recall	60,2	63,6	65,3	**66,7**	64,1	66,0
F1	65,5	67,1	69,7	**71,0**	68,1	69,7
Macro F1	79,8	80,6	82,2	**83,0**	81,3	82,2

Tab. 6.6: Desempeño de los modelos para la tarea de detección binaria de discurso de odio. Los modelos considerados son modelos *BETO* ajustados a dominio que consumen tres tipos de entrada: sin contexto, tweet, y tweet+cuerpo. Dos posibles entrenamientos fueron evaluados: sobre la tarea binaria (**Bin**) o sobre la tarea granular (**Gran**).

	Contexto	Comentario
1	Contó que era lesbiana, su papá le confesó que era gay y ahora su madre se enamoró de una mujer, lo que inspiró su segundo film	WTF. Mucho ESI los degeneró...
2		@usuario Esta familia tiene los genes alterados
3	Oscar González Oro ya está instalado en el Uruguay: Recuperé mi libertad”	Ahora quedate allá, y hablá mal d Macri d nuevo pa tener rating. Y opiná q los taxi boy uruguayos son mas educados q los escorts argentinos! Ano abierto!
4	¿Por qué un beso entre dos hombres los vuelve tan violentos?”: la vida después de haber sido víctima de ataques homofóbicos	Será xq va contra la naturaleza de la raza...
5	“¿Por qué no vemos médicos trans?”: el reclamo de un prestigioso cardiólogo para que América sea más inclusiva	Es difícil ser médico con la cabeza quemada
6		porque un enfermo no cura a otro enfermo
7	“Te amo”. La emotiva dedicatoria de Luis Novaresio a su pareja en su cumpleaños	🤮🤮🤮
8	Elizabeth Gómez Alcorta: “Por la pandemia, vamos a tener una suba de los femicidios y travesticidios”	Travesticidios... Osea asesinatos de tipos con peluca y tetas
9	Mariana Genesio Peña pasa su cuarentena total con guantes, barbijo y desnuda: “Mi cuerpo es el planeta Tierra”	Coronavirus nivel pelotudo en bolas
10		Con 3 piernas cualquiera es feliz!!!
11		pasa la cuarentena rascándose las bolas
12	Tras una ráfaga de más de 20 disparos asesinaron a una mujer trans en Rosario	Cómo no saco su escopeta y aplicó la defensa propia?!
13		@usuario Salió de caño... cuac!

Tab. 6.7: Falsos negativos para la característica LGBTI. Ninguno de los diez clasificadores que consumen contexto y texto indentificaron como discriminatorios a estos comentarios.

errores. En primer lugar, ciertos mensajes altamente ofensivos son complejos de entender: por ejemplo, los que tratan de “enfermos” o mencionan cuestiones de la genitalidad (ejemplos 5 y 6). Estas ofensas son realizadas por los usuarios de maneras tan sofisticadas que difícilmente un modelo de lenguaje pueda entender, como metáforas varias que hacen referencia a la genitalidad de una mujer trans (ejemplos 10 y 12) o bien refiriéndose al objetivo de la ofensa con un género distinto al autopercibido (ejemplos 8 y 9),

Algunos comentarios y sus contextos –como los ejemplos 3, 7, 9, 10 y 11– omiten información acerca de la sexualidad o género sobre quienes versa la noticia. Esta información faltante no permite a los clasificadores (ni tampoco a un humano que carezca de esta información) entender completamente el caracter discriminatorio de los mensajes.

En el caso de MUJER, restringimos nuestro análisis a aquellos ejemplos que

	Contexto	Comentario
1	Martha Rosenberg: “En situación de pandemia, legalizar el aborto es más urgente que nunca”	Quien es esta vieja?. No debería estar tejiendo?
2	Mara Gómez: la historia de la primera futbolista trans en el torneo argentino	Feminismo pierde de nuevo... ya le metieron un tipo... jaja punto para el patriarcado...
3	Tras una ráfaga de más de 20 disparos asesinaron a una mujer trans en Rosario	Las feministas en modo error 404 al no saber si celebrar o ofenderse
4	El desesperado pedido de Actrices Argentinas ante la violencia de género en cuarentena: “Es urgente”	Que risa me dan las feministas!!! Ignorantes.
5	Leche de cucaracha, la nueva bebida nutritiva: ¿quién se anima a probarla?	No me digas q la hija de CFK está embarazada y ya sale leche por esos senos
6	Los fans de Florencia Kirchner le piden casamiento por Instagram	Zoofilia
7		Hdp tienen que tener estomago para querer casarse con terrible adefesio
8	Rosario: para sacar una licencia de conducir habrá que hacer un curso de perspectiva de género	Te quieren adoctrinar desde cualquier ámbito, y se están metiendo en todo para que empieces a hablar como el orto, como a ellas les gusta.
9		El que choca más feministas le dan más años de licencia
10	Por qué los países liderados por mujeres parecen haber respondido mejor a la crisis del coronavirus	Son mujeres inteligentes que se dejan asesorar de sus esposos
11	Joe Biden presentó su nuevo equipo de comunicación compuesto enteramente por mujeres	Será equipe conche seque?

Tab. 6.8: Falsos negativos para la característica MUJER. Ninguno de los diez clasificadores que consumen contexto y texto (ajustados a dominio) lograron identificar como discriminatorios a estos comentarios para ninguna otra característica.

no son detectados por otras características ya que, por lo observado en la Figura 6.5, muchos ejemplos están en la frontera de otras categorías y son detectados por ellas. La Tabla 6.8 muestra una selección de estos falsos negativos, donde podemos apreciar que algunos ejemplos están en el borde borde de ser simplemente ofensivos (ejemplo 1 y posiblemente 6 y 7) o contienen mensajes irónicos complejos de descifrar (ejemplo 3, que las feministas celebren por la muerte de una mujer trans, o ejemplo 9, hablar de chocar mujeres en un curso de manejo). Esto es esperable dado el acuerdo relativamente bajo para la categoría MUJER en el etiquetado de este dataset reportado en la Tabla 5.7.

En la Tabla 6.9 podemos observar una selección de ejemplos donde el clasificador contextualizado acierta en su predicción mientras que el modelo sin contexto se equivoca, separados en falsos negativos (el modelo no contextualizado no logra detectarlos) y falsos positivos (el modelo no contextualizado predice equivocadamente que son discriminatorios). En el caso de LGBTI, la información contextual permite

	Tipo	Contexto	Comentario
CRIMINAL			
1	FN	Una policía baleó y mató a un joven de 17 años que la atacó con una tijera en Moreno url	👏 excelente!
2		El polémico cortejo del ladrón asesinado por el jubilado en Quilmes	Le hubieran puesto una bomba al cortejo
3	FP	Ivana Nadal se cansó de la criticaran y sorprendió con su respuesta: "Gracias a Dios a fin de año me voy del país"	Esa si es una buena noticia.
4		¿Se va del país? Juana Viale estaría tramitando la ciudadanía uruguaya	Q bueno ? Una mierda menos . Q se quede alla
LGBTI			
6	FN	"Te amo". La emotiva dedicatoria de Luis Novaresio a su pareja en su cumpleaños	Definitivamente no acepto esta degeneración repugnante de la humanidad.
7		Mara Gómez: la historia de la primera futbolista trans en el torneo argentino	Sigue siendo HOMBRE, que por GENÉTICA, no por una ideología u orientacion sexual, GE-NÉ-TI-CA, es más fuerte que la mujer.
8	FP	URGENTE: Un hombre se incrustó con su auto en la puerta de la Embajada de China y aseguró que tenía explosivos	No es hombre . Es un boludo
9		Detuvieron al hombre que admitió violar a su hija en audios de WhatsApp y fue tendencia en redes: un tío lo entregó	Degenerado 🤮
CLASE o RACISMO			
10	FN	Las organizaciones sociales salieron al cruce de la acusación de Sergio Berni por la toma de tierras: "La falta de vivienda no se resuelve con balas"	Si, los podemos cagar matando a todos y listo
11		La temporada de verano en la Costa Atlántica empezó con un corte total en la Ruta 2: organizaciones sociales piden canastas navideñas	Hay que desparasitar urgente el país.
12		La pregunta billonaria: ¿China debería pagar el costo de la pandemia?	Si obviamente y desaparecer de la faz de la tierra. Mira el quilombo que armó. Se nos están muriendo todos...
13	FP	Javier Milei confirmó que va "a militar en política" junto a José Luis Espert para que "en 35 años la Argentina sea primera potencia mundial"	Otro parásito
14		Martha Rosenberg: "En situación de pandemia, legalizar el aborto es más urgente que nunca"	La verdad que sí...así se dejan de reproducir!!!

Tab. 6.9: Ejemplos donde el clasificador contextualizado acierta y el no contextualizado falla. FN marca que el clasificador no contextualizado no detecta el comentario como discriminatorio (ni para la característica marcada ni para otras) mientras que el contextualizado sí lo hace; FP es al revés, que el clasificador no contextualizado marca erróneamente el comentario como discriminatorio

desambiguar casos como los 7 y 8, muy similares pero con un contexto sumamente distinto. Un problema que se puede apreciar es que el clasificador no contextuali-

zado –al ser entrenado con datos etiquetados de manera contextualizada– aprende correlaciones espurias como que las celebraciones son discriminatorias (ejemplo 16).

La comparativa entre los clasificadores entrenados sobre la tarea binaria y la tarea granular es más difícil de interpretar. Si bien se observa que entre los falsos negativos del clasificador binario se encuentra una proporción más alta de tweets racistas, es difícil elucidar una razón detrás de esto. En la Tabla C.1 del Apéndice C se encuentra una muestra de ejemplos en los cuales el clasificador granular acertó y el binario falló.

6.6. Discusión

Para analizar el impacto del contexto en la detección de discurso de odio, planteamos dos tareas de clasificación sobre el conjunto de datos construido en el Capítulo 5: la tarea de detección binaria, donde predecimos si un comentario contiene discurso de odio; y la tarea de detección granular, donde se deben predecir las características protegidas ofendidas (si agrede a las mujeres, al colectivo LGBTI, si es racista, etc). Propusimos clasificadores basados en *BETO* que consumieron tres tipos de entrada: el comentario sin contexto, el comentario con el contexto del tweet al que responden, y por último el comentario más el tweet al que responde junto al texto del artículo periodístico.

Para ambas tareas, pudimos observar en los experimentos realizados que el contexto en su versión simple (sólo el tweet de la noticia) brinda una mejora moderada pero estadísticamente significativa en la tarea de detección binaria de discurso de odio (alrededor de 3 puntos F1) y una mejora considerable en la tarea granular (alrededor de 6 puntos de Macro F1) contra los clasificadores que no tienen información contextual. Esto indicaría que el contexto puede ser aprovechado para mejorar los algoritmos de detección de discurso de odio. Si bien este resultado podría estar en aparente contradicción con trabajos recientes que no encontraron ninguna mejora en el uso del contexto en la detección de toxicidad [121], se puede señalar que la detección del discurso de odio es una de las formas más complejas de este comportamiento. Como tal, el contexto podría permitir –para este subconjunto de contenido tóxico– que los clasificadores tengan más información para predecir si el texto dado es discriminatorio o no. Otra razón detrás de este resultado es el dominio de nuestro conjunto de datos: mientras que Pavlopoulos et al. [121] usaron el título de un thread de Wikipedia Talk Pages y parcialmente el hilo conversacional, nosotros sólo usamos el título y el cuerpo del artículo como contexto para los comentarios de los usuarios. Más recientemente, Xenos et al. [170] han observado que los algoritmos de detección de toxicidad pueden aprovechar esta información adicional al restringir el análisis a un subconjunto de comentarios sensibles al contexto (ver Secciones 6.1 y 5.1 para más información).

En nuestros experimentos, la utilización de un contexto más largo (el artículo de la noticia) no reportó mejoras en el rendimiento de los clasificadores. Este resultado sería compatible con el hecho de que los respuestas de los usuarios a artículos periodísticos suelen estar basadas en poquísima información como el titular o en este caso el tweet acerca de la noticia. Sin embargo, los humanos solemos tener acceso a un contexto mucho más rico –muchas veces equivalente a haber leído la nota– y

también a conocimiento adicional del mundo real, algo que se ha mostrado que los modelos pre-entrenados de lenguaje carecen [16, 96]. Una posible razón adicional a esto puede ser atribuido a que el modelo pre-entrenado que usamos para codificar el texto del artículo (*BETO*) fue pre-entrenado sobre textos más cortos. Teniendo esto en cuenta, realizamos el ajuste de dominio usando los textos de artículos periodísticos, pero aún así el desempeño del clasificador que consume esta entrada se mantuvo por debajo del que consumía sólo el tweet original. Trabajo futuro debería estudiar formas de incorporar esta información de manera que pueda ser utilizada adecuadamente por los modelos.

El análisis del error realizado dio muestras de que la tarea de detección de discurso de odio es difícil para muchas características, aún considerando la mejora que reporta la utilización de contexto. Un caso notable observado en nuestros experimentos es la discriminación contra el colectivo LGBTI. En las instancias del dataset –y en muchos de los comentarios en los que el algoritmo de detección falla– puede verse que las agresiones contra este colectivo y sus miembros son sumamente sofisticadas, lejos de las agresiones meramente basadas en palabras o expresiones insultantes. Los clasificadores entrenados, aún en sus mejores versiones, obtuvieron una baja performance en la detección de este fenómeno (alrededor de 48 puntos de F1) comparada con la tasa de detección humana (casi 60 puntos). Estos resultados marcan la no trivialidad de esta tarea, indicando la necesidad de analizar este punto más detenidamente debido a la complejidad de estos mensajes, que suelen reunir ironía, metáforas, y otros artilugios que hacen difícil su detección.

Una particularidad de un subconjunto considerable de comentarios homofóbicos y transfóbicos en noticias periodísticas es que suelen ser dirigidos contra un individuo integrante de la comunidad LGBTI. Esta información –su pertenencia al colectivo– no está en todos los casos disponible en el contexto de la noticia o no puede llegar a ser captado por nuestros modelos, hecho que dificulta la detección de este tipo de agresiones. Este problema es generalizable a otras características protegidas además de LGBTI, y puede también marcar una posibilidad de mejora con la incorporación de conocimiento externo a los modelos (como es propuesto por Liu et al. [94]) para ayudar a los clasificadores a mejorar su tasa de detección.

En el caso de la detección de agresiones misóginas (MUJER) obtuvimos también tasas de desempeño muy bajas. Analizando los errores, pudimos observar que tenemos casos complejos de descifrar en los datos como, por ejemplo, ataques velados a mujeres víctimas de violación (llamarlas mentirosas). Estos casos caen en las agresiones individualizadas recién descriptas, las cuales pueden requerir conocimiento adicional del mundo real. De todas maneras, la baja tasa de acuerdo entre humanos pone una cota baja al desempeño máximo de los algoritmos automáticos.

Algo que debe ser tenido en cuenta para matizar estos resultados es que utilizamos un amplio espectro de características protegidas. Incluso, la característica más beneficiada por la adición del contexto es CRIMINAL, una categoría que puede ser considerada ad-hoc para este experimento. En contrapeso, otras características no convencionales son poco beneficiadas por el contexto (como discurso de odio en base a la apariencia, opinión política y discapacidad). Otro subtipo de discurso de odio muy beneficiado por la adición de contexto es el dirigido contra la comunidad LGBTI, a pesar de las dificultades señaladas en su detección.

En los experimentos observamos que los clasificadores mejoran levemente su performance en términos de detección de discurso de odio al ser entrenados para la tarea granular. Si bien la mejora es marginal (cerca de un punto de F1) y no es apreciable de manera subjetiva mediante un análisis de error, una posible razón detrás de esto es que la señal más precisa acerca de la categoría ofendida puede ayudar a distinguir mejor las fronteras de este fenómeno. Este resultado es coherente con lo observado en la Sección 4.6, donde la adición de otras variables a predecir no empeoraban el desempeño de los modelos, y con Gertner et al. [59] donde se reporta una mejora al modelar –de manera no supervisada– las características ofendidas. De todas formas, aún cuando no existiese mejora en la detección, poder tener una salida más interpretable y granular es mejor que simplemente obtener una predicción binaria.

Una limitación importante de este estudio es que todos los modelos de clasificación fueron entrenados sobre datos etiquetados observando el contexto. Algunos ejemplos de errores que observamos en el clasificador no contextualizado –celebraciones marcadas como discurso de odio– dan cuenta del problema que esto genera. Un estudio más completo del impacto del contexto debería incluir datos que sean etiquetados sin observar ningún tipo de contexto como es realizado en Pavlopoulos et al. [121], algo que por limitaciones de tiempo y recursos no fue posible realizar en esta tesis.

En el terreno de la aplicación, un problema práctico de este resultado es que no siempre tenemos un contexto disponible para un texto dado. Incluso si podemos encontrarlo, muchas veces este contexto puede no ser en forma de artículo de noticias sino como un hilo de conversación o incluso de alguna otra representación. Teniendo en cuenta alguna de las consideraciones hechas en esta discusión, una línea de investigación podría evaluar la incorporación de distintos tipos de contexto, desde más mensajes en el hilo de la conversación, conocimiento estructurado que consuma algún modelo *knowledge-aware* (*ERNIE* [177] o variantes de BERT que inyectan extractos de grafos de conocimiento [49, 94]) o bien una combinación de diversas fuentes de información.

6.7. Conclusiones

Hemos analizado aquí el impacto de la utilización del contexto en la detección automática de discurso de odio, realizando para ello experimentos de clasificación sobre el conjunto de datos construido en el capítulo anterior. Los resultados de estos experimentos dan indicios de que cierta información contextual puede ser de ayuda para mejorar la capacidad de detección de estos algoritmos. Si bien en nuestros experimentos el contexto más pequeño (el tweet del artículo de la noticia) fue el que mejor resultados obtuvo, una línea de trabajo futuro podría explorar formas de incorporar otras fuentes de información.

Del análisis de error, se puede apreciar que algunas categorías del discurso de odio se muestran esquivas para los algoritmos de detección del estado del arte. Uno de estos casos son los mensajes abusivos contra la comunidad LGBTI, que contienen mensajes semánticamente complejos, con carga irónica y metáforas que son difícilmente interpretables para los clasificadores basados en modelos de lenguaje del estado del arte. A pesar de estas limitaciones, la detección de discurso de odio

contra la comunidad LGBTI fue una de las más beneficiadas por la adición de contexto.

Podemos concluir que los datasets de discurso de odio deberían –en la medida de lo posible– contener **información contextual** sobre los comentarios analizados. Esta información puede darse en forma de artículos de noticias, como un hilo de conversación, o incluso como otras formas –por ejemplo, como una base o grafo de conocimiento. Sobre esto, trabajo futuro debería explorar el impacto de utilizar esta información adicional para integrarla en algoritmos de detección de discurso de odio. La evidencia de los experimentos realizados –por ahora preliminares, y con las limitaciones marcadas en la discusión– indica que los modelos del estado del arte pueden utilizar esta información para mejorar la detección de discurso de odio en redes sociales. En segundo lugar, los datasets de discurso de odio deberían incluir **información granular** acerca de las características atacadas –y no sólo una etiqueta binaria– ya que por un lado esto mejora la interpretabilidad de los algoritmos de detección, y, por el otro, resultados preliminares de este estudio indican que utilizar información detallada de las característica ofendidas mejora marginalmente el rendimiento en la detección en general.

Finalmente, un aspecto que introdujimos en este capítulo fue el de adaptar un modelo de lenguaje pre-entrenado a su dominio, siendo en nuestro caso los comentarios sobre notas periodísticas en Twitter. Las mejoras que reportó la utilización de estas técnicas fueron considerables, en consonancia con otros trabajos recientes. Pasamos a continuación a estudiar estas técnicas en el marco más general de la clasificación de textos sociales.

Parte IV

ADAPTACIÓN DE DOMINIO

7. ADAPTACIÓN DE DOMINIO

En este capítulo analizamos cómo mejorar la detección de discurso de odio desde una perspectiva más general, centrándonos en aplicar técnicas de **adaptación de dominio** a los algoritmos de clasificación que han sido considerados hasta el momento. Como hemos mencionado en la Sección 3.5, cuando hablamos de adaptación de dominio nos referimos al conjunto de técnicas y recursos destinados a tratar de que los algoritmos tengan un correcto desempeño en un subconjunto de tareas relacionadas entre sí [63]. En el caso concreto de las tareas cuyas instancias constan de analizar texto de usuarios en redes sociales u otros medios, Eisenstein [46] describió a la adaptación de dominio en términos de la necesidad de construir herramientas propias para este tipo de texto, muy distinto al lenguaje formal proveniente de otras fuentes.

Las técnicas de representación modernas –desde los embeddings hasta los modelos de lenguaje– suelen ser entrenadas sobre conjuntos de datos que se suponen lo suficientemente generales. Fuentes usuales son Wikipedia, que comprende textos de carácter enciclopédico, o Common Crawl, que es una recopilación de datos de distintos sitios web. El uso del lenguaje en estas fuentes suele guardar una considerable discordancia con el de muchas tareas de interés en NLP, como son aquellas basadas en textos provenientes de ciertos nichos donde el uso del lenguaje es muy específico. Ejemplos de esto son los documentos médicos, trabajos científicos, entre otros. A cada uno de estos grupos de textos con cierta relación se los denomina –de manera poco precisa– **dominios**. Entre estos, el contenido informal de las redes sociales tiene variedades lingüísticas muy particulares, con mucha jerga, expresiones coloquiales, errores ortográficos y demás que diferencian el uso del lenguaje de los textos fuentes mencionados.

Continuamos en este Capítulo alguna de las ideas y observaciones que hemos analizado previamente acerca de la adaptación de dominio. En primer lugar, en el Capítulo 3 se ha observado que las técnicas de representación –desde los word-embeddings hasta los modelos de lenguajes– tienen un buen desempeño sobre tareas de textos de redes sociales son entrenadas sobre este mismo dominio. Para el idioma español, sin embargo, no existe ningún modelo de lenguaje fácilmente accesible de estas características, como BERTweet para el inglés o AlBERTo en italiano. Describimos entonces el proceso de entrenamiento desde cero de un modelo de lenguaje sobre tweets en español, al cual llamamos *RoBERTuito*. Evaluamos su rendimiento compilando todas las tareas analizadas en esta tesis, y mostramos que es superior a todos los modelos considerados hasta el momento.

En segundo lugar, retomamos otro enfoque para adaptar las técnicas existentes al dominio de redes sociales. En el Capítulo 6 conseguimos mejorar la performance de los algoritmos de clasificación mediante la continuación del pre-entrenamiento del modelo de lenguaje sobre un gran conjunto de datos de noticias y comentarios no etiquetados, los cuales fueron recolectados como parte del proceso de construcción del dataset usado. Trabajo reciente muestra que esta técnica es generalizable a muchos dominios –como textos médicos, científicos, entre otros– y resulta en mejoras

consistentes del rendimiento de los algoritmos de clasificación basados en *BERT* y similares. En nuestro caso, analizamos el efecto de continuar el pre-entrenamiento ya no sólo sobre la tarea de detección contextualizada de discurso de odio sino sobre todas las tareas que hemos visto en los diferentes capítulos.

Finalmente, realizamos una comparación entre estas dos aproximaciones a adaptar nuestros algoritmos a este medio, algo de lo que no tenemos conocimiento se haya realizado hasta el momento, al menos para el idioma español. Este punto es de interés dado el gran costo que tiene entrenar modelos basados en Transformers, y sirve para verificar si el salto en el rendimiento de los modelos generados para dominios específicos puede subsanarse mediante una alternativa más económica como es la continuación del pre-entrenamiento.

Comenzamos en la Sección 7.1 haciendo un racconto de las técnicas de adaptación de dominio y modelos pre-entrenados. Describimos a continuación la construcción y el entrenamiento de *RoBERTuito* [127]. Finalmente, detallamos los experimentos de adaptación utilizando *BETO* como modelo de lenguaje a adaptar para luego comparar su desempeño contra *RoBERTuito*.

7.1. Trabajo previo

La definición de qué es un **dominio** en NLP suele ser relativamente amplia y poco precisa. Una posible aproximación a este concepto es la de un conjunto de textos que guardan cierta similaridad respecto al tópico o género; al medio utilizado; el público al cual apuntan, entre otras cosas. Algunos ejemplos de dominios podrían ser los artículos de noticias, las novelas u otros libros de ficción, discursos políticos, comentarios de redes sociales, entre otros [68]. Subdisciplinas del procesamiento del lenguaje natural tienen su eje en tratar estas distintas categorías de documentos atendiendo sus particularidades: BioNLP, SocialNLP, entre otras nuevas denominaciones.

Goodfellow et al. [63] definen la adaptación de dominio como una situación similar a la de Transfer Learning: dado un modelo que fue entrenado para una tarea T y una distribución de datos P_1, se lo quiere utilizar sobre la misma tarea T pero esta vez con una distribución de entrada P_2 bajo la asunción de que ambas distribuciones son relativamente similares. Un escenario posible es el de un clasificador de polaridad entrenado sobre comentarios acerca de reviews de libros, el cual queremos utilizar para analizar reviews de productos electrónicos. Glorot et al. [61] es uno de los primeros trabajos que aplica la idea de adaptación de dominio sobre modelos de Deep Learning en NLP. Los autores emplean *stacked denoising auto-encoders* para aprender características no supervisadas de los textos; para atacar la diferencia de distribuciones de entrada, realizan pre-entrenamiento no supervisado para los dominios analizados. Desde una óptica diferente, Eisenstein [46] describe puntualmente la adaptación de dominio para textos de redes sociales como "adaptar las herramientas al texto" (social), contraponiendo esto al concepto de *normalización* de la entrada, que sería intentar "adaptar el texto a las herramientas". Dentro de las tareas de extracción de opiniones, la adaptación resulta importante ya que expresiones en distintos ámbitos pueden tener sentidos distintos: decir "leé el libro" en una reseña de un libro Amazon puede ser algo positivo, mientras que en el comentario de una

película puede ser considerado negativo [117].

Dentro de la ola de modelos pre-entrenados que sacudió el mundo de NLP, la técnica de *ULM-FIT* [77] (descripta en la Sección 2.4.1) contempla tres etapas: pre-entrenamiento, ajuste de dominio, y ajuste discriminativo. En la etapa intermedia, se adaptan los pesos de la red neuronal utilizando de manera no-supervisada el texto del dataset, realizando una continuación del pre-entrenamiento de modelado de lenguaje. Los modelos basados en Transformers como BERT [45] y subsiguientes eliminaron esta etapa intermedia, dejando sólo la adaptación de los pesos sobre las etiquetas supervisadas. Recientemente, se ha observado que reintroducir esta adaptación no supervisada es algo beneficioso.

Siguiendo esta idea, y restringiendo nuestro interés a las técnicas actuales de clasificación, puede ser beneficioso ajustar un modelo de lenguaje a un dominio distinto al que fue utilizado en su pre-entrenamiento: puntualmente, queremos adaptar *BERT* (entrenado en Wikipedia) a textos de carácter más informal. Si bien se ha observado que los modelos de lenguaje basados en Transformers son mucho más robustos frente a los cambios de dominio que otros algoritmos previos [73], todavía siguen sufriendo cuando los datos analizados difieren fuertemente de los utilizados en el entrenamiento. Ruder [142] hace un repaso extenso de los últimos avances en las técnicas de adaptación de dominio utilizando modelos de lenguaje del estado del arte.

Gururangan et al. [68] analizan el impacto de continuar el pre-entrenamiento para los modelos de lenguaje basados en Transformers. En su estudio, consideran aplicar esta técnica sobre diversos dominios en inglés: biomédico, reviews de películas, papers de cs. de la computación (CS), y noticias. A diferencia de *ULM-FIT*, que sólo plantea el ajuste de pesos de acuerdo al conjunto de datos de una tarea particular, los autores plantean dos alternativas:

- *Domain Adaptation*: ajustar el modelo de lenguaje sobre un extenso conjunto de datos no etiquetado, usualmente el sobrante del proceso de recolección que no es anotado.
- *Task Adaptation*: ajustar el modelo de lenguaje sobre el dataset, de la misma manera que Howard and Ruder [77].

Usando como modelo base a *RoBERTa*, los autores reportan que los clasificadores aumentan su rendimiento para diversas tareas de cada dominio de manera significativa, tanto en el caso de realizar adaptación de dominio como en la adaptación a la tarea. Si bien las mayores ganancias se obtienen para la segunda, algunos resultados de ese mismo trabajo indicarían que este tipo de pre-entrenamiento puede dañar la generalización, posiblemente debido a un sobreajuste a las instancias del dataset.

Posteriormente al estallido de los modelos de lenguaje basados en Transformers, algunos trabajos se han dedicado a entrenar directamente estos modelos sobre un dominio de interés particular y no en textos genéricos como Wikipedia. Por ejemplo, *SciBERT* [12], *MediBERT* [137] y *LegalBERT* [32] están entrenados sobre textos científicos, médicos y legales respectivamente. AlBERTo [133] es uno de los primeros modelos entrenados directamente sobre tweets, particularmente para italiano; el ya mencionado *BERTweet* [113] prosiguió esta línea de investigación, siendo construido

Nombre	Idioma	Dominio	Familia
SciBERT	inglés	Papers	BERT
ClinicalBERT	inglés	noticias médicas	BERT
MediBERT	inglés	registros médicos	BERT
LegalBERT	inglés	legislación, contratos	BERT
BERTweet	inglés	tweets, algunos sobre COVID	RoBERTa
AlBERTo	italiano	tweets	RoBERTa
TwilBERT	español	tweets	∼BERT

Tab. 7.1: Modelos pre-entrenados sobre dominios no canónicos. En familia nos referimos a qué tipo de pre-entrenamiento es realizado en el modelo de lenguaje: BERT es MLM + NSP, RoBERTa sólo MLM. En el caso de TwilBERT, se usa un símil BERT ya que no usan exactamente NSP sino una tarea muy parecida.)

mediante un pre-entrenamiento similar al de *RoBERTa* [95] sobre cerca de 850M tweets en inglés, una parte de ellos relacionados a la pandemia del COVID-19. En español tenemos el modelo TwilBERT [62]; sin embargo, tiene algunas limitaciones: en primer lugar, no queda claro cuánto tiempo de entrenamiento recibió ni si los datos fueron suficientes; en segundo, usaron un modo de entrenamiento basado en una variante de la tarea NSP (ver Subsección 2.4.4) cuando numerosos trabajos muestran que el tipo de entrenamiento basado en RoBERTa (sólo tarea MLM) mejora el desempeño en las tareas finales. Finalmente, su modelo no es accesible mediante el model hub de huggingface, limitando seriamente su acceso. La Tabla 7.1 lista algunos de estos modelos.

Algunas oportunidades de mejora de lo estudiado en Gururangan et al. [68] son, en primer lugar y siguiendo la regla de Bender [14], realizar el estudio en un idioma distinto al inglés. En segundo lugar, estudiar el dominio de tareas en textos de redes sociales, algo no realizado en dicho trabajo. Finalmente, es de interés realizar una comparación de la performance de modelos adaptados al dominio contra aquellos que son entrenados desde cero, dado el enorme costo computacional, energético y ambiental que implica esto último [156].

7.2. Modelo pre-entrenado sobre tweets

En esta sección describimos el proceso de construcción de *RoBERTuito*. Entrenamos tres versiones de este modelo: una versión que preserva las mayúsculas del texto (nombrada **cased**); una versión que convierte todo a minúsculas (**uncased**); y una versión que convierte todo a minúsculas y elimina las tildes (**deacc**). El español normativo prescribe el uso de tildes en ciertos casos para señalar la acentuación de una palabra, algo que por lo general suele pasarse por alto en los textos escritos en redes sociales. Una hipótesis de trabajo es que eliminar esta información (tildes y mayúsculas) puede ayudar al rendimiento de los modelos en las tareas finales, al ser su uso tan inconsistente en el texto informal.

7.2.1. Recolección de tweets

A continuación describimos el proceso de recolección de tweets que utilizamos para entrenar *RoBERTuito*. El stream de API de acceso gratuito de Twitter (también conocida como *Spritzer*) es una subconjunto generado en tiempo real de alrededor del 1 % de los tweets. Esta muestra es supuestamente aleatoria, aunque algunos estudios han mostrado algunas preocupaciones acerca de posibles formas de manipularla [130]. Muestras no representativas y sesgadas pueden afectar al modelo en tareas finales, algunos de estos errores pudiendo ser dañinos generando sesgos raciales o de género. Por ello, publicamos el conjunto de datos para ser inspeccionado, y queda como trabajo futuro un análisis más detallado de sus instancias.

Descargamos primeramente una colección de Spritzer subida a Archive.org que data de Mayo de 2019 [1]. Filtramos aquellos tweets cuya metadata indicase que su idioma no sea el español, y nos guardamos los usuarios que los generaron. Sobre este conjunto de usuarios, usamos la API de Twitter para descargar todos sus tweets. De este proceso recolectamos alrededor de 622 millones de posteos de cerca 432 mil usuarios.

Finalmente, nos quedamos sólo con aquellos tweets que tengan 6 o más tokens, usando para esto el tokenizador entrenado en BETO [31], sin contar repeticiones de emojis y haciendo el preprocesado descripto en capítulos anteriores: reemplazamos los caracteres hasta un máximo de tres, convertimos los nombres de usuarios a un token especial `@usuario`, convertimos los emojis a una representación textual, y partimos los hashtags en lo posible (ver Sección 3.5). De este proceso obtuvimos 500 millones de tweets, ordenados en 1,000 archivos para facilitar la lectura en procesos posteriores. El repositorio de la recolección de tweets puede encontrarse en `https://github.com/finiteautomata/spritzer-tweets`.

Algo a destacar es que este proceso de recolección permitió que los datos contengan texto con *code-switching*[2] o incluso tweets de otros idiomas, ya que solo requerimos que la publicación en la muestra original esté en español. Mientras que otros trabajos como Nguyen et al. [113] requirieron que cada tweet estuviera en inglés, permitimos que se incluyeran otros idiomas en los datos previos al entrenamiento. Una estimación aproximada de la distribución lingüística utilizando el módulo de detección de idioma de *fasttext* [81] indicó que el 92 % de los datos están en español, el 4 % en inglés, el 3 % en portugués y el resto en otros idiomas.

7.2.2. Arquitectura y entrenamiento

Para cada una de las versiones de *RoBERTuito* (cased, uncased, deacc) entrenamos tokenizadores usando el algoritmo *SentencePiece* [88] en los tweets recopilados, disponiendo de un vocabulario de 30,000 tokens en todos los casos. Usamos la librería *tokenizers* [3] que proporciona implementaciones rápidas en el lenguaje de programación *Rust* para muchos algoritmos modernos de tokenización.

[1] `https://archive.org/details/archiveteam-twitter-stream-2019-05`

[2] Texto o comunicación verbal que fusiona dos idiomas, por ejemplo en spanglish u otras mezclas lingüísticas

[3] `https://github.com/huggingface/tokenizers`

Hiperparámetro	Valor
#Cabezas	12
#Capas	12
Tamaño oculto	768
Tamaño intermedio	3,072
Función de Activación	GeLU
#Vocabulario	30,000
Probabilidad MLM	0,15
Tamaño de secuencia	128
Batch size	4,096
Learning Rate	$3{,}5 * 10^{-4}$
Decay	0,10
β_1	0,90
β_2	0,98
ϵ	10^{-6}
Pasos de warmup	36,000 (6 %)

Tab. 7.2: Hiperparámetros utilizados en el entrenamiento de *RoBERTuito*. Los valores de β y ϵ refieren a los hiperparámetros de Adam

Se utilizó una arquitectura *RoBERTa* base para el modelo, con 12 capas de auto atención, 12 cabezas de atención y tamaño intermedio 768, de la misma manera que *BERTweet*. Entrenamos *RoBERTuito* sobre la tarea de MLM, en la misma línea de *RoBERTa* y *BERTweet*, sin tener en cuenta la tarea de predicción de la siguiente oración usada en *BERT* u otras tareas de orden de tweets (como la usada en Gonzalez et al. [62]). Teniendo en cuenta los hiperparámetros de *RoBERTa* y *BERTweet*, decidimos utilizar un tamaño de batch size grande para nuestro entrenamiento. Si bien se recomienda un tamaño de 8,192 en Liu et al. [95] y Nguyen et al. [113], debido a las limitaciones de recursos decidimos aumentar el número de actualizaciones utilizando un tamaño de lote de 4,096. La Tabla 7.2

Para comprobar la convergencia, primero entrenamos un modelo para 200,000 pasos de optimización. Al comprobar que convergió (y obtuvo buenos resultados en las tareas del benchmark que describimos a continuación), procedimos al entrenamiento completo de los tres modelos. El proceso de pre-entrenamiento tomó aproximadamente tres semanas en una TPU *v3-8* y una máquina pre-emptible *e2-standard-16* en Google Cloud Platform (GCP), ambos recursos provistos por el programa Google TPU Research Cloud. Nuestro código está basado en la biblioteca *huggingface's transformers* [168] y su implementación de *RoBERTa*. La Tabla 7.3 muestra los resultados del entrenamiento en términos de pérdida de entropía cruzada y perplejidad.

7.2.3. Evaluación

Para analizar la performance de este modelo, usamos un conjunto de tareas sobre textos generados en redes sociales en español, siguiendo lo hecho en Gonzalez et al. [62] y Polignano et al. [133]. Las tareas elegidas son todas las que analizamos en

Model	Train loss	Eval loss	Eval ppl
cased	1,864	1,753	5,772
uncased	1,940	1,834	6,259
deacc	1,951	1,826	6,209

Tab. 7.3: Resultados del pre-entrenamiento para las tres versiones de *RoBERTuito*. La función de costo utilizada es la entropía cruzada de la tarea de MLM.

esta tesis hasta el momento:

1. Análisis de sentimientos (Capítulo 3)
2. Análisis de emociones (Capítulo 3)
3. Detección de ironía (Capítulo 3)
4. Detección de discurso de odio (Capítulo 4)
5. Detección contextualizada de discurso de odio (Capítulo 6)

Para más detalles sobre los conjuntos de datos y cuestiones puntuales de cada tarea, referimos a los capítulos correspondientes. Comparamos el rendimiento de *RoBERTuito* contra los siguientes modelos de lenguaje pre-entrenados disponibles en español:

- *BETO* [31], tanto en versión cased como uncased.
- *RoBERTa$_{ES}$* (o RoBERTa$_{BNE}$) [69], un modelo RoBERTa entrenado sobre una base de datos de 500GB de todos los sitios *.es*
- *BERTin*[4], otro modelo RoBERTa entrenado en el contexto de un evento de la comunidad Flax/Jax [5], en el cual los autores exploraron diferentes estrategias de muestreo para entrenar este modelo en relativamente poco tiempo sobre la sección en español del corpus *mc4*, creado para entrenar T5 [136].

Cada uno de estos modelos comparte una arquitectura similar a *RoBERTuito* y una cantidad comparable de parámetros. Seguimos las prácticas estándares para el ajuste de los modelos, descriptas en anteriores capítulos y en Devlin et al. [45]. Para las tareas de clasificación, ajustamos los modelos para por 5 epochs con un learning rate triangular de $5 * 10^{-5}$ y un warmup de 10 % de los pasos de entrenamiento. Seleccionamos el modelo que mejor resultado obtuvo al final de cada epoch según la métrica de cada tarea.

7.2.4. Resultados

La Tabla 7.4 muestra los resultados de la evaluación de los modelos seleccionados para las cinco tareas de clasificación propuestas, expresados como la media ± desviación de diez ejecuciones de los experimentos. Podemos observar que en la mayoría

[4] `https://huggingface.co/bertin-project/bertin-roberta-base-spanish`

[5] `https://discuss.huggingface.co/t/open-to-the-community-community-week-using-jax-flax-for-nlp-cv/7104`

Modelo	CONTEX	ODIO	SENTIM	EMOCIÓN	IRONÍA
RoBERTuito$_U$	**59,3 ± 0,4**	**80,1 ± 1,0**	**70,7 ± 0,4**	**55,1 ± 1,1**	73,6 ± 0,8
RoBERTuito$_D$	**59,3 ± 0,6**	79,8 ± 0,8	70,2 ± 0,4	54,3 ± 1,5	**74,0 ± 0,6**
RoBERTuito$_C$	59,0 ± 0,5	79,0 ± 1,2	70,1 ± 1,2	51,9 ± 3,2	71,9 ± 2,3
RoBERTa$_{ES}$	57,7 ± 0,4	76,6 ± 1,5	66,9 ± 0,6	53,3 ± 1,1	72,3 ± 1,7
BERTin	55,7 ± 0,8	76,7 ± 0,5	66,5 ± 0,3	51,8 ± 1,2	71,6 ± 0,8
BETO$_U$	59,1 ± 0,6	75,7 ± 1,2	64,9 ± 0,5	52,1 ± 0,6	70,2 ± 0,8
BETO$_C$	58,2 ± 0,7	76,8 ± 1,2	66,5 ± 0,4	52,1 ± 1,2	70,6 ± 0,7

Tab. 7.4: Resultados de los experimentos de clasificación sobre el benchmark de tareas sociales. CONTEX es la tarea de detección contextualizada de discurso de odio, ODIO es detección de discurso de odio, SENTIM, EMOCIÓN E IRONÍA son análisis de sentimiento, emociones e ironía. Resultado expresado en porcentaje de la métrica correspondiente a cada tarea y como la media ± desviación de diez corridas de los experimentos. U, C, y D significan *uncased, cased y deacc* respectivamente. Más grande es mejor.

de los casos, las tres configuraciones de *RoBERTuito* obtienen resultados por encima de los otros modelos, en particular para las tareas de discurso de odio y análisis de sentimiento. El único caso donde esto no ocurre es en la tarea de detección de discurso de odio contextualizado, donde si bien hay una mejora, es marginal y no significativa.

Analizamos las diferencias entre los tres modelos de *RoBERTuito* mediante un test de Kruskal-Wallis [87] para las performances de cada tarea. Los resultados muestran diferencias significativas entre el desempeño de los tres modelos de *RoBERTuito* para todas las tareas analizadas ($H(3) = 6{,}88, p < 0{,}05$ para Discurso de odio, $H(3) = 9{,}90, p < 0{,}01$ para Análisis de sentimiento, $H(3) = 11{,}85, p < 0{,}01$ para Análisis de emociones, $H(3) = 11{,}54, p < 0{,}01$ para Detección de ironía), con la excepción de la tarea de discurso de odio contextualizado ($H(3) = 3{,}59, p > 0{,}15$).

Para verificar las diferencias significativas entre las performances de los tres modelos para las 4 tareas mencionadas, realizamos un análisis post-hoc con un test de Dunn (con corrección de Benjamini-Hochberg). Exceptuando la tarea de análisis de sentimientos, la versión *cased* muestra siempre diferencias significativas contra las versiones *uncased* o *deacc*. Sin embargo, no se encuentran diferencias significativas entre las versiones *uncased* y *deacc*.

Este resultado puede leerse de dos maneras: primero, que una normalización más fuerte (remover las tildes) del texto de entrada en español no produce una mejora significativa en el rendimiento de los modelos; también, que mantener las tildes en el texto de entrada no es beneficioso ni perjudicial para el rendimiento del modelo.

La Figura 7.1 muestra la distribución del número de tokens en el texto de entrada agrupados por tarea. Podemos observar que los modelos de *RoBERTuito* tienen representaciones más compactas que *BETO* y *RoBERTa-BNE*. *BERTin*, a pesar de su menor rendimiento en general, tiene un tamaño medio comparable al de nuestro modelo. Entre los modelos de *RoBERTuito*, podemos observar que la versión **deacc** tiene una longitud media ligeramente menor en comparación con la versión **uncased**. En suma, esto indicaría que los modelos de *RoBERTuito* logran codificar de manera más eficiente los tweets de los distintos conjuntos de datos considerados.

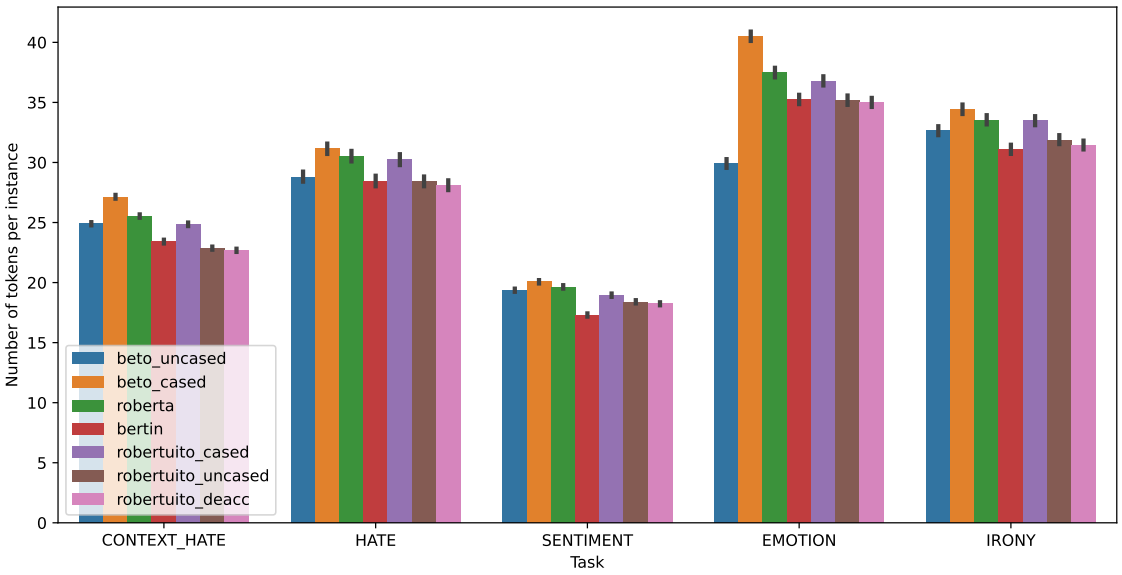


Fig. 7.1: Distribución de la cantidad de tokens por instancia para los tokenizadores de cada modelo. Las barras están agrupadas por tarea y muestran la media de la distribución junto a su intervalo de confianza a 95 %. Más chico es mejor.

7.3. Adaptación de modelos pre-entrenados

Acabamos de observar que entrenar un modelo desde cero sobre tweets resulta en una mejor performance para un conjunto de tareas de dicho dominio. Cabe preguntarse **¿puede ser esta mejora replicada continuando el pre-entrenamiento de otro modelo de lenguaje?** Esta pregunta tiene –más allá del interés teórico de si un modelo de lenguaje entrenado en un dominio distinto puede adaptarse con éxito a un dominio diferente– dos consideraciones prácticas:

- Para lenguajes de recursos relativamente bajos o bien laboratorios menos favorecidos, entrenar un modelo desde cero como realizamos en la anterior sección puede ser prohibitivo en términos económicos.
- Reducir los inmensos costos computacionales de los modelos actuales de NLP puede ser de interés, no sólo en términos económicos sino también ambientales [16].

Para analizar si es posible replicar el rendimiento de un modelo especialmente diseñado para un dominio, continuamos el pre-entrenamiento de un modelo *BETO* sobre textos generados por usuario –tweets para nuestro caso– y probamos su performance sobre el benchmark de tareas sociales descripto en la sección anterior.

7.3.1. Metodología

Para realizar la adaptación de dominio, seguimos las recomendaciones de Gururangan et al. [68], tomando un gran conjunto de datos no anotado de textos sociales y corriendo la tarea de Masked Language Modeling sobre estos. En términos de ese

trabajo, realizamos una forma de *Domain Adaptation Pre-training* (*DAPT*), que consta de usar un conjunto de datos grande y relacionado a nuestra tarea final. Utilizamos para esto los mismos datos recolectados para entrenar a *RoBERTuito*, descriptos en la Sección 7.2.1.

Usamos como modelos base las versiones *cased* y *uncased* de BETO, y continuamos el pre-entrenamiento sobre estos datos, descartando la tarea NSP. En lugar de correr por 12,500 pasos de optimización como es sugerido en Gururangan et al. [68], optimizamos este hiperparámetro probando con 2,500, 5,000, 10,000 y 20,000 pasos de optimización. Para cada modelo nos quedamos con la configuración que obtuvo el mejor resultado en términos del benchmark analizado.

Para entrenar estos modelos, usamos una TPU v2-8, donde cada paso de optimización tomó alrededor de 2,5 segundos. Usamos una configuración similar a la descripta para el entrenamiento de *RoBERTuito* (ver Tabla 7.2), con un learning rate levemente superior ($5 * 10^{-4}$) y limitando también la longitud de secuencia a 128 tokens.

7.3.2. Resultados

En la Figura 7.2 se ilustra el desempeño sobre el benchmark de tareas para los modelos de lenguaje *BETO* y *RoBERTuito*, así como también para las versiones con ajuste de dominio de *BETO*. Para ambos casos, aumentar el pre-entrenamiento pareciera coincidir con una mejor performance, aunque dentro de los modelos *uncased*, el que fue optimizado por 5,000 pasos pareciera haber empeorado su rendimiento general. Esta tendencia, sin embargo, no se cumple en el caso de la tarea de detección contextualizada de discurso de odio. Una posible razón detrás de esto es que la tarea planteada tiene diferencias con el dominio sobre el cual ajustamos: utilizamos pares de tweets, uno de ellos (el contexto) proveniente de un medio periodístico. También puede argumentarse que el dominio de tweets en general es demasiado amplio [46]: para el caso de nuestra tarea, el pre-entrenamiento sobre datos generados de la distribución general de tweets no pareciera ser beneficioso.

Modelo	CONTEX	ODIO	SENTIM	EMOCIÓN	IRONÍA
RoBERTuito$_U$	$59{,}3 \pm 0{,}4$	$80{,}1 \pm 1{,}0$	$70{,}7 \pm 0{,}4$	$55{,}1 \pm 1{,}1$	$73{,}6 \pm 0{,}8$
RoBERTuito$_D$	$59{,}3 \pm 0{,}6$	$79{,}8 \pm 0{,}8$	$70{,}2 \pm 0{,}4$	$54{,}3 \pm 1{,}5$	$74{,}0 \pm 0{,}6$
RoBERTuito$_C$	$59{,}0 \pm 0{,}5$	$79{,}0 \pm 1{,}2$	$70{,}1 \pm 1{,}2$	$51{,}9 \pm 3{,}2$	$71{,}9 \pm 2{,}3$
BETO$_U$+FT	$58{,}8 \pm 0{,}3$	$77{,}5 \pm 1{,}5$	$68{,}0 \pm 0{,}4$	$55{,}3 \pm 0{,}9$	$71{,}7 \pm 0{,}5$
BETO$_C$+FT	$57{,}2 \pm 0{,}6$	$77{,}7 \pm 0{,}9$	$68{,}6 \pm 0{,}5$	$51{,}7 \pm 0{,}9$	$73{,}0 \pm 0{,}4$
BETO$_U$	$59{,}1 \pm 0{,}6$	$75{,}7 \pm 1{,}2$	$64{,}9 \pm 0{,}5$	$52{,}1 \pm 0{,}6$	$70{,}2 \pm 0{,}8$
BETO$_C$	$58{,}2 \pm 0{,}7$	$76{,}8 \pm 1{,}2$	$66{,}5 \pm 0{,}4$	$52{,}1 \pm 1{,}2$	$70{,}6 \pm 0{,}7$

Tab. 7.5: Resultados de la evaluación de modelos pre-entrenados y modelos ajustados en dominio para el benchmark de tareas sociales: CONTEXT es contextualized hate speech, HATE es hate speech detection sobre el dataset de hatEval, SENTIMENT, EMOTION e IRONY son análisis de sentimiento, emociones e ironía sobre los corpus de TASS. Todos los scores son Macro F1s. beto-cased-ft y beto-uncased-ft son modelos adaptados al dominio sociall. Score es la media de cada fila.

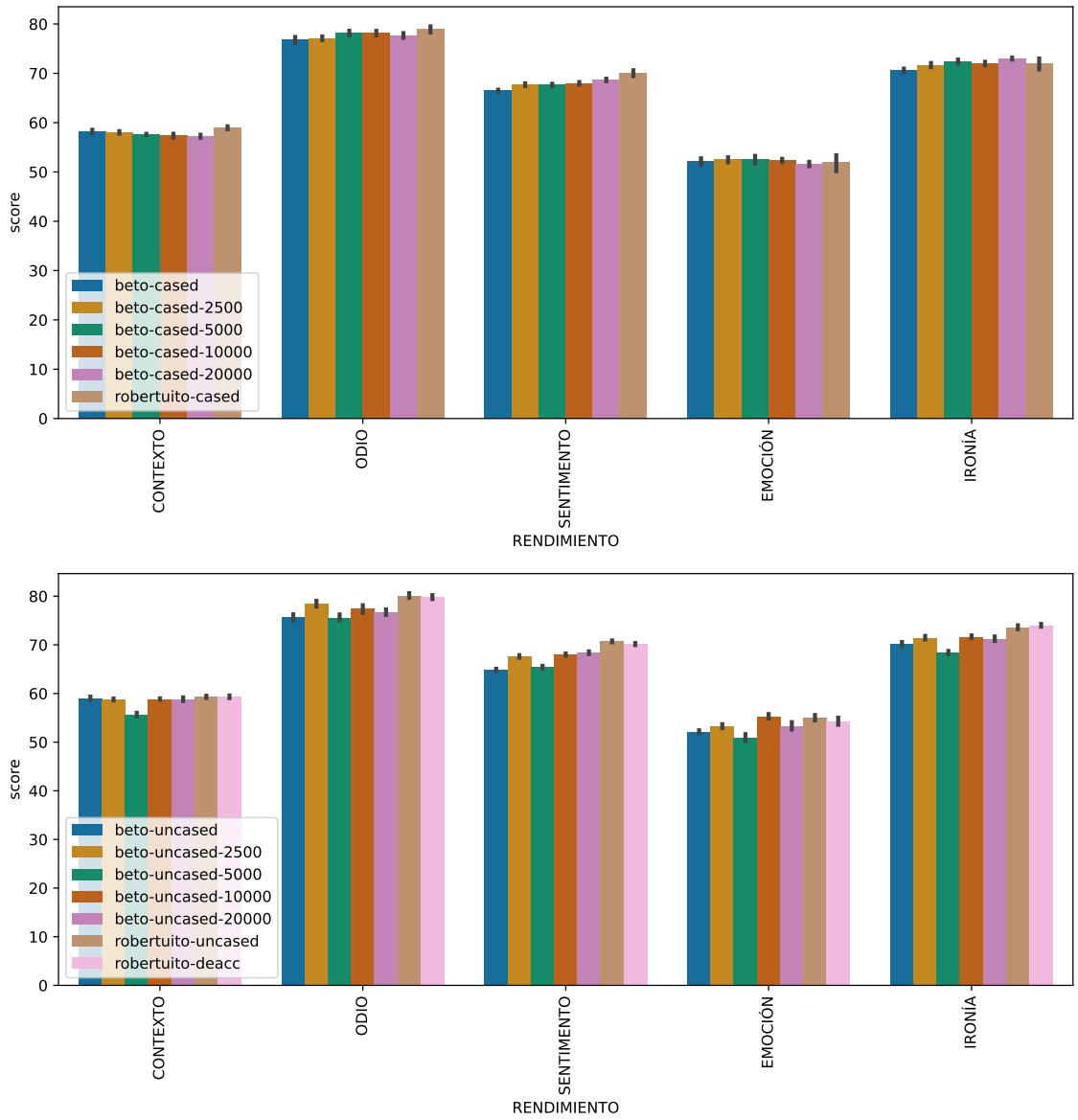


Fig. 7.2: Resultados sobre el benchmark de los modelos BETO y *RoBERTuito* (en versiones cased y uncased). Las barras están agrupadas por tarea y muestran la media de la performance sobre 15 corridas, junto a su intervalo 95 %. El número al final de los modelos indica la cantidad de pasos de optimización realizados en el ajuste de dominio. En tonos azules las variantes de *RoBERTuito*. Más grande es mejor

	Modelos					
	BETO		*RoBERTuito*$_U$		*RoBERTuito*$_C$	
Métrica	¬FT	FT	¬FT	FT	¬FT	FT
LLAMA	63,80	68,47	67,19	69,74	66,63	**70**,**12**
MUJER	41,07	42,06	42,99	**44**,**29**	41,08	43,68
LGBTI	45,13	48,23	48,45	47,94	45,56	**51**,**36**
RACISMO	68,79	**72**,**05**	67,70	69,78	66,60	68,94
CLASE	49,13	**51**,**15**	46,96	47,06	47,01	48,65
POLITICA	57,90	62,48	63,00	62,87	62,19	**63**,**16**
DISCAPACIDAD	58,49	**60**,**89**	57,43	60,64	58,00	59,01
APARIENCIA	74,13	76,63	73,83	74,74	74,71	**76**,**68**
CRIMINAL	65,03	**69**,**94**	62,61	66,63	61,28	65,65
Macro F1	58,16	**61**,**32**	58,91	60,41	58,12	60,80
Macro Precision	64,16	**70**,**21**	60,11	61,63	60,15	62,30
Macro Recall	53,97	55,09	58,38	59,62	56,91	**59**,**80**

Tab. 7.6: Resultados de los experimentos de clasificación para la tarea *granular* de detección de discurso de odio, expresados en porcentajes de F1 para las características. Cada modelo consume el comentario analizado y el tweet de la noticia. Consideramos tres modelos: *BETO*, *RoBERTuito*$_U$ (uncased) y *RoBERTuito*$_C$ (cased). Para cada uno, tenemos dos versiones: una sin ajustar al dominio (¬FT) y ajustada a dominio (FT) de acuerdo a lo descripto en la Sección 6.3. En negrita, los mejores resultados.

La Tabla 7.5 muestra los resultados, expresados nuevamente como medias y desviaciones estándar, para los modelos considerados. Seleccionamos como modelos ajustados a dominio (indicados con $+FT$) a los que obtuvieron mejores resultados entre aquellos que entrenamos con distinta cantidad de pasos de optimización: 10,000 pasos de optimización para *BETO* versión uncased, y 20,000 para versión cased [6]. Haciendo una comparación entre los modelos *BETO* y *RoBERTuito* en versiones *uncased*, vemos para la tarea de detección de discurso de odio que la brecha es de alrededor de 4,4 puntos de F1, mientras que la versión FT achica esa diferencia a 2,26 puntos F1, una reducción del 48 % de la diferencia de performance. En el caso de análisis de sentimiento, pasamos de un gap de 5,8 puntos de F1 a uno de 2,7, una reducción del 53 %; para análisis de emociones, esta diferencia pasa de 3 puntos de F1 a 0, logrando de hecho mejores resultados. Finalmente, en el caso de detección de ironía, el gap de 3,4 puntos de F1 pasa a 1,9, una reducción de 44 %.

7.4. Adaptación de dominio para detección contextualizada de discurso de odio

Para la tarea de detección contextualizada de discurso de odio, el rendimiento de los clasificadores basados en *BETO* se deteriora al realizar la adaptación de dominio sobre tweets en solitario. Sin embargo, los resultados de la Sección 6.4 indicaron una mejora al continuar el pre-entrenamiento sobre pares de tweets que contengan texto y contexto. Una posible conclusión de ello es que el dominio de nuestro conjunto de

[6] Incluimos en el Apéndice D la tabla completa de resultados para cada uno de los pasos

datos construido en el Capítulo 5 es suficientemente distinto en sus instancias al de las demás tareas consideradas.

Teniendo eso en cuenta, efectuamos experimentos de clasificación complementarias realizando una adaptación a dominio de *RoBERTuito*. Efectuamos el mismo procedimiento descripto en la Sección 6.3.1 sobre las versiones *cased* y *uncased* de nuestro modelo, y corrimos el experimento sobre la tarea granular considerando únicamente el contexto simple (sólo el tweet de la noticia).

La Tabla 7.6 contiene los resultados para cada característica y modelo, tanto en sus versiones no ajustadas a dominio (¬FT) como aquellas que sí fueron ajustadas (FT). Mientras que *RoBERTuito* obtiene mejores resultados en la columna ¬FT, al atravesar el proceso de adaptación de dominio, *BETO* obtiene mejores resultados que nuestro modelo. Para la versión uncased, algunas características empeoran su rendimiento luego de la adaptación, aunque esto no ocurre para la versión cased.

7.5. Discusión

En primer lugar, el modelo pre-entrenado sobre tweets construido en este capítulo (*RoBERTuito*) presenta mejoras significativas para casi todas las tareas analizadas en español, con picos de hasta casi 4 puntos de F1 para la tarea de Análisis de Sentimientos contra la mejor versión de *BETO*. De los distintos tipos de normalización de texto utilizados en *RoBERTuito* (*cased*, *uncased* y *deacc*), podemos observar que los modelos *uncased* y *deacc* obtienen mejores performances en general. Si bien el modelo *uncased* muestra una ligera performance superior al modelo *deacc*, esta mejora no es significativa, y esto indicaría que remover tildes no reporta una degradación de la performance. Esto es esperable ya que usualmente no se utiliza de manera consistente esta marcación en el español "vulgar" de las redes sociales, aunque, de todas formas, pruebas más extensivas son necesarias sobre otras tareas y modelos para verificar esta hipótesis.

Una de las limitaciones de esta comparación es que *RoBERTuito* fue entrenado sólo por 600 mil pasos de optimización, contra los casi 900 mil pasos de *BETO*, y el millón de pasos de *BERTweet*. Hay que observar que la optimización de *BETO* se da con un batch size menor (512 vs 4,096) y la de *BERTweet* se hace con un batch size mayor (∼7,000). En términos de cantidad de tokens procesados en el pre-entrenamiento, la comparación puede no ser del todo justa, aunque de todas formas ilustra que el pre-entrenamiento sobre este dominio particular es efectivo.

Con respecto a los experimentos de adaptación de dominio, podemos observar que adaptando *BETO* sobre un conjunto de tweets en español obtenemos una mejora en todos las tareas contra la versión base. Comparado con *RoBERTuito*, las versiones adaptadas logran recortar alrededor del 50 % de la brecha de rendimiento entre ambos, incluso reduciéndolo a cero en algunos casos. Esta comparación, sin embargo, no es del todo justa ya que *BETO* fue pre-entrenado de una manera distinta que *RoBERTuito*. Por cuestiones de tiempo no pudieron ser realizados sobre la versión de *RoBERTa* en español[7] pero trabajo futuro debería usar como base este modelo. Otra opción que no tuvimos en cuenta en este trabajo y que podría reducir más la

[7] Principalmente, ya que éste modelo y *bertin* fueron lanzados mientras realizábamos estos experimentos

diferencia es la de agregar vocabulario en el ajuste de dominio, algo que Howard and Ruder [77] realizan en su implementación de ULM-FIT.

Una consideración práctica de ajustar los modelos de lenguaje es que esta técnica permite mejorar el desempeño de una manera relativamente económica, sin tener que efectuar un costoso pre-entrenamiento desde cero. En términos concretos, un ajuste de dominio puede realizarse utilizando una placa de GPU en uno o dos días, mientras que pre-entrenar un modelo desde cero requiere acceso a un hardware más oneroso. Algunos trabajos recientes [79] muestran alternativas para construir desde cero modelos de lenguaje basados en Transformers bajo escenarios de recursos reducidos, ajustando varios hiperparámetros y usando algunas técnicas de optimización como LAMB [172]. Los recursos mencionados en esos trabajos, lamentablemente, están lejos del alcance de los disponibles de laboratorios no tan favorecidos. En este contexto, continuar el pre-entrenamiento para un dominio particular aparece como una alternativa mucho más factible.

Una pequeña discusión aparte merece la tarea de detección contextualizada introducida en el Capítulo 6. La mejora al utilizar *RoBERTuito* en esta tarea no pareciera ser significativa, a diferencia de los demás casos. Esto puede deberse a que las instancias de este conjunto de datos tienen una estructura bastante diferente de la de las demás: cada una consta de dos tweets, donde el contexto suele ser un titular de diarios formulado como un tweet, y el comentario en cuestión. El titular de un diario tiene una forma mucho más cercana al dominio del pre-entrenamiento de *BETO*, mitigando una de las posibles mejoras de *RoBERTuito*. Más aún, observamos que realizar un ajuste de dominio sobre tweets aislados empeora el rendimiento sobre esta tarea; haciendo este ajuste de la misma manera que en el capítulo anterior, tampoco se logró obtener mejores resultados que con *BETO*.

7.6. Conclusiones

En este capítulo hemos abordado la tarea de mejorar el rendimiento de la detección de discurso de odio en el contexto más general de tareas de clasificación sobre textos sociales en español. Para ello, utilizamos como benchmark varias de las tareas que vimos en esta tesis: detección de discurso de odio (en sus dos versiones, no contextualizada y contextualizada), análisis de sentimiento, análisis de emociones, y detección de ironía.

En primer lugar, y en la corriente de modelos pre-entrenados sobre distintos dominios, generamos un nuevo y valioso recurso para la clasificación de textos sociales: *RoBERTuito*, un modelo de lenguaje basado en *RoBERTa* sobre tweets en español. Para ello, recolectamos un gran corpus de tweets en español, y utilizando las TPU provistas por Google realizamos el pre-entrenamiento de este modelo. Los experimentos de clasificación sobre el conjunto de tareas arrojaron que *RoBERTuito* obtiene mejores significativas sobre otros modelos en español. Así mismo, observamos que remover tildes en el preprocesado no reporta una degradación significativa en la performance.

Por otro lado, exploramos una técnica de ajuste de dominio sobre modelos actuales para comparar la ganancia de rendimiento y compararla contra *RoBERTuito*. Para ello, tomamos los modelos de *BETO* (en sus versiones cased y uncased) y co-

rrimos la tarea de MLM sobre los tweets recolectados para entrenar nuestro modelo anterior. Si bien la performance de estos modelos ajustados mejora con respecto de *BETO*, se mantiene por debajo de *RoBERTuito*, aunque recortando considerablemente la brecha de performance entre ambos modelos. De todas formas, este análisis puede ser de consideración para aquellos lenguajes con menos recursos que no pueden pre-entrenar modelos de lenguaje desde cero.

Resta como trabajo futuro realizar estos experimentos sobre tareas más desafiantes, y también realizar los ajustes sobre los modelos *RoBERTa* en español para hacer una comparación más justa. Así mismo, explorar otras alternativas de mejora sobre la tarea de detección contextualizada de discurso de odio, elusiva para ambas técnicas consideradas.

Todos estos experimentos han sido realizados en español, y sus recursos publicados. El modelo puede ser encontrado en el hub de Huggingface [8], como así también el código para entrenarlo y para correr el benchmark con otros modelos pre-entrenados [9], y en un futuro la base de datos de tweets en español.

7.7. Notas

En Pérez et al. [127] puede encontrarse la descripción de la construcción de *RoBERTuito*. En el Apéndice D puede encontrarse la tabla completa de resultados para las distintas cantidades de pasos de optimización, junto a experimentos adicionales sobre las habilidades multilinguales de *RoBERTuito*.

[8] `https://huggingface.co/pysentimiento/robertuito-base-uncased`
[9] Ambos en [10]

CONCLUSIONES

8. CONCLUSIONES

En esta tesis, hemos abordado la tarea de la detección de discurso de odio, intentando hacer avanzar el estado del arte basado mayormente en la clasificación binaria de este fenómeno sobre comentarios de usuarios en redes sociales. En ese sentido, propusimos una extensión de la tarea agregándole un marco contextual a cada instancia analizada. Esta información es usualmente descartada en la literatura del tema, que ha estado centrada en el análisis de comentarios aislados. Para estudiar esta extensión, construimos un conjunto de datos de respuestas de usuarios a artículos periodísticos argentinos en la red social Twitter. Luego de hacer una reseña del trabajo previo, tuvimos el cuidado de construir este recurso de manera interdisciplinaria –en la frontera del derecho, la sociología y el procesamiento del lenguaje natural– y teniendo en cuenta el componente cultural del discurso de odio, evitando recaer en el etiquetado mediante plataformas de terceros que limitan esta posibilidad.

En base a los experimentos de clasificación realizados sobre este conjunto de datos, hemos podido brindar cierta evidencia de que el contexto –en este caso, en forma de tweet de medio periodístico– puede aprovecharse para identificar discursos de odio mejorando el rendimiento de clasificadores basados en técnicas del estado del arte. Si bien no se observó lo mismo al utilizar un contexto más largo –el artículo periodístico completo– trabajo futuro debería explorar si existen formas útiles de incorporar esta información al clasificador. De manera heurística, podríamos argumentar que los humanos tenemos acceso a contextos muchos más ricos, accediendo por otras vías a información sobre la noticia y el comentario en cuestión, incorporando conocimiento del mundo real –como por ejemplo, si la persona sobre la que habla la noticia posee cierta característica protegida no mencionada. Los clasificadores del estado del arte carecen de esta información, con lo cual una posible línea de investigación puede ser la de incorporar este conocimiento dentro de los algoritmos.

Este resultado –sobre los beneficios de utilizar información contextual– va en línea con algunos trabajos recientes en NLP que muestran que la utilización de más de una fuente de información puede ser beneficiosa para ciertas tareas. Esto es algo esperable ya que nuestro entendimiento dista de ser descontextualizado –sólo sobre un comentario o un texto aislado– sino que incorpora diversos elementos: desde el tópico de la conversación, quiénes son los interlocutores, conocimiento externo, entre otras cuestiones.

Un punto adicional que observamos es que la predicción de múltiples características –además de la mera existencia del discurso de odio– no sólo no empeora el rendimiento de los algoritmos sino que lo mejora parcialmente para las técnicas del estado del arte. Es decir, si en vez de sólo predecir que existe o no discurso de odio predecimos más características –como ser la o las características ofendidas, si existe un llamado a la acción violenta, si está dirigido a un grupo o un individuo– podemos, por un lado, mejorar la interpretabilidad y la riqueza de la salida de los algoritmos de detección; y por otro, mejorar su rendimiento al entrenarlos sobre una señal más rica. De esto, se desprende que es conveniente generar recursos que

contengan anotaciones más detalladas –no sólo la etiqueta binaria– y en lo posible marcando las características protegidas que se estén vulnerando en cada instancia.

Respecto a las limitaciones de este trabajo y sus conclusiones, una cuestión particular es que el conjunto de datos mencionado fue etiquetado considerando en todo momento el contexto del comentario. Una comparación justa entre algoritmos que incorporen contexto contra algoritmos que no lo hagan debería incluir un conjunto de entrenamiento etiquetado de manera descontextualizada, para así poder ajustar los clasificadores de acuerdo a estos dos tipos de anotaciones.

Una limitación más general –ya no sólo de los resultados de esta tesis sino del área del procesamiento del lenguaje natural– versa sobre el actual estado del arte, basado en modelos de lenguajes neuronales como BERT, GPT, y compañía. Si bien es innegable el avance que han supuesto estos modelos pre-entrenados, habiendo logrado resultados superadores en casi toda tarea de NLP, no podemos dejar de observar que, en términos de lo mencionado por Judea Pearl [122], estos sistemas aún están en una etapa meramente asociacional. Muy a pesar de que muchos trabajos hablen sobre cierto entendimiento por parte de estos algoritmos (por ejemplo, el benchmark General Language Understanding Evaluation, GLUE), estos sólo detectan regularidades en los datos, sin efectuar ningún razonamiento sobre ellos.

Para nuestro problema concreto, un clasificador puede detectar que decirle "sos hombre" a un artículo relacionado a una mujer (quizás trans) conlleva discurso de odio contra la comunidad LGBTI. Sin embargo, este mismo mensaje ofuscado de alguna manera (por ejemplo, preguntándole el nombre, pidiéndole el documento, o alguna otra forma que no hayamos observado en los datos) logra burlar a nuestros sistemas ya que están exclusivamente basados en detectar regularidades y no pueden efectuar ningún razonamiento simbólico entre la equivalencia de estos mensajes. Ligado a este ejemplo, algunos trabajos han puesto en duda los avances recientes en el área, indicando que los actuales algoritmos basados en modelos de lenguaje –aún en sus formas más complejas y sobreparametrizadas con miles de millones de parámetros– no son más que "loros estocásticos", muy hábiles en detectar regularidades y hacernos creer que llevan adentro algún tipo de entendimiento [15, 16]. Sin embargo, el entrenamiento sobre la mera forma del lenguaje –Terabytes de texto no etiquetado– no conlleva ningún tipo de comprensión sino sólo un aprendizaje sobre su distribución.

Volviendo al fenómeno estudiado en esta tesis, aún cuando gran parte del discurso de odio se expresa en forma de insultos o expresiones ofensivas muy características que son detectables por los algoritmos actuales, hay un subconjunto de estos mensajes que necesitan conocimiento del mundo real, de la relación entre interlocutores, y muchas veces de realizar algún tipo de razonamiento. Para los enfoques actuales de NLP, este tipo de deducciones están fuera de alcance, y hemos observado en el análisis de error algunos casos donde los clasificadores fallaban en la detección de comentarios abiertamente discriminatorios pero que guardaban algún tipo de razonamiento, metáfora, o dificultad adicional. Parafraseando a Mitchell [110], podemos decir que **la detección de discurso discriminatorio es más difícil de lo que creemos**.

¿Significa esto que los algoritmos actuales no son de ningún uso? En absoluto. Los actuales algoritmos de detección, aún con sus defectos y siendo bastante rudi-

mentarios, logran captar parte del lenguaje discriminatorio que observamos en redes sociales. Sin embargo, es necesario entender sus limitaciones: a medida que estos sistemas puedan encontrar regularidades con más detalle –y potencialmente sean usados para moderar este tipo de agresiones– muchos usuarios expresarán este tipo de ofensas de manera más sofisticada para lograr esquivar su escrutinio, apelando a metáforas, mensajes indirectos, multimodalidad, etc. Asimismo, un punto no explorado en este trabajo y que limita la aplicación para casos graves de discursos de odio es la limitada interpretabilidad de los algoritmos basados en redes neuronales, un mal endémico a toda el área de Inteligencia Artificial. Teniendo estas cuestiones en mente, planteamos que agregar más contexto e información del mundo real a nuestros algoritmos puede ayudarlos a mitigar parcialmente sus limitaciones.

Para cerrar, un eje que atraviesa este trabajo es que fue realizado íntegramente en español y atendiendo la realidad sociocultural de Argentina. La inmensa mayoría de la literatura sobre este tema es en inglés, algo que está muy alejado no sólo desde lo lingüístico sino también desde el plano cultural. Los discursos de odio están situados dentro de las realidades sociales de cada región, por lo que es necesario estudiarlos atendiendo las distintas problemáticas características y no sólo considerando la variable idioma. Como consecuencia de esto, esta tesis intenta aportar a balancear la asimetría de recursos tanto en el área particular y específica de detección de discurso de odio como así también en la de NLP en general.

Apéndice

A. DISCURSO DE ODIO

A.1. Tratados internacional sobre libertad de expresión y discurso de odio

A.1.1. Libertad de expresión

La Convención Americana de Derechos Humanos (CADH) establece que:

> 1. Toda persona tiene derecho a la libertad de pensamiento y de expresión. Este derecho comprende la libertad de buscar, recibir y difundir informaciones e ideas de toda índole, sin consideración de fronteras, ya sea oralmente, por escrito o en forma impresa o artística, o por cualquier otro procedimiento de su elección.
>
> 2. El ejercicio del derecho previsto en el inciso precedente no puede estar sujeto a previa censura sino a responsabilidades ulteriores, las que deben estar expresamente fijadas por la ley y ser necesarias para asegurar:
>
> a) el respeto a los derechos o a la reputación de los demás, o
>
> b) la protección de la seguridad nacional, el orden público o la salud o la moral públicas. (CADH, Artículo 13)

En Estados Unidos, la primer enmienda protege este derecho humano, mientras que en la Unión Europea, legislación similar ofrece protección a la libertad de expresión. Finalmente, la declaración universal de los derechos humanos de la ONU [1] menciona tanto en su preámbulo como en el artículo 19

> Todo individuo tiene derecho a la libertad de opinión y de expresión; este derecho incluye el de no ser molestado a causa de sus opiniones, el de investigar y recibir informaciones y opiniones, y el de difundirlas, sin limitación de fronteras, por cualquier medio de expresión.ONU (Declaración Universal de los Derechos Humanos)

Otro documento conocido como el Pacto Internacional de Derechos Civiles y Políticos (ICCPR por sus siglas en inglés), sancionado en 1966 en la Asamblea de las Naciones Unidas y ratificado por 166 países, incluye en su artículo 19:

> 1. Nadie podrá ser molestado a causa de sus opiniones.
>
> 2. Toda persona tiene derecho a la libertad de expresión; este derecho comprende la libertad de buscar, recibir y difundir informaciones e ideas de toda índole, sin consideración de fronteras, ya sea oralmente, por escrito o en forma impresa o artística, o por cualquier otro procedimiento de su elección.
>
> 3. El ejercicio del derecho previsto en el párrafo 2 de este artículo entraña deberes y responsabilidades especiales. Por consiguiente, puede estar sujeto a ciertas restricciones, que deberán, sin embargo, estar expresamente fijadas por la ley y ser necesarias para:
>
> a) Asegurar el respeto a los derechos o a la reputación de los demás;
>
> b) La protección de la seguridad nacional, el orden público o la salud o la moral públicas. (Artículo 19 de la ICCPR)

Este último apartado ilustra que la libertad de expresión no es un derecho completamente irrestricto. El ejercicio de los derechos e igualdad ante la ley de otros marca este límite. Citando nuevamente al Pacto de San José de Costa Rica:

[1] `https://www.un.org/es/about-us/universal-declaration-of-human-rights`

Palabra	#tweets
puta	2226
callate	1223
migra	802
perra	640
arabe	483
zorra	367
sudac	355
ceuta	310
acoso	284
polla	272

Tab. A.1: Palabras del dataset y su cantidad de incidencia en tweets del dataset de HatEval.

> 1. Los Estados Partes en esta Convención se comprometen a respetar los derechos y libertades reconocidos en ella y a garantizar su libre y pleno ejercicio a toda persona que esté sujeta a su jurisdicción, sin discriminación alguna por motivos de raza, color, sexo, idioma, religión, opiniones políticas o de cualquier otra índole, origen nacional o social, posición económica, nacimiento o cualquier otra condición social.Artículo 1 (Pacto San José de Costa Rica, CADH)

y a la Declaración Universal de los Derechos Humanos de la ONU:

> Todos los seres humanos nacen libres e iguales en dignidad y derechos y, dotados como están de razón y conciencia, deben comportarse fraternalmente los unos con los otros.

A.1.2. Discurso de odio

Una recopilación de definiciones realizada por un informe de la UNESCO

> Mientras que el sistema interamericano de derechos humanos ha desarollado determinados estándares, no existe una definición universalmente aceptada de "discurso de odio" en el derecho internacional. Según un reciente informe emitido por la UNESCO que estudió las distintas definiciones de discurso de odio en el derecho internacional, el concepto con frecuencia se refiere a "expresiones a favor de la incitación a hacer daño (particularmente a la discriminación, hostilidad o violencia) con base en la identificación de la víctima como perteneciente a determinado grupo social o demográfico. Puede incluir, entre otros, discursos que incitan, amenazan o motivan a cometer actos de violencia. No obstante, para algunos el concepto se extiende también a las expresiones que alimentan un ambiente de prejuicio e intolerancia en el entendido de que tal ambiente puede incentivar la discriminación, hostilidad y ataques violentos dirigidos a ciertas personas (Countering Hate Speech Online, Gagliardone et al. [55])

A.2. Incidencia de keywords en el dataset

La Tabla A.1 muestra un listado de palabras obtenido por observación de instancias del dataset de Basile et al. [10]. Podemos observar que algunas palabras (como *puta, callate, migrante, árabe*) tienen una altísima tasa de incidencia en tweets, dando cuenta de un posible sesgo de recolección de los datos.

B. CONSTRUCCIÓN DE DATASET CONTEXTUALIZADO DE DISCURSO DE ODIO

En este apéndice describiremos algunos pormenores de la construcción del conjunto de datos del Capítulo 5.

B.1. Distribución de datos recolectados

La Figura B.1a muestra la distribución temporal de los artículos, sin aplicar ningún filtro por palabras, mientras que B.1b muestra aquellas relacionadas al COVID-19 utilizando filtrando aquellos que mencionen alguna de las siguientes palabras: *coronavirus, encierro, síntomas, covid, fase, fiebre, cuarentena, infectados, distanciamiento, normalidad, Wuhan, aislamiento*. Podemos observar dos caídas. Hay un pequeño pozo en mayo 2020 que se debió a problemas técnicos de nuestros servidores de recolección. Por otro lado, observamos que algunos medios (particularmente La Nación) parecieran mencionar menos directamente al COVID (al menos con los términos referidos anteriormente) hasta un nuevo pico cerca de fin de año, coincidente con un nuevo rebrote del virus en este país. Sin embargo, estas mediciones pueden contener artefactos del método de filtrado: muchas notas contienen links a otras con sus títulos, pudiendo interferir en estas estimaciones.

B.2. Recursos utilizados

El etiquetado del conjunto de datos consumió alrededor de 450 horas de trabajo entre todos los anotadores. El gasto total de anotación fue de 125,000 pesos argentinos, lo cual al tipo de cambio de ese entonces equivale a alrededor de 1,400 dólares estadounidenses.

B.3. Manual de criterios de anotación

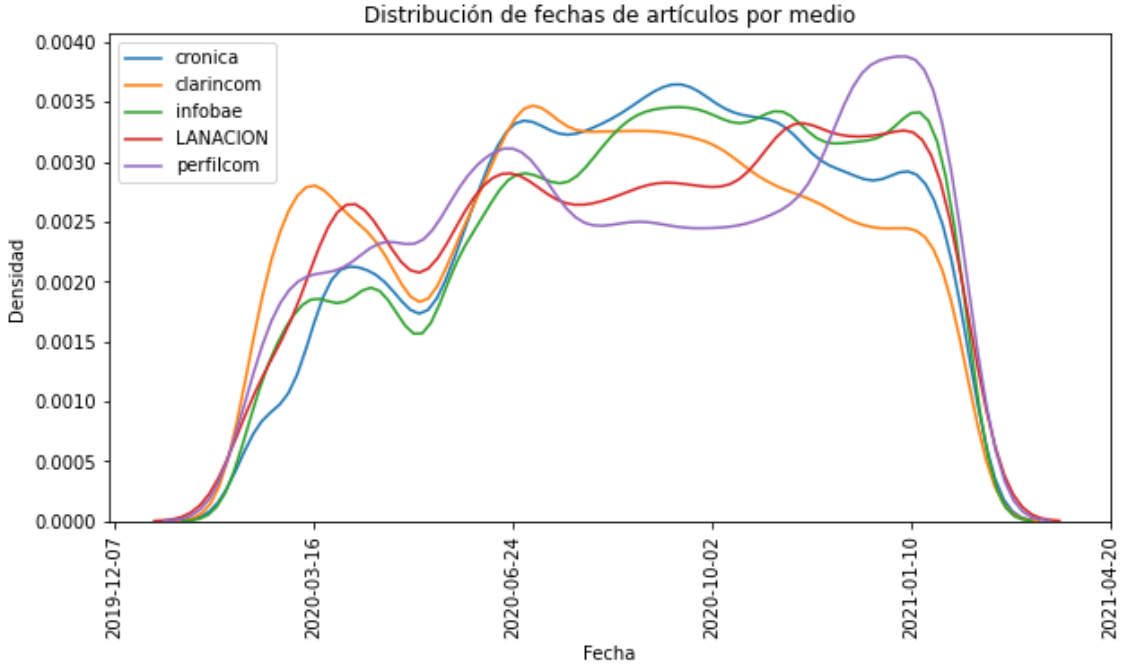
B.3.1. Presencia de lenguaje discriminatorio

Entendemos que hay discurso discriminatorio en el tweet si contiene declaraciones de carácter intenso y posiblemente irracional de rechazo, enemistad y aborrecimiento contra un individuo o contra un grupo, siendo estos objetivos de estas expresiones por poseer (o aparentar poseer) una característica protegida.

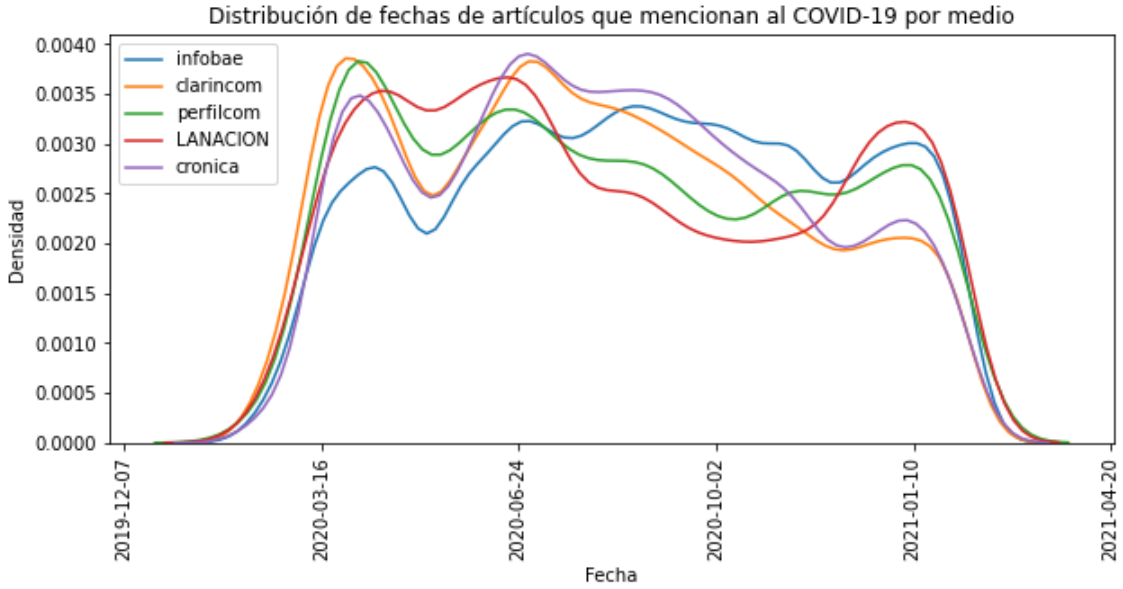
Este discurso puede manifestarse de manera explícita (insultos directos), celebraciones sobre asesinatos u otros crímenes, o bien otras expresiones más veladas. Lo que queremos captar es la intención del autor del tweet. El carácter discriminatorio de un mensaje está dado tanto por el contexto (en este caso, el tweet original del medio periodístico y posiblemente la nota) y el contenido del tweet en sí mismo. Por ejemplo, un comentario que diga “excelente” sin contexto es una cosa, y decir eso mismo en una nota que relata un femicidio, o un asesinato es otra muy distinta.

Las características protegidas que vamos a tener en cuenta son las siguientes:

1. Sexo (Mujeres, concretamente)
2. Género o identidad sexual (Colectivo LGBTI)
3. Ser inmigrantes, extranjeros, pueblos aborígenes u otras nacionalidades (Xenofobia, racismo)
4. Situación socioeconómica o por barrio de residencia
5. Poseer discapacidades, problemas salud mental o de adicción al alcohol u otros estupefacientes



(a) Distribución temporal de artículos recolectados, sin aplicar ningún filtro



(b) Distribución temporal de artículos recolectados que mencionan COVID-19 o algún término relacionado

Fig. B.1: Gráficos de distribución de los datos recolectados.

6. Opinión o ideología política
7. Aspecto o edad (mayormente, gordofobia/gerontofobia)
8. Antecedentes penales o estar privado de la libertad

Es decir, para considerar un mensaje como discriminatorio, debe cumplir que el discurso discriminatorio está orientado hacia un individuo o grupo de al menos una (aunque posiblemente más de una) característica protegida.

Consideramos que el mensaje del tweet (a la vez que el receptor del odio) es el que determina si puede o no ser considerado discriminatorio y hacia qué grupo está dirigido. Esto puede no necesariamente coincidir con el destinatario explícito del mensaje: por ejemplo, si alguien le dice a Susana Giménez "judía sionista hdp", a pesar de no ser Susana Giménez judía, se puede considerar esto como discurso de odio contra las minorías religiosas y/o discurso xenófobo.

B.3.2. Llamado a la acción

Entendemos que un tweet (que contiene discurso discriminatorio) llama a la acción si contiene alguna incitación a tomar algún tipo de medida contra el sujeto o grupo ofendido. Esta medida puede ser de carácter violento ("hay que matarlos ya" "pongámosles una bomba") o de carácter menos violento ("hay que dejar de comprarles a estos chinos ladrones")

Estos tweets nos interesan particularmente porque son los más peligrosos y dañinos: los que llaman a tomar algún tipo de represalia contra la persona o el grupo en cuestión.

B.3.3. Características protegidas

Finalmente, para cada tweet deberemos marcar qué grupo o característica protegida es atacado. En este caso, necesariamente un grupo/característica debe ser seleccionado:

Usaremos una notación abreviada en la interfaz de etiquetado, en la que algunos de los grupos o características mencionadas fueron reagrupadas de la siguiente manera:

1. MUJER: por su sexo
2. LGBTI: por género o identidad sexual
3. RACISMO: Por ser inmigrantes, extranjeros, pueblos aborígenes u otras nacionalidades (Xenofobia, racismo)
4. POBREZA: Por situación socioeconómica o por barrio de residencia.
5. DISCAPAC: Por tener discapacidades, problemas salud mental o de adicciones
6. POLITICA: Por su opinión o ideología política
7. ASPECTO: Por su aspecto o edad
8. CRIMINAL: Por sus antecedentes o situación penal (presos)

A su vez, agregamos la categoría "OTROS". Esta categoría es excepcional, y debería utilizarse sólo si algún tipo de discriminación no está contemplado en estas categorías.

Respecto a la discriminación de carácter político tiene que ser algo más que una mera opinión sino tener una componente irracional, de descalificación y de aborrecimiento considerable sobre un individuo o una facción política.

No se contempla dentro de las categorías protegidas a las profesiones. Es decir, no tenemos en cuenta el discurso contra científicos, médicos, o periodistas; de esto último hay bastante material agresivo en los comentarios.

Lineamientos generales

El discurso discriminatorio no es sólo discurso ofensivo contra una persona o grupo con alguna de las características protegidas. Tiene que apelar a su condición de mujer, inmigrante, LGBTI, etc. para que lo consideremos así.

Por ejemplo: si alguien agrede a una mujer, a un inmigrante, o a alguien de la comunidad LGBTI, no necesariamente está incurriendo en un discurso discriminatorio salvo que apele a algo que remita a su característica como tal.

Expresiones de aprobación ante noticias de crímenes o acciones contra persona o grupo de las características protegidas son consideradas discriminatorias.

Ejemplos

Violan a la reconocida actriz XXXX YYYY Asesinan a un comerciante chino por creer que tenía Coronavirus Motín y muerte en la prisión de Marcos Paz

Comentarios de contenido discriminatorio: (emoji de aplausos) - uno menos - bravo! -

Si no queda claro que haya un mensaje discriminatorio o parece de carácter difuso o demasiado tangencial, entonces etiquetar como no discriminatorio

MUJER

Insultar a una mujer sin hacer ninguna referencia particular a su condición de mujer no es suficiente para ser considerado discurso discriminatorio

Como regla: si el mismo insulto o agresión aplicase contra un hombre, entonces no debiéramos considerarlo como discriminación

Insultos contra las expresiones políticas del movimiento de las mujeres son consideradas en esta categoría: si se las insulta como feminazis, aborteras, pañuelito verde, etc

Apelaciones a su apariencia o aspecto propias de una mujer son consideradas en esta categoría. En este punto consideramos comentarios cosificadores

Insultar como “vieja” a una persona no califica como misoginia. Usar para ese caso la categoría ASPECTO que contempla la gerontofobia

Contexto: Nati Jota furiosa por los comentarios que recibe en las redes.

Comentarios discriminatorios:

- Pero si sos de plástico nena! (opina de manera denigratoria de su apariencia)
- Flor de gato!
- Miauuu!
- A esta sólo se la conoce por su cuerpo y ahora se hace la santa. Andá a estudiar
- Le damos hasta que San Lorenzo vuelva a Boedo
- Y esta rubia tarada quién es?

Ofensivos pero no misóginos:

- Comentarios ofensivos pero no sexistas:
- Callate forra (ofensivo pero no particularmente sexista)
- Y esta quién es? A quién le importa?
- Quién?
- Nati cuánto?
- Esta también recibe sobres?
- Otra descerebrada más (súper agresivo, pero es un comentario que podría recibir un hombre, aunque en menor cantidad…)

- Andá a cagar..boluda..conductora de 4ta..
- HDP
- Boba

LGBTI

Expresiones de disgusto o rechazo del estilo "qué asco" frente a integrantes de la colectividad deben ser marcados como discurso discriminatorio

Apelaciones a la anatomía o del estilo "biologicistas" deben ser marcados como discurso discriminatorio

Contexto: Anibal Pachano sobre la cuarentena: "este virus nos está destruyendo a los actores"

Comentarios discriminatorios:

- Qué asco este sujeto
- Y a este trolo quién le pidió su opinión?
- Me desagrada

Ofensivos pero no discriminatorios:

- Ajá, y?
- No me importa tu opinión, gracias
- Se acuerdan cuando este atacaba a Cristina? Andá a llorarle al gato, Aníbal
- Pelotudazo!

Contexto: Famosa actriz trans dio una entrevista a La NACION: "Como mujer, estoy en la mejor etapa de mi vida"

- Muy lindo todo, pero sos hombre
- "Mujer" dice jajajaj naciste y morirás macho
- Qué asco este trolo
- 🤮
- Una mujer con sorpresa
- Trava horrendo
- Perdón pero qué nombre dice tu DNI?

Ofensivos pero no discriminatorios:

- Quién?
- Y a quién le interesa lo que diga esta mujer?
- Que se vaya a cagar
- Agradecele al gobierno de Néstor que tenés derechos

RACISMO

- Asignar a todo un pueblo con una característica o actividad particular negativa entra en esta categoría
- Cualquier llamada a expulsar, matar, tirar bombas, destruir, aniquilar, etc a un pueblo, inmigrantes o país entra en esta categoría. Más aún, deben ser marcadas como "llamadas a la acción"
- Teorías conspirativas (hasta cierto punto) o acusaciones a un Gobierno/Estado no deberían entrar en esta categoría
- Acusaciones respecto al virus

Contexto: Rescatan en China a cuarenta gatitos bebé que iban a ser utilizados en restaurantes

Ejemplos discriminatorios:

- Chinos asquerosos
- Malparidos! Chinos de mierda
- Sigan desparramando pestes hijos de puta!
- Por qué no se comen entre ellos?
- País horrible y enfermo
- Estos chinos nos diseminan su peste por todo el mundo

Ejemplos discriminatorios y que llaman a la acción:

- Son una plaga estos chinos, hay que matarlos a todos
- Ojalá les caigan veinte bombas
- Hay que barrer ese país de la faz de la tierra

Ejemplos no racistas:

- Son comunistas, qué querés
- Lo rescataron a Mauri?
- El gobierno chino está intentando controlar todo, cuidado
- Mienten con que controlaron el virus, cómo no van a mentir con esto también?
- De terror!

Contexto: Tomas de terrenos: se profundizan las tomas en todo Buenos Aires

Comentarios discriminatorios:

- **Rajen a todos los peruanos de las tomas! (llama a la acción, además)**
- **Increíble, cualquier paraguayo o bolita de mierda te saca la casa**

POBREZA

- Expresiones denigrantes sobre la condición socioeconómica son consideradas discriminatorias: referirse a las personas sin empleo como "vagos", a otros como "(chori)planeros"
- Respecto a la gente que posee planes de asistencia social, si es una opinión general sin ser ofensiva no se considera discriminatoria (del estilo "hay que reducir el gasto social y los planes", "perdieron la cultura del trabajo")

Contexto: Presupuesto: aumentó el gasto en planes asistenciales durante la pandemia
Ejemplos discriminatorios:

- Basta de mantener vagos!
- Cansada de los planeros
- Che laburar estos atorrantes ni en pedo no?
- PARASITOS

Ejemplos no discriminatorios:

- La gente que trabaja y aporta impuestos es cada vez menos. Estamos al horno

POLITICA

- Apreciaciones derogatorias sobre la posición política son consideradas discriminatorias : zurdo/a, bolchevique, peroncho, gorila, kuka, etc
- Acusaciones de corrupción o de "recibir sobres" no son consideradas discriminatorias
- Tampoco aquellas expresiones que traten de inútiles a funcionarios
- Tratar de viejo/a, gordo/a, u otras cuestiones físicas deben ser marcados en las categorías respectivas, no acá

Contexto: Aumentó el gasto en planes asistenciales durante la pandemia
Ejemplos discriminatorios:

- BASTA ZURDOS DE ROBARNOS
- Bolcheviques de mierda

Ejemplos no discriminatorios:

- La gente que trabaja y aporta impuestos es cada vez menos. Estamos al horno
- Qué gobierno de inútiles
- Son unos delincuentes
- Hijos de mil puta!
- Siguen volando los sobres para el Congreso
- Siga siga la impresión

Ejemplos no discriminatorios

ASPECTO

- Apreciaciones denigrantes sobre la apariencia de una persona y/o su edad
- Principalmente, tenemos en mente la gordofobia y gerontofobia, pero puede referir a otras características físicas (por ejemplo, la altura).
- En casos en las cuales haya solapamiento con mujer, marcar ambas

Contexto: Luis Brandoni: "No convoqué el banderazo"
Ejemplos discriminatorios:

- Viejo de mierda!
- Qué decrépito impresentable que es este señor
- Estás gagá, pelotudo

Contexto: Jorge Lanata vuelve a la televisión
Ejemplos discriminatorios:

- Gordo chanta otra vez volvés a vender pescado podrido?
- porque no te vas vos tambien con todos bola de sebo!!!1

CRIMINAL

Cualquier comentario que celebre acciones contra criminales o personas privadas de su libertad (golpizas, asesinatos, muerte en motines, etc) entra en esta categoría. En este ítem muchas veces veremos que son llamados a la acción: el de "matarlos", llamar a reducir sus derechos, etc

Contexto: Enfrentamiento entre policías y ladrones en Recoleta: un ladrón muerto
Ejemplos discriminatorios:

- Uno menos!
- Excelente!
- 👏
- Que pena, pobrecito

Ejemplos que además llaman a la acción:

- MUY BIEN! Felicitaciones al policía, hay que liquidarlos sin piedad

Ejemplos no discriminatorios:

- Cómo puede ser que nuestra Ministra no haga nada?
- La policía actuó correctamente.

Contexto: Motín por el Coronavirus en Olmos: 3 muertos
Ejemplos discriminatorios y que llaman a la acción

- Hay que rociar con nafta todas las cárceles
- Soltemos 3 o 4 infectados con COVID en cada cárcel y problema solucionado
- Paredón y listo

DISCAPACIDAD

Referencias peyorativas de adicciones a drogas, alcohol u otros estupefacientes. También referencias peyorativas a la salud mental de la persona en cuestión. Decir "está loco" no entra acá.

Contexto: Malena Pichot sale a cruzar a Baby Etchecopar
Ejemplos discriminatorios:

- Callate faloperita!
- No probaste dejando la merca sucia?

OTROS

Esta categoría está reservada para cualquier otro tipo de discriminación que no esté contemplada en las categorías mencionadas Insultos a profesiones (científicos, periodistas, por ejemplo) no entran en este apartado ESTA CATEGORIA ES SUMAMENTE EXCEPCIONAL. NO USAR INDISCRIMINADAMENTE

C. DETECCIÓN CONTEXTUALIZADA DE DISCURSO DE ODIO

C.1. Análisis comparativo entre clasificadores granulares y binarios

		LGBTI	
1	FN	Mara Gómez: la historia de la primera futbolista trans en el torneo argentino	Ponga huevos, Mara ponga huevos...
2	FN	Graciana Peñafort: "La marcha me dio mucha pena y tuvo un nivel de convocatoria menor al esperado"	Pena das vos, termotanque de lipídos y déficit fiscal
3	FP	"No soy feminista, soy mujer ", la frase de Viviana Canosa que generó polémica	La condición de mujer no la exime de ser una pelotuda y terrible hija de puta. Asi como tampoco la exime de vivir en un mundo con menos oportunidades, menores remuneraciones y más riesgos x el solo hecho de ser mujer, aunque no quiera luchar x eso.
4	FP	T ras una ráfaga de más de 20 disparos asesinaron a una mujer trans en Rosario	Con que le dispararon? Alta minigun tiene que ser para que 20 tiros sean solo una ráfaga y no una fullauteada
		RACISMO	
1	FN	Es hija de chinos, llevó merienda al colegio para compartir y se la rechazaron por temor a contagiarse de coronavirus	Con lo hijos de puta, maleducados e intransigentes que son los chinos, si fuese al revés ya nos hubiesen embarcado a la estratosfera. Nadie ha visto como echan a los chicos pobres de los super por el miedo a que los afanen? Hablando de prejuicios
1	FN	Multitudes en las calles y discotecas abarrotadas: la fiesta de Wuhan tras el año de la pandemia que se inició en uno de sus "mercados húmedos"	Los odio
1	FN	Su novia es mexicana y Migraciones le exige casi 50 mil pesos para dejarla ingresar a la Argentina	Mira si le cobran eso a cada venezolano colombiano peruano chino boliviano que vienen al país a chupar sangre ,pagamos la deuda externa y nos volvemos potencia mundial
1	FP	URGENTE: Un hombre se incrustó con su auto en la puerta de la Embajada de China y aseguró que tenía explosivos	No es hombre . Es un boludo
1	FP	Su novia es mexicana y Migraciones le exige casi 50 mil pesos para dejarla ingresar a la Argentina	Que me traiga una botellas de tequila del bueno
1	FP	El principal gremio docente nacional rechazó el regreso a las clases presenciales	Banda de VAGOS
1	FP	China: identificaron otro virus çon potencial para convertirse en pandemia"nuevo virus china transportado por cerdos podría infectar ahumanos las actuales vacunas podrían adaptarse más en la nota	PERO LA PUTA MADRE VIEJO

Tab. C.1: Ejemplos donde el clasificador granular acierta y el binario falla. FN marca que el clasificador binario no detecta el comentario como discriminatorio mientras que el contextualizado sí lo hace; FP es al revés, que el clasificador binario marca erróneamente el comentario como discriminatorio. Agrupadas de acuerdo a ciertas características; en el caso de los falsos positivos esta categoría es especulativa

La Tabla C.1 contiene ejemplos de discurso de odio donde el clasificador entrenado sobre la tarea granular acierta, y el clasificador entrenado sobre la tarea binaria falla. La categoría FN indica que el clasificador binario no detecta el comentario como discriminatorio, mientras que la categoría FP indica que el clasificador binario marca erróneamente el comentario como discriminatorio. No pudimos encontrar ningún patrón detrás de estos ejemplos.

D. ADAPTACIÓN DE DOMINIO

D.1. Algunos detalles técnicos

Una pequeña observación técnica es el que el código de entrenamiento es una adaptación de los ejemplos de la librería *huggingface/transformers*, ya que la inmensa cantidad de datos que manejamos(cerca de 500gb) no es manejada adecuadamente por esta librería. Dejamos a quien esté interesado esta aclaración para adaptarla cuando este problema sea resuelto.

D.2. Tabla completa de resultados

En la Tabla D.1 tenemos los resultados para todos los modelos considerados en el benchmark de adaptación de dominio, referidos en la sección 7.3.2. Notamos, para compacidad, con subíndice *U, C, D* a las versiones *uncased*, *cased* y *deacc*. Así mismo, notamos con $10K$ (por ejemplo) a aquel modelo con adaptación de dominio por 10,000 pasos según descripto en la sección 7.3.

Podemos observar que, observando el score general, el mejor modelo adaptado a dominio para las versiones *uncased* es beto_{U10K}, y para las versiones *cased* es $BETO_{C5K}$; si bien en este último caso tiene una performance muy similar al de 20K pasos (de hecho, omitiendo la tarea de discurso de odio contextualizado gana por mínimo margen el de 20K).

D.3. Evaluación multilingual

Evaluamos *RoBERTuito* en español, como ya hemos visto en el Capítulo 7. Adicionalmente, debido al proceso de recolección de datos, nuestro conjunto de pre-entrenamiento contiene tweets en otros idiomas, potencialmente en una mezcla de ellos. Teniendo eso en cuenta, evaluamos el modelo en otras dos configuraciones: inglés y code-switching inglés-español. La Tabla D.2 resume todas las tareas en las que evaluamos *RoBERTuito*.

Para **inglés**, probamos *RoBERTuito* en tres tareas: análisis de emociones, detección de discursos de odio y análisis de sentimientos. Para el análisis de emociones y el discurso de odio usamos las secciones en inglés de los conjuntos de datos antes mencionados (*EmoEvent* y *HatEval*), mientras que para el análisis de sentimientos usamos el dataset *SemEval 2017 Task-4* [141], que comparte las mismas etiquetas que el conjunto de datos correspondiente español (negativo, neutro, positivo). En este caso, comparamos las habilidades de *RoBERTuito* en inglés con modelos monolingües, *BERT*, *RoBERTa* y *BERTweet*; y también contra modelos multilingües como *XLM-R* [39] y *mBERT*. Si bien todos estos modelos comparten una arquitectura base, los diferentes tamaños de vocabulario y la cantidad de parámetros hacen que la comparación no sea tan directa.

Finalmente, evaluamos las habilidades de code-switching de nuestro modelo en el *Linguistic Code-Switching Evaluation Benchmark* (LinCE) [2]. LinCE comprende cinco tareas para datos de código conmutado en varios pares de idiomas (español-inglés, hindi-inglés, árabe estándar moderno-árabe egipcio, árabe-inglés, entre otros), muchas de las cuales formaron parte de tareas compartidas anteriores. Evaluamos *RoBERTuito* en tres tareas diferentes del benchmark: POS tagging [3], reconocimiento de entidad nombrada (NER) y análisis de sentimientos [119]. Como el proceso de recopilación de datos se centró en los usuarios de habla hispana, algunos de los cuales también hablan inglés y spanglish [1], probamos *RoBERTuito* en la subsección español-inglés del benchmark.

Este benchmark tiene un sistema de evaluación centralizado, no liberando etiquetas doradas para el subconjunto de pruebas. Evaluamos nuestros modelos en los conjuntos de datos de desarrollo y comparamos nuestros resultados con los proporcionados por Winata et al. [167], que logra el mejor rendimiento para el etiquetado NER y POS. Como modelos competidores de *RoBERTuito* para la

[1] La mezcla morfosintáctica de español e inglés

Modelo	CHATE	HATE	SENTIMENT	EMOTION	IRONY	score
robertuito$_U$	$59{,}3 \pm 0{,}4$	$80{,}1 \pm 1{,}0$	$70{,}7 \pm 0{,}4$	$55{,}1 \pm 1{,}1$	$73{,}6 \pm 0{,}8$	67,8
robertuito$_D$	$59{,}3 \pm 0{,}6$	$79{,}8 \pm 0{,}8$	$70{,}2 \pm 0{,}4$	$54{,}3 \pm 1{,}5$	$74{,}0 \pm 0{,}6$	67,5
robertuito$_C$	$59{,}0 \pm 0{,}5$	$79{,}0 \pm 1{,}2$	$70{,}1 \pm 1{,}2$	$51{,}9 \pm 3{,}2$	$71{,}9 \pm 2{,}3$	66,4
beto$_{U10K}$	$58{,}8 \pm 0{,}3$	$77{,}5 \pm 1{,}5$	$68{,}0 \pm 0{,}4$	$55{,}3 \pm 0{,}9$	$71{,}7 \pm 0{,}5$	66,3
beto$_{U20K}$	$58{,}8 \pm 0{,}7$	$76{,}8 \pm 1{,}2$	$68{,}4 \pm 0{,}5$	$53{,}3 \pm 1{,}6$	$71{,}2 \pm 0{,}9$	65,7
beto$_{C5K}$	$57{,}6 \pm 0{,}2$	$78{,}1 \pm 1{,}0$	$67{,}7 \pm 0{,}4$	$52{,}5 \pm 1{,}6$	$72{,}4 \pm 0{,}9$	65,7
beto$_{C20K}$	$57{,}2 \pm 0{,}6$	$77{,}7 \pm 0{,}9$	$68{,}6 \pm 0{,}5$	$51{,}7 \pm 0{,}9$	$73{,}0 \pm 0{,}4$	65,6
beto$_{C10K}$	$57{,}4 \pm 0{,}8$	$78{,}2 \pm 0{,}9$	$68{,}0 \pm 0{,}6$	$52{,}4 \pm 0{,}6$	$72{,}0 \pm 0{,}7$	65,6
beto$_{C2{,}5K}$	$58{,}0 \pm 0{,}5$	$77{,}1 \pm 0{,}7$	$67{,}7 \pm 0{,}6$	$52{,}5 \pm 1{,}0$	$71{,}7 \pm 0{,}8$	65,4
RoBERTa$_{ES}$	$57{,}7 \pm 0{,}4$	$76{,}6 \pm 1{,}5$	$66{,}9 \pm 0{,}6$	$53{,}3 \pm 1{,}1$	$72{,}3 \pm 1{,}7$	65,3
beto$_C$	$58{,}2 \pm 0{,}7$	$76{,}8 \pm 1{,}2$	$66{,}5 \pm 0{,}4$	$52{,}1 \pm 1{,}2$	$70{,}6 \pm 0{,}7$	64,8
bertin	$55{,}7 \pm 0{,}8$	$76{,}7 \pm 0{,}5$	$66{,}5 \pm 0{,}3$	$51{,}8 \pm 1{,}2$	$71{,}6 \pm 0{,}8$	64,5
beto$_U$	$59{,}1 \pm 0{,}6$	$75{,}7 \pm 1{,}2$	$64{,}9 \pm 0{,}5$	$52{,}1 \pm 0{,}6$	$70{,}2 \pm 0{,}8$	64,4
beto$_{U5K}$	$55{,}7 \pm 0{,}7$	$75{,}6 \pm 1{,}2$	$65{,}4 \pm 0{,}5$	$50{,}9 \pm 1{,}4$	$68{,}4 \pm 0{,}7$	63,2
beto$_{U2{,}5K}$	$58{,}8 \pm 0{,}4$	$78{,}4 \pm 1{,}1$	$67{,}6 \pm 0{,}5$	$53{,}3 \pm 0{,}8$	$71{,}5 \pm 0{,}7$	65,9

Tab. D.1: Resultados de la evaluación de modelos pre-entrenados y modelos ajustados en dominio para el benchmark de tareas sociales: CHATE es contextualized hate speech, HATE es hate speech detection sobre el dataset de hatEval, SENTIMENT, EMOTION e IRONY son análisis de sentimiento, emociones e ironía sobre los corpus de TASS. Todos los scores son Macro F1s. beto-cased-ft y beto-uncased-ft son modelos adaptados al dominio social. Score es la media de cada fila.

Idioma	Tareas	Tipo de tareas	Dataset	Tamaño
Español	Análisis de Sentimientos	Text Classification	TASS 2020 Task A	14,500
	Análisis de emociones		TASS 2020 Task B	8,400
	Discurso de odio		HatEval	6,600
	Irony Detection		IrosVA 2019	9,000
Inglés	Análisis de Sentimientos	Text Classification	SemEval 2017 Task 4	61,900
	Análisis de emociones		TASS 2020 Task B	7,303
	Discurso de odio		HatEval	13,000
Español-Inglés	Análisis de Sentimientos	Text Classification	LinCE	18,789
	POS tagging	Text Labelling		42,911
	NER	Text Labelling		67,233

Tab. D.2: Tareas de evaluación para *RoBERTuito*. Las tareas se agrupan por configuración: tareas solo en español, tareas solo en inglés y tareas de código mixto español-inglés.

evaluación español-inglés, tenemos mBERT, XLM-R (tanto en arquitectura base como grande) y los modelos monolingües *BERT* y *BETO*.

D.3.1. Resultados

La Tabla D.3 muestra los resultados de la evaluación de los modelos seleccionados para las tres tareas en inglés. Podemos observar que *RoBERTuito* supera tanto a mBERT como a XLM-R, que son los otros modelos multilinguales evaluados para las tareas. En comparación con los modelos monolinguales en inglés, los resultados de *RoBERTuito* son similares a los de *RoBERTa* y ligeramente superiores a *BERT*. Como era de esperar, *BERTweet* obtiene los mejores resultados.

Modelo	Odio	Sentim	Emoción
BERTweet	**55,3**	**70,3**	42,8
RoBERTuito	54,2	68,4	44,1
RoBERTa	45,8	69,5	**46,3**
BERT	48,9	68,9	42,8
mBERT*	43,3	66,6	40,4
XLM-R_{BASE}*	45,7	68,0	35,7

Tab. D.3: Resultados de la evaluación de las tres tareas de clasificación en inglés. Los resultados se expresan como la puntuación media de Macro F1 de 10 ejecuciones de los experimentos de clasificación. * marca modelos multilinguales

Model	Sentiment	NER	POS
RoBERTuito	**60,6**	68,5	97,2
XLM-R_{LARGE}	–	**69,5**	**97,2**
XLM-R_{BASE}	–	64,9	97,0
C2S mBERT	59,1	64,6	96,9
mBERT	56,4	64,0	97,1
BERT	58,4	61,1	96,9
BETO	56,5	–	–

Tab. D.4: Resultados de la evaluación para las tareas de código mixto de la sección español-inglés del benchmark LinCE. Los resultados se toman de la clasificación oficial del benchmark. El rendimiento de Análisis de Sentimientos se mide con Macro F1, NER con Micro F1 y la POS tagging con accuracy. C2S es un acrónimo de Char2Subword BERT

La Tabla D.4 muestra los resultados de la clasificación del LinCE benchmark [2] para las tres tareas seleccionadas: análisis de sentimientos, NER y POS tagging. Para la primer tarea obtiene los mejores resultados en términos de Micro F1. Para las otras dos tareas, obtiene la segunda posición, donde un modelo XLM-R_{LARGE} [167] tiene los mejores resultados. Entre los modelos comparados, *RoBERTuito* tiene 108 millones de parámetros, mientras que XLM-R_{LARGE} suma alrededor de cinco veces este número, lo que hace que nuestro modelo sea el más eficiente en términos de tamaño para esta subsección del benchmark.

[2] `https://ritual.uh.edu/lince/leaderboard`

Bibliografía

[1] Josh Adams and Vincent J Roscigno. White supremacists, oppositional culture and the world wide web. *Social Forces*, 84(2):759–778, 2005.

[2] Gustavo Aguilar, Sudipta Kar, and Thamar Solorio. LinCE: A Centralized Benchmark for Linguistic Code-switching Evaluation. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1803–1813, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://www.aclweb.org/anthology/2020.lrec-1.223`.

[3] Fahad AlGhamdi, Giovanni Molina, Mona Diab, Thamar Solorio, Abdelati Hawwari, Victor Soto, and Julia Hirschberg. Part of speech tagging for code switched data. In *Proceedings of the Second Workshop on Computational Approaches to Code Switching*, pages 98–107, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5812. URL `https://www.aclweb.org/anthology/W16-5812`.

[4] Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. Automatic identification and classification of misogynistic language on twitter. In *International Conference on Applications of Natural Language to Information Systems*, pages 57–64. Springer, 2018.

[5] Aymé Arango, Jorge Pérez, and Barbara Poblete. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd international acm sigir conference on research and development in information retrieval*, pages 45–54, 2019.

[6] Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *International Conference on Learning Representations*, 2017.

[7] Article 19. Hate speech explained: A toolkit. Technical report, Article 19, London, UK, London, UK, 2015.

[8] Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 759–760. International World Wide Web Conferences Steering Committee, 2017.

[9] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

[10] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Rangel, Paolo Rosso, and Manuela Sanguinetti. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019)*. Association for Computational Linguistics, 2019.

[11] Peter W. Battaglia, Jessica B. Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, Caglar Gulcehre, Francis Song, Andrew Ballard, Justin Gilmer, George Dahl, Ashish Vaswani, Kelsey Allen, Charles Nash, Victoria Langston, Chris Dyer, Nicolas Heess, Daan Wierstra, Pushmeet Kohli, Matt Botvinick, Oriol Vinyals, Yujia Li, and Razvan Pascanu. Relational inductive biases, deep learning, and graph networks, 2018.

[12] Iz Beltagy, Kyle Lo, and Arman Cohan. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1371. URL `https://aclanthology.org/D19-1371`.

[13] Emily Bender. The benderrule: On naming the languages we study and why it matters. *The Gradient*, 2019.

[14] Emily M Bender. On achieving and evaluating language-independence in nlp. *Linguistic Issues in Language Technology*, 6(3):1–26, 2011.

[15] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.463. URL `https://aclanthology.org/2020.acl-main.463`.

[16] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? . In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021.

[17] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The journal of machine learning research*, 3:1137–1155, 2003.

[18] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[19] Michał Bilewicz and Wiktor Soral. Hate speech epidemic. the dynamic effects of derogatory language on intergroup relations and political radicalization. *Political Psychology*, 41:3–33, 2020.

[20] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006. ISBN 0387310738.

[21] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[22] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008, 2008.

[23] Emily Blout and Patrick Burkart. White supremacist terrorism in charlottesville: Reconstructing 'unite the right'. *Studies in Conflict & Terrorism*, pages 1–22, 2020.

[24] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[25] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450366755. doi: 10.1145/3308560.3317593. URL `https://doi.org/10.1145/3308560.3317593`.

[26] Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR, 2018.

[27] Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. Class-based n-gram models of natural language. *Computational linguistics*, 18(4): 467–480, 1992.

[28] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[29] Pete Burnap and Matthew L Williams. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2):223–242, 2015.

[30] Carlos Arcila Calderón, Gonzalo de la Vega, and David Blanco Herrero. Topic modeling and characterization of hate speech against immigrants on twitter around the emergence of a far-right party in spain. *Social Sciences*, 9(11):188, 2020.

[31] José Canete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. Spanish pre-trained bert model and evaluation data. *PML4DC at ICLR*, 2020.

[32] Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.findings-emnlp.261. URL `https://aclanthology.org/2020.findings-emnlp.261`.

[33] Wanxiang Che, Yijia Liu, Yuxuan Wang, Bo Zheng, and Ting Liu. Towards better UD parsing: Deep contextualized word embeddings, ensemble, and treebank concatenation. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 55–64, Brussels, Belgium, October 2018. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/K18-2005`.

[34] Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics. doi: 10.3115/v1/D14-1179. URL `https://aclanthology.org/D14-1179`.

[35] Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

[36] Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroğlu, and Marco Guerini. Conan-counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, 2019.

[37] CIDH. Discurso de odio y la incitación a la violencia contra las personas lesbianas, gays, bisexuales, trans e intersex en américa. Technical report, Comisión Interamericana sobre Derechos Humanos, 2015.

[38] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *Journal of machine learning research*, 12(ARTICLE):2493–2537, 2011.

[39] 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. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747.

[40] Miguel Ángel García Cumbreras, Julio Villena-Román, Eugenio Martínez Cámara, Manuel Carlos Díaz-Galiano, Maria Teresa Martín-Valdivia, and Luis Alfonso Urena López. Overview of tass 2016. In *TASS@ SEPLN*, pages 13–21, 2016.

[41] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[42] Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. In *ICWSM*, 2017.

[43] Flor Miriam Plaza del Arco, M. Dolores Molina-González, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120, 2021. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2020.114120. URL https://www.sciencedirect.com/science/article/pii/S095741742030868X.

[44] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848.

[45] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[46] Jacob Eisenstein. What to do about bad language on the internet. In *Proceedings of the 2013 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 359–369, 2013.

[47] Paul Ekman. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200, 1992.

[48] Jeffrey L Elman. Finding structure in time. *Cognitive science*, 14(2):179–211, 1990.

[49] Keyur Faldu, Amit Sheth, Prashant Kikani, and Hemang Akabari. Ki-bert: Infusing knowledge context for better language and domain understanding. *arXiv preprint arXiv:2104.08145*, 2021.

[50] Elisabetta Fersini, Maria Anzovino, and Paolo Rosso. Overview of the task on automatic misogyny identification at ibereval. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018), co-located with 34th Conference of the Spanish Society for Natural Language Processing (SEPLN 2018). CEUR Workshop Proceedings. CEUR-WS. org, Seville, Spain*, 2018.

[51] Elisabetta Fersini, Debora Nozza, and Paolo Rosso. Overview of the evalita 2018 task on automatic misogyny identification (ami). *Proceedings of the 6th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA'18), Turin, Italy. CEUR. org*, 2018.

[52] Jill Filipovic. Blogging while female: How internet misogyny parallels real-world harassment. *Yale JL & Feminism*, 19:295, 2007.

[53] Paula Fortuna and Sérgio Nunes. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30, 2018.

[54] Damián Furman, Santiago Marro, Cristian Cardellino, Diana Popa, and Laura Alonso Alemany. You can simply rely on communities for a robust characterization of stances. *Florida Artificial Intelligence Research Society*, 34(1), 2021.

[55] Iginio Gagliardone, Danit Gal, Thiago Alves, and Gabriela Martinez. *Countering online hate speech*. Unesco Publishing, 2015.

[56] Lei Gao and Ruihong Huang. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria, September 2017. INCOMA Ltd. doi: 10.26615/978-954-452-049-6_036. URL https://doi.org/10.26615/978-954-452-049-6_036.

[57] Manuel García-Vegaa, Manuel Carlos Díaz-Galianoa, Miguel Á García-Cumbrerasa, Flor Miriam Plaza del Arcoa, Arturo Montejo-Ráeza, Salud María Jiménez-Zafraa, Eugenio Martínez Cámarab, César Antonio Aguilarc, Marco Antonio, Sobrevilla Cabezudod, et al. Overview of tass 2020: introducing emotion detection. 2020.

[58] Manas Gaur, Ugur Kursuncu, Amit Sheth, Ruwan Wickramarachchi, and Shweta Yadav. Knowledge-infused deep learning. In *Proceedings of the 31st ACM Conference on Hypertext and Social Media*, HT '20, page 309–310, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450370981. doi: 10.1145/3372923.3404862. URL https://doi.org/10.1145/3372923.3404862.

[59] Abigail Gertner, John Henderson, Elizabeth Merkhofer, Amy Marsh, Ben Wellner, and Guido Zarrella. MITRE at SemEval-2019 task 5: Transfer learning for multilingual hate speech detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 453–459, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2080. URL https://aclanthology.org/S19-2080.

[60] Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon Univ Pittsburgh Pa School of Computer Science, 2010.

[61] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[62] Jose Angel Gonzalez, Lluís-F Hurtado, and Ferran Pla. Twilbert: Pre-trained deep bidirectional transformers for spanish twitter. *Neurocomputing*, 426:58–69, 2021.

[63] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[64] Eduardo Graells-Garrido, Ricardo Baeza-Yates, and Mounia Lalmas. *How Representative is an Abortion Debate on Twitter?*, page 133–134. Association for Computing Machinery, New York, NY, USA, 2019. ISBN 9781450362023. URL https://doi.org/10.1145/3292522.3326057.

[65] Edel Greevy and Alan F Smeaton. Classifying racist texts using a support vector machine. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 468–469. ACM, 2004.

[66] John J Gumperz. Contextualization revisited. *The contextualization of language*, 22:39–53, 1992.

[67] Raj Kumar Gupta and Yinping Yang. CrystalNest at SemEval-2017 task 4: Using sarcasm detection for enhancing sentiment classification and quantification. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 626–633, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2103. URL https://aclanthology.org/S17-2103.

[68] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.740. URL https://aclanthology.org/2020.acl-main.740.

[69] Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquín Silveira-Ocampo, Casimiro Pio Carrino, Aitor Gonzalez-Agirre, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. Spanish language models, 2021.

[70] Bo Han and Timothy Baldwin. Lexical normalisation of short text messages: Makn sens a# twitter. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 368–378, 2011.

[71] Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. Racism is a virus: Anti-asian hate and counterspeech in social media during the covid-19 crisis, 2021.

[72] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[73] Dan Hendrycks, Xiaoyuan Liu, Eric Wallace, Adam Dziedzic, Rishabh Krishnan, and Dawn Song. Pretrained transformers improve out-of-distribution robustness. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2744–2751, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.244. URL https://aclanthology.org/2020.acl-main.244.

[74] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9 (8):1735–1780, 1997.

[75] J Horrigan. Online shopping, pew internet & american life project report. *Washington, DC: Pew Research Center*, pages 1–42, 2008.

[76] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138*, 2017.

[77] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1031. URL https://aclanthology.org/P18-1031.

[78] Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. WikiConv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1305. URL https://aclanthology.org/D18-1305.

[79] Peter Izsak, Moshe Berchansky, and Omer Levy. How to train bert with an academic budget. *arXiv preprint arXiv:2104.07705*, 2021.

[80] Edwin Jain, Stephan Brown, Jeffery Chen, Erin Neaton, Mohammad Baidas, Ziqian Dong, Huanying Gu, and Nabi Sertac Artan. Adversarial text generation for google's perspective api. In *2018 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1136–1141, 2018. doi: 10.1109/CSCI46756.2018.00220.

[81] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 427–431, Valencia, Spain, April 2017. Association for Computational Linguistics. URL `https://aclanthology.org/E17-2068`.

[82] Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. The gab hate corpus: A collection of 27k posts annotated for hate speech. 2018.

[83] Heather Hensman Kettrey and Whitney Nicole Laster. Staking territory in the "world white web" an exploration of the roles of overt and color-blind racism in maintaining racial boundaries on a popular web site. *Social Currents*, 1(3):257–274, 2014.

[84] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[85] Adam Klein. From twitter to charlottesville: Analyzing the fighting words between the alt-right and antifa. *International Journal of Communication*, 13:22, 2019.

[86] Klaus Krippendorff. *Content analysis: An introduction to its methodology*. Sage publications, 2018.

[87] William H Kruskal and W Allen Wallis. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association*, 47(260):583–621, 1952.

[88] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-2012. URL `https://aclanthology.org/D18-2012`.

[89] Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. Race: Large-scale reading comprehension dataset from examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 785–794, 2017.

[90] Mirko Lai, Viviana Patti, Giancarlo Ruffo, and Paolo Rosso. Stance evolution and twitter interactions in an italian political debate. In *International Conference on Applications of Natural Language to Information Systems*, pages 15–27. Springer, 2018.

[91] Thomas K Landauer and Susan T Dumais. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, 104(2):211, 1997.

[92] Pierre Lévy. *Cyberculture*, volume 4. U of Minnesota Press, 2001.

[93] Rensis Likert. A technique for the measurement of attitudes. *Archives of psychology*, 1932.

[94] Weijie Liu, Peng Zhou, Zhe Zhao, Zhiruo Wang, Qi Ju, Haotang Deng, and Ping Wang. K-bert: Enabling language representation with knowledge graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:2901–2908, 04 2020. doi: 10.1609/aaai.v34i03.5681.

[95] 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. *arXiv preprint arXiv:1907.11692*, 2019.

[96] Robert Logan, Nelson F. Liu, Matthew E. Peters, Matt Gardner, and Sameer Singh. Barack's wife hillary: Using knowledge graphs for fact-aware language modeling. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5962–5971, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1598. URL `https://aclanthology.org/P19-1598`.

[97] Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, 2015.

[98] Franco M. Luque and Juan Manuel Pérez. Atalaya at TASS 2018: Sentiment analysis with tweet embeddings and data augmentation. In *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN, TASS@SEPLN 2018, co-located with 34nd SEPLN Conference (SEPLN 2018), Sevilla, Spain, September 18th, 2018.*, pages 29–35, 2018. URL `http://ceur-ws.org/Vol-2172/p1_atalaya_tass2018.pdf`.

[99] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL `http://www.aclweb.org/anthology/P11-1015`.

[100] Karla Mantilla. Gendertrolling: Misogyny adapts to new media. *Feminist Studies*, 39(2): 563–570, 2013.

[101] Eugenio Martínez-Cámara, Yudivián Almeida Cruz, Manuel C. Díaz-Galiano, Suilan Estévez Velarde, Miguel Á. García-Cumbreras, Manuel García-Vega, Yoan Gutiérrez Vázquez, Arturo Montejo Ráez, André Montoyo Guijarro, Rafael Muñoz Guillena, Alejandro Piad Morffis, and Julio Villena-Román. Overview of TASS 2018: Opinions, health and emotions. In Eugenio Martínez-Cámara, Yudivián Almeida Cruz, Manuel C. Díaz-Galiano, Suilan Estévez Velarde, Miguel Á. García-Cumbreras, Manuel García-Vega, Yoan Gutiérrez Vázquez, Arturo Montejo Ráez, André Montoyo Guijarro, Rafael Muñoz Guillena, Alejandro Piad Morffis, and Julio Villena-Román, editors, *Proceedings of TASS 2018: Workshop on Semantic Analysis at SEPLN (TASS 2018)*, volume 2172 of *CEUR Workshop Proceedings*, Sevilla, Spain, September 2018. CEUR-WS.

[102] Ramón Martín-Brufau, Carlos Suso-Ribera, and Javier Corbalán. Emotion network analysis during covid-19 quarantine - a longitudinal study. *Frontiers in Psychology*, 11:2802, 2020. ISSN 1664-1078. doi: 10.3389/fpsyg.2020.559572. URL `https://www.frontiersin.org/article/10.3389/fpsyg.2020.559572`.

[103] Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305, 2017.

[104] Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, 1943.

[105] Reid McIlroy-Young and Ashton Anderson. From "welcome new gabbers" to the pittsburgh synagogue shooting: The evolution of gab. In *Proceedings of the international aaai conference on web and social media*, volume 13, pages 651–654, 2019.

[106] Stephen Merity, Nitish Shirish Keskar, and Richard Socher. Regularizing and optimizing lstm language models. In *International Conference on Learning Representations*, 2018.

[107] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[108] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[109] Marvin Minsky and Seymour Papert. Perceptrons. 1969.

[110] Melanie Mitchell. Why ai is harder than we think. *arXiv preprint arXiv:2104.12871*, 2021.

[111] Saif M Mohammad and Peter D Turney. Nrc emotion lexicon. *National Research Council, Canada*, 2, 2013.

[112] Hamdy Mubarak, Kareem Darwish, and Walid Magdy. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56, Vancouver, BC, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-3008. URL https://aclanthology.org/W17-3008.

[113] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen. BERTweet: A pre-trained language model for English Tweets. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 9–14, 2020.

[114] Reynier Ortega-Bueno, Francisco Rangel, D Hernández Farıas, Paolo Rosso, Manuel Montes-y Gómez, and José E Medina Pagola. Overview of the task on irony detection in spanish variants. In *Proceedings of the Iberian languages evaluation forum (IberLEF 2019), co-located with 34th conference of the Spanish Society for natural language processing (SEPLN 2019). CEUR-WS. org*, volume 2421, pages 229–256, 2019.

[115] Tin Htet Paing. Zuckerberg urged to take genuine steps to stop use of fb to spread hate in myanmar. *The Irrawaddy*. URL https://www.irrawaddy.com/news/burma/zuckerberg-urged-to-take-genuine-steps-to-stop-use-of-fb-to-spread-hate-in-myanmar.html.

[116] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREC*, volume 10, pages 1320–1326, 2010.

[117] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1–2):1–135, jan 2008. ISSN 1554-0669. doi: 10.1561/1500000011. URL https://doi.org/10.1561/1500000011.

[118] Ankur Parikh, Oscar Täckström, Dipanjan Das, and Jakob Uszkoreit. A decomposable attention model for natural language inference. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2249–2255, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1244. URL https://aclanthology.org/D16-1244.

[119] Parth Patwa, Gustavo Aguilar, Sudipta Kar, Suraj Pandey, Srinivas PYKL, Bj.ºrn Gamb.ªck, Tanmoy Chakraborty, Thamar Solorio, and Amitava Das. Semeval-2020 task 9: Overview of sentiment analysis of code-mixed tweets. In *Proceedings of the 14th International Workshop on Semantic Evaluation (SemEval-2020)*, Barcelona, Spain, December 2020. Association for Computational Linguistics.

[120] Ioannis Pavlopoulos. Aspect based sentiment analysis. *Athens University of Economics and Business*, 2014.

[121] John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. Toxicity detection: Does context really matter? *arXiv preprint arXiv:2006.00998*, 2020.

[122] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[123] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001, 2001.

[124] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[125] Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. Detecting and monitoring hate speech in twitter. *Sensors*, 19(21):4654, 2019.

[126] Juan Manuel Pérez and Franco M. Luque. Atalaya at SemEval 2019 task 5: Robust embeddings for tweet classification. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 64–69, Minneapolis, Minnesota, USA, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/S19-2008. URL `https://aclanthology.org/S19-2008`.

[127] Juan Manuel Pérez, Damián A Furman, Laura Alonso Alemany, and Franco Luque. Robertuito: a pre-trained language model for social media text in spanish. *arXiv preprint arXiv:2111.09453*, 2021.

[128] Matthew Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. Semi-supervised sequence tagging with bidirectional language models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1756–1765, 2017.

[129] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. *CoRR*, abs/1802.05365, 2018. URL `http://arxiv.org/abs/1802.05365`.

[130] Jürgen Pfeffer, Katja Mayer, and Fred Morstatter. Tampering with twitter's sample api. *EPJ Data Science*, 7(1):50, 2018.

[131] Flor Miriam Plaza del Arco, Carlo Strapparava, L. Alfonso Urena Lopez, and Maite Martin. EmoEvent: A multilingual emotion corpus based on different events. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1492–1498, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL `https://aclanthology.org/2020.lrec-1.186`.

[132] Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523, 2021.

[133] Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, and Valerio Basile. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *6th Italian Conference on Computational Linguistics, CLiC-it 2019*, volume 2481, pages 1–6. CEUR, 2019.

[134] James Pustejovsky and Amber Stubbs. *Natural Language Annotation for Machine Learning: A guide to corpus-building for applications*. .ºReilly Media, Inc.", 2012.

[135] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training.

[136] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL `http://jmlr.org/papers/v21/20-074.html`.

[137] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-bert: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, 4(1):1–13, 2021.

[138] Howard Rheingold. *The virtual community: Finding commection in a computerized world*. Addison-Wesley Longman Publishing Co., Inc., 1993.

[139] Alan Ritter, Sam Clark, Oren Etzioni, et al. Named entity recognition in tweets: an experimental study. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 1524–1534, 2011.

[140] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386, 1958.

[141] Sara Rosenthal, Noura Farra, and Preslav Nakov. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/S17-2088. URL `https://aclanthology.org/S17-2088`.

[142] Sebastian Ruder. Recent Advances in Language Model Fine-tuning. `http://ruder.io/recent-advances-lm-fine-tuning`, 2021.

[143] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[144] Koustuv Saha, Eshwar Chandrasekharan, and Munmun De Choudhury. Prevalence and psychological effects of hateful speech in online college communities. In *Proceedings of the 10th ACM conference on web science*, pages 255–264, 2019.

[145] Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*, 2017.

[146] Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. An italian twitter corpus of hate speech against immigrants. In *Proceedings of the eleventh international conference on language resources and evaluation (LREC 2018)*, 2018.

[147] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, 2020.

[148] Helmut Schmid. Improvements in part-of-speech tagging with an application to german. In *In Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, 1995.

[149] Anna Schmidt and Michael Wiegand. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10, 2017.

[150] Mike Schuster and Kaisuke Nakajima. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE, 2012.

[151] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

[152] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. SemEval-2020 task 8: Memotion analysis- the visuo-lingual metaphor! In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.99. URL `https://aclanthology.org/2020.semeval-1.99`.

[153] Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. Defining and detecting toxicity on social media: Context and knowledge are key. *arXiv preprint arXiv:2104.10788*, 2021.

[154] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL `https://aclanthology.org/D13-1170`.

[155] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[156] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

[157] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112, 2014.

[158] Wilson L Taylor. "cloze procedure": A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433, 1953.

[159] Mike Thelwall. Social networks, gender, and friending: An analysis of myspace member profiles. *Journal of the American Society for Information Science and Technology*, 59(8): 1321–1330, 2008.

[160] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[161] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL `https://aclanthology.org/W18-5446`.

[162] William Warner and Julia Hirschberg. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26. Association for Computational Linguistics, 2012.

[163] Alex Warofka. An independent assessment of the human rights impact of facebook in myanmar. *Facebook Newsroom, November*, 5, 2018.

[164] Zeerak Waseem. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5618. URL `https://aclanthology.org/W16-5618`.

[165] Zeerak Waseem and Dirk Hovy. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93, 2016.

[166] Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. Implicitly abusive language–what does it actually look like and why are we not getting there? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, 2021.

[167] Genta Indra Winata, Samuel Cahyawijaya, Zihan Liu, Zhaojiang Lin, Andrea Madotto, and Pascale Fung. Are multilingual models effective in code-switching? In *Proceedings of the Fifth Workshop on Computational Approaches to Linguistic Code-Switching*, pages 142–153, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.calcs-1.20. URL `https://aclanthology.org/2021.calcs-1.20`.

[168] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL `https://aclanthology.org/2020.emnlp-demos.6`.

[169] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. 2016.

[170] Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.woah-1.15. URL `https://aclanthology.org/2021.woah-1.15`.

[171] Mariona Taulé y Alejandro Ariza y Montserrat Nofre y Enrique Amigó y Paolo Rosso. Overview of detoxis at iberlef 2021: Detection of toxicity in comments in spanish. *Procesamiento del Lenguaje Natural*, 67(0):209–221, 2021. ISSN 1989-7553. URL `http://journal.sepln.org/sepln/ojs/ojs/index.php/pln/article/view/6390`.

[172] Yang You, Jing Li, Sashank Reddi, Jonathan Hseu, Sanjiv Kumar, Srinadh Bhojanapalli, Xiaodan Song, James Demmel, Kurt Keutzer, and Cho-Jui Hsieh. Large batch optimization for deep learning: Training bert in 76 minutes. *arXiv preprint arXiv:1904.00962*, 2019.

[173] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. A survey of sentiment analysis in social media. *Knowledge and Information Systems*, 60(2):617–663, 2019.

[174] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Predicting the type and target of offensive posts in social media, 2019.

[175] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval), 2019.

[176] Min-Ling Zhang and Zhi-Hua Zhou. A review on multi-label learning algorithms. *Knowledge and Data Engineering, IEEE Transactions on*, 26:1819–1837, 08 2014. doi: 10.1109/TKDE.2013.39.

[177] Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. Ernie: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1441–1451, 2019.

[178] Ziqi Zhang, David Robinson, and Jonathan Tepper. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European Semantic Web Conference*, pages 745–760. Springer, 2018.

[179] Yukun Zhu, Ryan Kiros, Rich 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. In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[180] Frederik Zuiderveen Borgesius, Judith Möller, Sanne Kruikemeier, Ronan Ó Fathaigh, Kristina Irion, Tom Dobber, Balazs Bodo, and Claes H de Vreese. Online political microtargeting: promises and threats for democracy. *Utrecht Law Review*, 14(1):82–96, 2018.