PONTIFICIA UNIVERSIDAD CATÓLICA DEL PERÚ

Escuela de Posgrado

Detección de ciberbullying en español para el dominio de corpus de texto teatrales aplicado a redes sociales usando transferencia de aprendizaje y validación adversarial

Trabajo de investigación para obtener el grado académico de Maestro en Informática que presenta:

Esli Samuel Marquez Zavaleta

Asesor:

Héctor Erasmo Gómez Montoya

Lima, 2024

Informe de Similitud

Yo, **Héctor Erasmo GÓMEZ MONTOYA**, docente de la Escuela de Posgrado de la Pontificia Universidad Católica del Perú, asesor de el trabajo de investigación titulado "Detección de ciberbullying en español para el dominio de corpus de texto teatrales aplicado a redes sociales usando transferencia de aprendizaje y validación adversarial" de el autor **Esli Samuel Marquez Zavaleta**, dejo constancia de lo siguiente:

- El mencionado documento tiene un índice de puntuación de similitud de 08%. Así lo consigna el reporte de similitud emitido por el software *Turnitin* el 24/07/2024.
- He revisado con detalle dicho reporte y el trabajo de investigación, y no se advierte indicios de plagio.
- Las citas a otros autores y sus respectivas referencias cumplen con las pautas académicas.

Lugar y fecha:

San Miguel, 24 de Julio de 2024.

Apellidos y nombres del asesor: **GÓMEZ MONTOYA, Héctor Erasmo**	
DNI: 70599170	Firma
ORCID: 0000-0002-1338-3392	

Agradecimientos

El trabajo realizado ha sido posible gracias al apoyo de grandes personas a las cuales extiendo mi más profundo agradecimiento en estas líneas.

A Dios, porque me muestra siempre que sus tiempos son perfectos y más aún sus planes.

A mis padres, a mi esposa y a mi familia que siempre están allí brindándome su apoyo incondicional para poder cumplir todos mis objetivos personales y profesionales.

A mi asesor, por sus enseñanzas, recomendaciones y paciencia para elaborar y conseguir los objetivos de este trabajo de investigación.

A cada uno de mis amigos profesionales, quienes brindaron sus conocimientos con la clasificación y evaluación de las oraciones, sin duda fue un arduo trabajo.

ÍNDICE

Spanish Detection of cyberbullying for the domain of theatrical text corpus applied to social networks using transfer learning and adversarial validation

Detección de ciberbullying en español para el dominio de corpus de texto teatrales aplicado a redes sociales usando transferencia de aprendizaje y validación adversarial

Esli Samuel Marquez, Hector Erasmo Gómez
Ingeniería Informática, Pontificia Universidad Católica del Perú
{esmarquez, hector.gomez}@pucp.edu.pe

Resumen: El aprendizaje de los modelos de detección de ciberbullying en redes sociales depende significativamente del conjunto de datos en cual fue entrenado lo que puede limitar su capacidad de generalización a otros conjuntos de datos. Este estudio propone un enfoque innovador utilizando transferencia de aprendizaje. Se desarrolló un modelo robusto de detección de ciberbullying basado en guiones teatrales, que ofrecen contextos ricos y variados. Para ello, se creó un corpus en español a partir de estos guiones, el cual fue meticulosamente etiquetado por expertos. Posteriormente, el modelo fue entrenado con este corpus para establecer una base de conocimiento que se aplicó luego a otros corpus de redes sociales. Los resultados mostraron una exactitud del 83% en las pruebas realizadas. Complementamos dicho modelo con una validación utilizando ejemplos adversarios, a partir de técnicas de data aumentada generamos más oraciones para fortalecer su capacidad de generalización, mejorando su desempeño tanto en su corpus como en distintos dominios de ciberbullying.
Palabras clave: discurso de odio, transferencia de aprendizaje, validación adversarial, modelos de lenguaje.

Abstract: The learning of cyberbullying detection models in social networks depends significantly on the data set on which it was trained, which can limit its generalization capacity to other data sets. This study proposes an innovative approach using transfer learning. A robust cyberbullying detection model was developed based on theatrical scripts, which offer rich and varied contexts. To do this, a Spanish corpus was created from these scripts, which experts meticulously labeled. The model was then trained with this corpus to establish a knowledge base that was then applied to other social media corpora. The results showed an accuracy of 83% in the tests carried out. We complement this model with a validation using adversarial examples, using augmented data techniques we generate more sentences to strengthen its generalization capacity, improving its performance both in its corpus and in different cyberbullying domains.
Keywords: hate speech, transfer learning, adversarial validation, large language model

1 Introducción

El ciberbullying o ciberacoso es fundamentalmente el maltrato emocional y humillación hacia otras personas en las diferentes redes sociales. Los insultos, mensajes burlones, difamaciones y rumores son algunos de los recursos empleados para causar sufrimiento a través de las plataformas en línea.

Sin embargo, gracias al procesamiento de lenguaje natural (NLP) que ha ganado mucha relevancia en los últimos años, siendo una de las áreas tradicionales de la Inteligencia Artificial (Feng y Shi, 2021); es que las computadoras pueden comprender textos

escritos por humanos y a su vez crear textos que los humanos puedan entenderlos.

Los modelos y soluciones existentes para la detección de ciberbullying están basados en modelos supervisados, donde dependen de la cantidad y calidad de datos donde el etiquetado por los expertos puede llegar a tener un costo elevado, además la fuente de datos de los mismos son poco textuales, ya que las redes sociales tienen interfaces de programación (API) en cierta medida limitadas y con restricciones para obtener dicha información; lo que puede generar un conjunto de datos de baja calidad y en consecuencia un modelo de detección de ciberacoso poco robusto. (Emmery et al., 2020).

En efecto, la mayoría de los estudios revisados sobre la detección del ciberacoso o discurso de odio (hate speech) arrojan resultados muy altos en sus conjuntos de pruebas, que van del rango de 0.45 hasta un 0.95 (H. Rosa et al., 2019); sin embargo, el aprendizaje obtenido resulta ser limitado para una generalización en otros conjuntos de datos u en otros dominios (Arango et al., 2019).

El ciberbullying es expresado en diferentes dominios, no solo en redes sociales, como lo son guiones de teatro, guiones de películas, shows de comedia (stand-up comedy), comentarios en publicaciones, mensajes de texto, foros en línea o salas de chat. Dado el contenido de los mismos los guiones de teatro son textuales y contextuales por lo que aprovecharemos dicho conocimiento para aplicarlo en el dominio de las redes sociales.

El objetivo de la investigación es construir un modelo robusto aplicando transferencia de aprendizaje (transfer learning) para la detección de ciberbullying en redes sociales aplicando validación contra ejemplos adversarios. (Karan y Snajder, 2018). Para alcanzar dicho objetivo proponemos las siguientes tareas:

En primer lugar, crear un corpus de oraciones a partir de los guiones de teatro en español, etiquetados por expertos.

En segundo lugar, desarrollar un modelo para detectar el ciberbullying en guiones de teatro con el objetivo de aprovechar el aprendizaje obtenido y aplicarlo a conjuntos de datos provenientes de redes sociales.

Por último, validaremos nuestro modelo mediante ejemplos adversarios creando un conjunto de datos con diferentes canales de evasión de ciberbullying para que mejorar la capacidad de generalización entre dominios de ciberbullying.

En la siguiente sección, proporcionamos un resumen de la revisión de los trabajos relacionados al problema de estudio. En la Sección 3, explicamos nuestro conjunto de datos obtenidos. En la Sección 4, informamos sobre nuestros experimentos realizado para luego en la Sección 5, proporcionar discusiones sobre análisis de errores y escalabilidad. Concluimos el estudio en la última sección.

2 Trabajos relacionados

Resumimos los trabajos relacionados sobre los métodos, conjuntos de datos y transferencia de aprendizaje para la detección del discurso de ciberbullying.

2.1 Detección de ciberbullying

Múltiples investigaciones estudian diferentes métodos y modelos relacionados a la detección de ciberbullying, ya sea odio o lenguaje ofensivo, en diferentes idiomas como inglés, español y turco (Cuzcano et al., 2020; H. Rosa et al., 2019; Arce-Ruelas et al., 2022). Algunos de estos modelos estaban basados en léxicos, aplicando modelos tradicionales lo cual lleva a estricta dependencia de los mismos, y otros basados en la semántica aplicando redes convolucionales o redes recurrentes en un aprendizaje supervisado.

2.2 Transferencia de aprendizaje en detección de ciberbullying

Se encontraron diferentes estudios acerca de modelos de detección de ciberbullying, los cuales fueron entrenados en conjuntos de datos específicos y validados en otros conjuntos de datos con clases similares en el idioma inglés. Karan y Snajder (2018), Arango et al. (2019), y Markov et al. (2021) detallan en sus experimentaciones que los modelos basados en SVM (Support Vector Machine) no logran generalizarse fuera de su dominio original de datos. Sin embargo, mencionan mejores resultados cuando los modelos son entrenados con conjuntos de datos provenientes de la enciclopedia digital Wikipedia, donde en una primera observación el contenido es más textual y las oraciones son de mayor longitud.

De similar forma, Cagri et al. (2022) experimentan con buenos resultados la

transferencia de dominios siendo estos dominios clasificaciones de ciberbullying como: género, raza, religión, política y deportes, para el idioma inglés y turco, realizándolo con modelos de última generación (transformers) donde concluyen que superan a los modelos convencionales.

Diversos estudios realizan la transferencia de dominio utilizando conjuntos de datos de distintos idiomas, aplicando la transferencia interlingüística para disminuir la influencia del lenguaje en la detección de odio o ciberbullying. (Pamungkas y Patti, 2019; Basile et al., 2019; Nozza, 2021).

2.3 Validación adversarial

Existen múltiples artículos que exploran técnicas para combatir ataques adversarios en el campo del procesamiento de lenguaje natural. Por ejemplo, Cardoso et al. (2022) construyen corpus paralelos proponiendo diferentes canales de errores para su investigación evaluando la robustez de un modelo frente a ataques con ejemplos adversarios. Aunque aún no se ha aplicado específicamente al dominio del ciberbullying, este estudio propone canales para abordar la evasión de ciberbullying.

3 Conjunto de Datos

3.1 Construcción del conjunto de datos

Para la creación del conjunto de datos, recopilamos 404 guiones teatrales de la web, como se muestra en el Anexo 1, durante los meses de junio – setiembre del 2023, con algunas características de género como comedia-humor, drama, romance, ficción, bullying; de público como infantiles, adolescentes, adultos; de región como dialectos de Perú, Argentina, México y España. Todos estos guiones libres de derecho de autor y obtenidas utilizando estrategias como web scrapping y procesamiento manual.

Generamos el corpus aplicando una subdivisión por oraciones. Luego realizamos un primer preprocesamiento basado en bigramas y trigramas para remover las palabras o frases que no aportan a nuestro modelo como, por ejemplo: “acto i”, acto ii”, “autor:”, “ambientación”, “personajes”, “título de la obra”, “abre el telón”, etc. Como resultado, obtuvimos 39625 oraciones con 657770 tokens.

3.2 Etiquetado del conjunto de datos

Los anotadores de datos participantes en este estudio fueron seleccionados por su experiencia y conocimiento en psicología, salud mental, lingüística y ciencias sociales como se detallan en la Tabla 1.

N°	Edad	Género	Especialidad
1	32	Femenino	Psicología
2	26	Femenino	Psicología
3	35	Masculino	Psicología
4	25	Femenino	Enfermería
5	31	Femenino	Enfermería
6	30	Masculino	Educación
7	33	Femenino	Educación

Tabla 1: Datos demograficos de los anotadores

A través de una convocatoria interna se reclutaron 7 profesionales, todos con experiencia en situaciones relacionadas con el bullying/ciberbullying.

El etiquetado se centró en la detección de ciberbullying, abarcando un lenguaje ofensivo como burlas, insultos, humillación, incitación al odio o violencia hacia individuos o grupos de personas. Las anotaciones se clasificaron en dos categorías: "HATE" para contenidos ofensivos y "NORMAL" para aquellos que no lo eran, como se muestra en la Tabla 2.

Oraciones	Clase
“Solo espero que no me contagie los hongos ese mugroso”	HATE/ OFFENSIVE
“Mi hijo sólo ha venido para presentarme a su novieta, una chica sin clase y de familia desconocida.”	HATE/ OFFENSIVE
“¡Ya se te están notando las canas, por vieja que eres”	HATE/ OFFENSIVE
“Ya no llore, Don Pedro, que me va a hacer llorar.”	NORMAL
“Pero le voy a tener que pagar.”	NORMAL

Tabla 2: Oraciones del conjunto de datos.

3.3 Estadísticas de los datos

Luego de un análisis del corpus generado a partir de los guiones de teatro, presentamos en la Tabla 3, algunas características de dicho corpus y en la Figura 1, se visualiza un primer

análisis de las palabras mas usadas en las oraciones etiquetadas como “HATE”.

Definición	Cantidad
Cantidad de oraciones	39625
Oraciones con ciberbullying	3544
Número de tokens	657770
Tamaño de vocabulario	46071

Tabla 3: Estadísticas del conjunto de datos.

Figura 1: Nube de palabras de adjetivos calificativos con etiqueta HATE/OFFENSIVE.

4 Experimentaciones

Tenemos principalmente dos experimentos, el primero consiste en analizar el rendimiento de varios modelos para la detección de ciberbullying en nuestro conjunto de datos de guiones de teatro. El segundo experimento implica evaluar nuestro modelo en diferentes conjuntos de datos públicos de redes sociales mediante validación adversarial.

4.1 Detección de ciberbullying en los guiones de teatro

En primer lugar, se toma el conjunto de datos compuesto por los guiones de teatro en lenguaje natural como entrada y se produce un texto mejorado como resultado. Al preprocesar el conjunto de datos, lo cual implica la eliminación de datos incompletos, ruido, errores y otros elementos no deseados, para facilitar una comprensión más profunda de la información.

Para esta fase, se realizaron una serie de modificaciones en el tratamiento del texto, que consistieron en: 1) Eliminación de los nombres al inicio de las oraciones, por ejemplo “RAFAELA: Eso no va a pasar nunca.”, 2) Eliminación de texto que representan emojis 3) Eliminación de caracteres especiales.

4.1.1 Diseño Experimental

Con el objetivo de evaluar nuestro conjunto de datos, ajustamos diferentes modelos de lenguaje basado en Transformers pre-entrenados en el idioma español.

Utilizamos la implementación de HuggingFace (Wolf et al., 2020), y tras analizar diversos modelos, se decidió trabajar con los siguientes:

- roBERTa-BNE (Gutiérrez-Fandiño et al, 2021): un modelo basado en RoBERTa entrenada en el corpus de la Biblioteca Nacional de España, datos en español.
- BETO (Cañete et al., 2020): modelo variante del modelo BERT entrenado exclusivamente para el idioma español.
- RoBERTuito (Pérez, J.M. et al., 2021): un modelo basado en RoBERTa entrenado para contenido generado por usuarios en español, en 500 millones de tweets

La distribución del conjunto de datos para la experimentación es la siguiente: 80% para entrenamiento, 20% para validación. Aplicamos la misma configuración para cada uno de los modelos

4.1.2 Resultados Experimentales

Luego del procesamiento y validación, se obtuvieron los siguientes resultados para las clases definidas en el etiquetado, como se muestra en la Tabla 4 para cada uno de los modelos entrenados

Métrica	roBERTa-BNE	BETO	RoBERTuito
Precisión	0.917	0.911	0.913
F1-score	0.911	0.903	0.909
Sensibilidad	0.906	0.897	0.906

Tabla 4: Resultados de métricas obtenidas.

De acuerdo a los resultados de la Tabla 4, los modelos obtienen valores muy similares y comparables. Sin embargo, según la métrica f1-score, que es adecuada para casos de desequilibrio de clases, se seleccionó el modelo basado en roBERTa-BNE por tener un valor mayor. Este modelo se utilizó para continuar con el entrenamiento y su posterior aplicación a los conjuntos de datos de redes sociales listados en la Tabla 5.

4.2 Detección de ciberbullying en redes sociales.

En la Tabla 5, se visualizan los conjuntos de datos de las redes sociales que evaluaremos, destacando algunas de sus características

- SemEval-2019 Task 5 - spanish (Basile et al., 2019).
- HateCheck - Spanish (Röttger et al., 2022)
- HaterNet (Pereira-Kohatsu et al., 2019).

Conjunto de Datos	Cantidad de registros	Porcentaje de ciberbullying
SemEval-2019 Task5 – spanish	6600	41.2%
HateCheck - spanish	3745	70.3%
HaterNet	6000	26.1%

Tabla 5: Conjuntos de datos de redes sociales.

4.2.1 Canales de evasión de ciberbullying

Detallamos, los canales de evasión de ciberbullying para construir un corpus de oraciones adversariales.

- Error tipográfico (ErrTipo): En este canal, se modifican las palabras a través de replicar excesivamente letras o caracteres de una palabra. Solo se realiza esta alteración a las palabras catalogadas como sustantivo y/o adjetivos.

Algoritmo Error Tipográfico
Entrada: w w = "*anormal"* **Salida**: pEva 1. (vocal, posicion) = EncontrarUltimaVocal(w) vocal= 'a'; posicon = 6 2. pEva = w[0 : posicion− 1] + vocal*3 + w[posicion + 1 :] pEva = "anormaaaal" 3. **return** pEva

Tabla 6: Algoritmo error tipográfico.

- Ambigüedad fonema-grafema (Amb Homof): Este canal se fundamenta en el similar sonido de los grafemas (Cardoso et al., 2022) evadiendo las validaciones de dichas palabras.

Algoritmo Ambigüedad Fonema-grafema
Entrada: w w = "*cabezón"* **Salida**: pEva 1. grafema = EncontrarGrafema(w) grafema = "ca" 2. grafemaAmb = GrafemaAmbiguo(grafema) grafemaAmb = "k" 3. pEva = ReemplazarGrafema(w) pEva = "kbezón" 4. **return** pEva

Tabla 7: Algoritmo error tipográfico.

- Palabras coloquiales: En este canal, se basa en el uso de palabras con modismos del lenguaje español, para ello se llevó a cabo una recolección de dichas palabras para la creación de un diccionario de fuente propia.

Algoritmo Palabras coloquiales
Entrada: w w = "*casa"* **Salida**: pEva 1. pEva = EncontrarPalabraColoquial(w) pEva = "jato" 3. **return** pEva

Tabla 8: Algoritmo palabras coloquiales.

4.2.2 Diseño Experimental

En investigaciones similares a las que se describen en este artículo, se emplean un conjunto de datos de texto extraídos de redes sociales. Estos textos presentan rasgos distintivos debido al estilo de escritura característico de los usuarios, que puede incluir errores ortográficos, abreviaciones y otros aspectos de escritura peculiar. Estos textos suelen ser clasificados como de baja calidad lingüística. Por lo tanto, es necesario llevar a cabo un proceso de preprocesamiento similar a los estudiados con el fin de elevar la calidad de los datos.

En esta fase, los conjuntos de datos se preprocesan de la misma manera que los textos originales utilizados para crear el modelo. Además, todas las URL, menciones usuario y números se eliminaron. A continuación, se detallan los experimentos realizados.

- Experimento Base: Definimos el conjunto de datos SemEval-2019 de la

Tabla 5. como un corpus “base”. La distribución de los conjuntos de datos para la experimentación: 500 para pruebas, del restante 80% para entrenamiento, 20% para validación. Se realizó el entrenamiento del modelo con estos datos y se realizó pruebas en los otros 2 conjuntos de datos de la Tabla 5.

- Experimento Adversarial: Luego, se creó un corpus a partir del corpus “base” con cada uno de los canales de evasión de ciberbullying, obteniendo 18768 oraciones. La distribución realizada fue similar al experimento “base”, los mismos 500 oraciones de pruebas y 80% para entrenamiento, 20% para validación del nuevo conjunto de datos. Se realizó el entrenamiento de un modelo con estos datos y se realizó pruebas en los 2 conjuntos de datos adicionales de la Tabla 5.

4.2.3 Resultados Experimentales

Obtuvimos los siguientes resultados para cada uno de los experimentos propuestos (ver Tabla 9 y 10).

- Experimento Base.

Métrica	Validación	Pruebas
Exactitud	**0.8238**	**0.8360**
Precisión	0.7953	0.8125
F1-score	**0.7915**	**0.8161**
Sensibilidad	0.7876	0.8198

Tabla 9: Resultados obtenidos para experimento base.

Se realizó pruebas en los otros conjuntos de datos, los cuales el modelo no conocía, obteniendo en el corpus HateCheck, el 62.08% y en Haternet el 72.13% de predicciones correctas.

- Experimento Adversarial

Métrica	Validación	Pruebas
Exactitud	**0.9702**	**0.8520**
Precisión	0.9622	0.8558
F1-score	**0.9615**	**0.8279**
Sensibilidad	0.9609	0.8018

Tabla 10: Resultados obtenidos para experimento adversarial.

De igual forma se realizó pruebas en los otros conjuntos de datos, obteniendo en el corpus HateCheck, el 65.61% y en Haternet el 71.33% de predicciones correctas.

5 Análisis y Discusiones

Se tomó como conjunto de datos base el corpus de Basile et al. (2019), un corpus que principalmente tiene como destinatarios de odio a las mujeres e inmigrantes. A partir del aprendizaje obtenido con los guiones de teatro y el entrenamiento con dicho conjunto de datos se observó una mejora en el desempeño de los resultados obtenidos de la literatura en dicho corpus. Además, se obtuvieron resultados aún mejores cuando se validó con el modelo entrenado partir de datos adversariales.

Adicionalmente se realizaron pruebas en los conjuntos de datos restantes de la tabla 5, obteniendo resultados aceptables considerando que estos eran desconocidos para el modelo. Profundizando en ello, en la tabla 11 se observan algunos errores de clasificación al realizar pruebas en los conjuntos de datos Haternet y HateCheck enfocada a diversas formas de ciberbullying como etnia, raza, género, discapacidad, política y religión.

Oración	Etiqueta	Predicción
"el mejor momento del día es irme a la cama mamona"	Normal	Hate
“Buah yo no me haría nunca un twitter, eso está lleno de rojos de porqueria”	Hate	Normal

Tabla 11: Tabla de análisis de errores de predicción.

Una posible causa de esto podría ser la naturaleza del corpus base en comparación con los demás corpus. Por lo tanto, realizar experimentos de diferentes perspectivas o contextos, podría enriquecer mucho más al modelo.

6 Conclusiones

Construimos un conjunto de datos para la detección de ciberbullying en el lenguaje español basado en guiones de teatro y al criterio de expertos, para luego analizar el desempeño en los modelos de última generación y obtener un aprendizaje en cuanto a texto y contexto.

Se experimentó la aplicación del aprendizaje adquirido del análisis de guiones teatrales a la detección de ciberbullying en plataformas de redes sociales, observándose adaptabilidad y generalización del modelo en diferentes corpus.

Y de forma conjunta con una validación adversarial, la propuesta de canales para la creación de datos adversariales demostró la robustez del modelo para identificar posibles riesgos asociados con la evasión de mensajes de contenido de ciberbullying. Los resultados mostraron que al probar con nuevos corpus que no habían sido utilizados en el entrenamiento, el modelo mantuvo su efectividad sin necesidad de ser reentrenado con dichos corpus.

Finalmente, creemos en la mejora continua del modelo mediante la realización de más experimentos, como la definición de otro corpus base y la inclusión de datos más diversos y actualizados para un mejor entrenamiento. Asimismo, proponemos mejorar los canales de evasión, incluyendo el análisis del contexto para palabras coloquiales y la generación de más canales para la evaluación adversarial, ya que las formas de evasión del ciberbullying continúan incrementándose.

Bibliografía

Allen, J., D. Byron, M. Dzikovska, G. Ferguson, L. Galescu, y A. Stent. 2000. An Architecture for a Generic Dialogue Shell. *Natural Language Engineering*. To appear.

Arango, A., Perez, J., and Poblete, B. (2019). 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, SIGIR'19, pages 45–54, New York, NY, USA. Association for Computing Machinery.

Arce-Ruelas, K. I., Alvarez-Xochihua, O., Pellegrin, L., Cardoza-Avendaño, L., y González-Fraga, J. Ángel. (2022). Automatic Cyberbullying Detection: a Mexican case in High School and Higher Education students. IEEE Latin America Transactions, 20(5), 770–779.

Basile Valerio, Bosco Cristina, Fersini Elisabetta, Nozza Debora, Patti Viviana, Pardo Francisco, Rosso Paolo y Sanguinetti Manuela. 2019. 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, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Cagri Toraman, Furkan Şahinuç, and Eyup Yilmaz. (2022). Large-Scale Hate Speech Detection with Cross-Domain Transfer. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 2215–2225, Marseille, France. European Language Resources Association.

Cañete, J., Chaperon, G., Fuentes y Pérez, J. (2020). Spanish pre-trained bert model and evaluation data.

Cuzcano, X.M. y Ayma, V.H. (2020). A comparison of classification models to detect cyberbullying in the Peruvian Spanish language on twitter. International Journal of Advanced Computer Science and Applications, 11(10), 132-138.

Cardoso Gerardo, Gomez Erasmo, Oncevay Arturo. (2022). Corrector ortográfico neuronal para errores ortográficos multilingües adversarios para lenguas amazónicas peruanas.

De la Rosa, J., Ponferrada, E. G., Villegas, P., Salas, P. G. D. P., Romero, M., y Grandury, M. (2022). Bertin: Efficient pre-training of a spanish language model using perplexity sampling.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emmery, C., Verhoeven, B., De Pauw, G. et al. (2021). Current limitations in cyberbullying detection: On evaluation criteria, reproducibility, and data scarcity. Lang Resources & Evaluation 55, 597–633

Feng, H., y Shi, F. (2021). Deep Learning in Natural Language Processing, edited by Li Deng and Yang Liu. Singapore: Springer, 2018. ISBN 9789811052088. XVII 329

pages. Natural Language Engineering, 27(3), 373-375.

Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., y Villegas, M. (2021). Maria: Spanish language models

H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J. P. Carvalho, S. Oliveira y I. Trancoso, "Automatic cyberbullying detection: A systematic review". Computers in Human Behavior, vol. 93, pp. 333-345, 2019.

Karan, M. and Snajder, J. (2018). Cross-domain detection of abusive language online. In Proceedings of the 2nd Workshop on Abusive Language Online (ALW2), pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Markov, I., Ljubešić, N., Fišer, D., and Daelemans, W.(2021). Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 149–159, Online. Association for Computational Linguistics.

Nozza, D. (2021). Exposing the limits of zero-shot cross-lingual hate speech detection. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), pages 907–914, Online. Association for Computational Linguistics.

Pamungkas, E. W. and Patti, V. (2019). Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Pereira-Kohatsu, Juan Carlos, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. (2019). "Detecting and Monitoring Hate Speech in Twitter".

Pérez, J.M., Furman, D.A., Alemany, L.A., y Luque, F.M. (2021). RoBERTuito: a pre-trained language model for social media text in Spanish. International Conference on Language Resources and Evaluation.

Röttger, P., Seelawi, H., Nozza, D., Talat, Z., y Vidgen, B. (2022). Multilingual HateCheck: Functional Tests for Multilingual Hate Speech Detection Models.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). 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. Association for Computational Linguistics.

A Anexo 1: Sitios web de guiones teatrales.

1. Comediathèque. Obras en español. https://comediatheque.net/traductions/obras-en-espanol/
2. Obras de teatro cortas para jóvenes. https://www.obrascortas.com/obras-de-teatro-cortas-para-jovenes/
3. Obras de Teatro Cortas. https://obrasdeteatrocortas.org/
4. Guiones de teatro. Blogspot. http://guionesdeteatro.blogspot.com/2009/11/
5. Dramaturgia y escuela. Universidad Nacional de Cuyo. https://bdigital.uncu.edu.ar/objetos_digitales/2905/dramaturgiayescuela2.pdf
6. 10 guiones de obras de teatro cortas: Ejemplos gratis. https://es.slideshare.net/portesp/10-guiones-de-obras-de-teatro-cortas-ejemplos-gratis
7. Obras de teatro en PDF. Pedagogía Milennial. https://www.pedagogiamilennial.com/obras-de-teatro-en-pdf/
8. Club de Escritura. https://clubdeescritura.com/
9. Obra de Teatro Virtuales. https://obra-de-teatro-virtuales.webnode.es/
10. Relatos Cortos. https://relatoscortos.org/
11. Biblioteca Antológica. https://www.biblioteca-antologica.org/