Clasificación de comentarios de odio LGBT en Twitter

Paola Garcia, *Member, IEEE,* Mirella Jimenez, *Member, IEEE.*

***Abstract*—Las redes sociales han traído grandes beneficios a la humanidad, pues permiten la interacción rápida y sencilla entre diversos usuarios. Sin embargo, estas a menudo son utilizadas para emitir comentarios mal intencionados. La comunidad LGBT es una de las más perjudicadas, pues constantemente reciben comentarios o discursos de odio en las diversas redes sociales. El objetivo del presente trabajo es crear un modelo, mediante Aprendizaje Supervisado, que nos permita detectar los comentarios de odio dirigidos a la comunidad LGBT en Twitter. Los resultados experimentales arrojaron que con el uso del algoritmo de aprendizaje automático Super Vector Machine (SVM) y la técnica de extracción de características Term Frecuency (TF), se puede obtener un 73% de precisión en el modelo. Nuestro estudio tiene implicancias prácticas en la detección de comentarios de odio dirigido a la comunidad LGBT en Twitter y servirá de base para futuras investigaciones que decidan dirigir su trabajo en beneficio de otras comunidades y en distintas plataformas.**

***Index Terms*—Hate Speech, Twitter, Machine Learning, Term Frequency, SVM**

I. Introduccion

Las redes sociales permiten la interacción entre diversos usuarios, su uso es sencillo y representa una fuente de comunicación rápida, lo que resulta beneficioso para las personas. No obstante, estas tecnologías también son utilizadas como medio para emitir comentarios o discursos de odio, que generan un efecto no deseado como sentirse más negativos cuando los leen, lo que puede conllevar a generar síntomas de depresión o trastornos de ansiedad (Cuncic, 2021)[1]. Asimismo, según un estudio reciente, uno de cada cinco usuarios víctimas de acoso en línea(22%) informaron que su última experiencia de hostigamiento ocurrió en la sección de comentarios de las redes sociales, siendo nombrada como las sección más frecuente de acoso (Anderson, 2014)[2]. Por tal, los comentarios y/o discursos de odio en redes sociales son una problemática relevante a considerar y analizar.

Una de las comunidades más afectadas al momento de recibir comentarios de odio es la comunidad LGBT. Los comentarios de odio tienen como objetivo socavar la dignidad y valor de un ser humano perteneciente a un grupo social en particular (Ilga-Europe, s.f.)[3]. En específico, un mensaje negativo a la comunidad LGBT, sus partidarios o el resto de la sociedad, supone que un grupo social no merece reconocimiento, respeto o igualdad (Ilga-Europe, s.f.)[3]. El 23% de la población mayoritaria LGBT declaran haber sido expuestos a comentarios odiosos en redes sociales, lo que resulta más del doble de la población mayoritaria no LGBT (Hoeg, 2019)[4].

En el Perú, la mayoría de estos ciberataques se dan en redes sociales como Facebook o Twitter. Asimismo, estas plataformas (en particular Twitter) no revisan ni quitan el contenido potencialmente ofensivo, lo que resulta beneficioso y alentador para las personas acostumbradas a dejar este tipo de comentarios.

En el presente informe se realiza un análisis que permite detectar discursos de odio hacia la comunidad LGBT en redes sociales, siendo Twitter la red social elegida. En específico, nos centraremos en extraer comentarios y tweets a través de la técnica de 'scraping', con ayuda de palabras claves relacionadas al tema a tratar. La finalidad es obtener información directamente conectada con nuestro objetivo.

II. Estado del arte

Hate speech classification in social media using emotional analysis

El trabajo realizado por Martins, Gomes, Almeida, Novais y Heriques (2018)[5] busca detectar los discursos de odio en línea utilizando enfoques léxicos y emocionales, con el objetivo de automatizar su detección y, en consecuencia, su mitigación. Este estudio utiliza el conjunto de datos proporcionado por Davidson y Warmsley, que proporciona un conjunto de 24782 tweets previamente clasificados en: discurso de odio (1430), lenguaje ofensivo (19190), ninguno (4162). Con esto, construyeron el dataset con técnicas de

NLP, identificaron los N-gramas, removieron las stopwords y aplicaron TF-IDF para clasificar las palabras más relevantes del conjunto de datos, esto les hizo notar que las palabras más frecuentes eran agresivas u ofensivas. Además, en esto último, utilizaron una comparación de las palabras frecuentes con la lista de palabras identificadas en el léxico de Hatebase como de odio. Con ello, obtuvieron 975 tweets pre procesados, de la data, para cada categoría. Para calcular la intensidad de la emoción ira, utilizaron el paquete Syuzhet de R y el léxico NRC-Intensity. Aquí obtuvieron un primer resultado que muestra que el discurso de odio es menos intenso que el lenguaje ofensivo (0,29 frente a 0,51). A partir de ello, obtuvieron 14 dimensiones de análisis en un nuevo dataset, para después, seleccionar solo las dimensiones más útiles. Para la clasificación, utilizaron el programa Weka, con los algoritmos más importantes de clasificación de textos: SVM, Naive Bayes, Random Forest. Los resultados muestran que el algoritmo que mejor clasifica los discursos de odio fue el Random Forest (80.64%), seguido del SVM (80.56%) y, finalmente, el Naive Bayes (71.33%)

Automatic Hate Speech Detection using Machine Learning: A Comparative Study

En este artículo, Abro, Shaikh, Ali, Khan, Mujtaba y Hussain Khand (2020)[6] buscan comparar el rendimiento de tres técnicas de ingeniería de características y ocho algoritmos de aprendizaje automático para un conjunto de datos, con el fin de detectar discursos de odio.El conjunto de datos que utilizaron son tweets recopilados y etiquetados por CrowdFlower que se dividen en tres clases: tweets que incitan al odio (16%) , tweets no ofensivos (50%), tweets ofensivos que no incitan al odio (33%). En el preprocesamiento, cambiaron los tweets a minúsculas, eliminaron todas las URL, nombres de usuario, espacios en blanco, hashtags, puntuaciones y palabras de parada utilizando técnicas de concordancia de patrones de los tweets recogidos, y después realizaron la tokenización y el stemming de los tweets. Para aplicar las técnicas de machine learning, la data debía tener características numéricas para ser entendidos por las reglas de clasificación. Para ello, en el estudio utilizaron tres técnicas de ingeniería de características: N-gram con TFIDF, Word2vec y Doc2vec. Una vez con el dataset limpio y listo para aplicar machine learning, las técnicas que utilizaron para este estudio comparativo fueron NB, SVM , KNN , DT , RF, AdaBoost , MLP y LR. Los resultados muestran que la menor precisión (0,58), menor recall (0,57) y menor accuracy (57%) se encontró utilizando los clasificadores MLP y KNN y la técnica de ingeniería de características TF-IDF; y, los mejores resultados se obtuvieron utilizando como clasificador el SVM y como técnica de características el TF-IDF, recall (0.79), precisión (0.77), accuracy (79%).

Hate Speech Detection Using Natural Language Processing Techniques

Este trabajo, realizado por Biere (2018)[7], busca averiguar cómo las técnicas de Natural Languages Processing pueden contribuir a detectar los discursos de odio de las redes sociales (en específico Twitter). El trabajo se centra en explorar y aplicar un método eficaz para esta tarea de clasificación. La

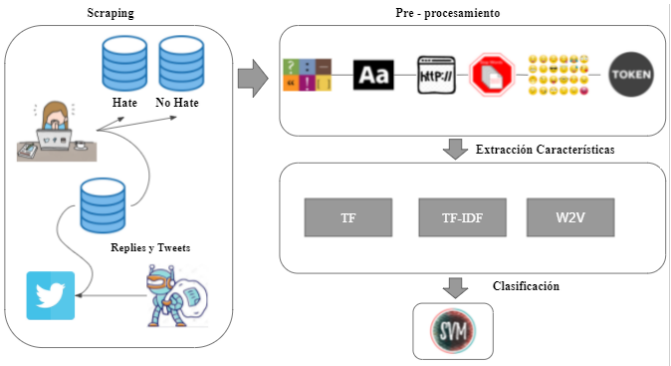

Fig. 1. Metodología de detección de discurso de odio.

data que se utilizó es el dataset distribuido por CrowdFlower, el que contiene 24,783 tweets en Inglés clasificados en: odio, ofensivo y ninguno. Para el preprocesamiento de los datos se realizó un proceso de normalización que incluía: remover caracteres especiales, pasar a minúscula, steaming para reducir la inflexión de palabras, para después tokenizar el texto. Para la clasificación se utilizaron librerías de redes neuronales y machine learning, en específico para el entrenamiento del modelo, se utilizó la librería Keras con Tensor Flow back-end. El método de aprendizaje profundo que se utilizó fue Convolutional neural network (CNN), pues su arquitectura permitió obtener un accuracy del 91%, respaldado además por trabajos previos, lo que denota un buen rendimiento. No obstante, concluyen que es necesario calidad en los datos para que el modelo elegido funcione de forma óptima.

III. Metodología

A. Proceso

En esta sección, se explica la arquitectura y las técnicas utilizadas con el objetivo de detectar discursos de odio en comentarios y tweets de Twitter. A través de 3 etapas, 'scraping' u obtencion de datos, pre-procesamiento, extracción de características, y clasificación de los datos.Como se puede observar en la figura N°1, Cada una de estas partes serán detalladas en las siguientes secciones.

B. Obteccion de la Data

Para nuestra primera etapa se recopiló tweets y comentarios de la red social Twitter a través de la técnica de scraping. Para este punto se utilizó la librería BeatifulSoap y Selenium de Python. Estas dos librerías apoyaron en la recolección de alrededor 43 000 discursos en total. Sin embargo, al pasar por un primer proceso de limpieza para obtener comentarios o tweets únicos se obtuvo un total de 15 000 discursos diferentes.

La extracción tuvo como inpust palabras claves como *'gay'*, *'LGBT'*, *'lesbiana'*, que hacen referencia directamente a la comunidad en estudio. Además, de recolectar información de usuarios como *@MarchaOrgulloPe* y *@LgbtPeru*. Los datos se almacenaron en un formato *'xlsx'* lo que posteriormente, facilitó la actividad de etiquetar manualmente cada documento con la finalidad de identificar discursos de odio y no odio.

Cada autor decidió etiquetar los 15 000 documentos a consideración propia, los cuales finalmente fueron comparados y clasificados. Cabe destacar, que los documentos etiquetados como discurso de odio fue nuevamente revisado por un integrante de la comunidad LGBT para una mayor precisión.

C. Pre-procesamiento

Una vez obtenido el dataset etiquetado, empezamos con el pre-procesamiento de los datos. Lo primero a realizar fue pasar todos los datos a minúscula, esto con el fin de tener las palabras de forma estándar. Después, se eliminó los urls de los tweets y comentarios, pues estos no eran necesarios para nuestro análisis. Como el objetivo es poder clasificar adecuadamente los comentarios ofensivos, consideramos que los emojis que se utilizan en las redes sociales debían ser considerados; para ello, con el uso de la librería "emoji" de python se pudo desmojizar cada dato y tenerlo en forma de escrita, según la emojipedia. Posteriormente, se eliminaron las etiquetas (@) y los hashtag (#) , y nos quedamos únicamente con las palabras. Para hacerlo más estándar, se procedió a eliminar las tildes del conjunto de datos. Finalmente, removimos los stopwords y nos quedamos con las palabras mayores a dos letras.

Cabe mencionar, que para el uso de técnica de extracción de características *'Word2Vec'* se aplicó la tokenización en esta etapa.

D. Extracción de características

En esta etapa, decidimos utilizar tres técnicas de extracción de características: *'Term Frecuency (TF)'*, *'Term Frequency — Inverse Document Frequency (TF-IDF)'* y *'Word2Vec'*. Esto con el fin de evaluar cuál de las tres te permitía obtener un mejor modelo en la etapa posterior.

El Term Frecuency (TF) nos permite medir la frecuencia con la que aparece un término en un documento determinado, lo que nos permite caracterizarlos. En teoría, si una palabra aparece más frecuentemente en un documento, más lo caracterizará.

La segunda técnica (TF-IDF) tiene mucha relación con la primera, la diferencia radica en que esta mide el peso de la palabra en el documento dado, considerando la frecuencia con la que se repite en todos los documentos. Es decir, su valor va a aumentar a medida que la palabra aparezca en el documento, pero será compensada por la frecuencia de la palabra en la colección de documentos.

Estas dos primeras técnicas son muy comunmente utilizadas para minería de textos y machine learning.

Por último, la tercera técnica de extracción de características usada fue Word2Vec el cual toma como input palabras de un corpus de textos y a través de ellas aprende a dar una representación vectorial. Este algoritmo crea vectores que se definen a través de la funcion de similitud de coseno y representan la similitud semántica entre palabras. Es decir, dos palabras similares estarán ubicadas una cerca a la otra, mientras que palabras diferentes se encontrarán a distancias lejanas en el espacio vectorial. Para este estudio y con el apoyo de redes neuronales al aplicar la técnica se obtuvo vectores característicos por cada documento.

	NO predictivo	SI Predictivo
Actual NO	TN	FP
Actual SI	FN	TP

Fig. 2. Matriz de confusión

TF	No Hate	Hate
No Hate	79	21
Hate	32	68

Fig. 3. Matriz de confusión tf

E. Clasificación

En esta sección, se procedió a utilizar *'Super Vector Machine (SVM)'*.Esta técnica consiste en un conjunto de algoritmos de aprendizaje supevisado que te permite realizar clasificación y regresión, dado una base de datos adecuadamente etiquetada. Además, para aplicar el algoritmo de clasificación, se utilizó la librería *sklearn*.

Lo primero que se hizo fue dividir la data utilizando la regla del 80-20; es decir, el 80% fue destinada para la parte del entrenamiento y el 20% restante para el testeo. Como en la parte de etiquetado se logró obtener 504 valores clasificados como "comentario o discurso de odio", 400 fueron destinados para entrenamiento y 100 para testeo. Lo mismo se hizo para la otra parte de la data, utilizando la misma cantidad para el entrenamiento y para el testeo.

Como se mencionó anteriormente, se evaluaron tres técnicas de extracción de características con el algoritmo de aprendizaje automático. La primera que se utilizó fue Term Frecuency (TF), luego TF-IDF y finalmente W2V. En total, se realizaron tres análisis, para comparar la eficacia de los modelos de clasificación.

El rendimiento del clasificador lo hemos evaluado utilizando la matriz de confusión, que nos arroja los verdaderos negativos (TN), falsos positivos (FP), falsos negativos (FN) y verdaderos positivos (TP), tal como se muestra en la figura N°2.

Asimismo, la medida que se utilizó fue la precisión o accuracy, pues midel el porcentaje de casos que el modelo ha acertado.

IV. Resultados

A. Resultados del modelo con la técnica TF

La primera técnica nos arrojó un accuracy del 73%. Esto quiere decir que el 73% de los datos fueron clasificados correctamente. Además, la matriz de confusión quedó como en la Figura N°3.

Se observa que de los valores de testeo, 79 que eran "No hate" fueron clasificados de forma correcta como "No hate" y 68 fueron clasificados de forma correcta como "Hate".

B. Resultados del modelo con la técnica TF - IDF

La segunda técnica nos arrojó un accuracy menor al de la primera, pues fue del 67%. Esto quiere decir que el 67% de

TF - IDF	No Hate	Hate
No Hate	69	31
Hate	41	59

Fig. 4. Matriz de confusión tfidf

W2V	No Hate	Hate
No Hate	70	30
Hate	31	69

Fig. 5. Matriz de confusión w2v

los datos fueron clasificados correctamente. Con ello, la matriz de confusión quedó con en la Figura N°4.

Se observa que de los valores de testeo, 69 que eran ”No hate” fueron clasificados de forma correcta como ”No hate” y 59 fueron clasificados de forma correcta como ”Hate”.

C. Resultados del modelo con la técnica Word2Vec

La última técnica nos arrojó un accuracy menor al de la primera, pero mayor a la segunda, pues fue del 69.5%. Esto quiere decir que el 69.5% de los datos fueron clasificados correctamente. Con ello, la matriz de confusión quedó con en la Figura N°5.

Se observa que de los valores de testeo, 70 que eran ”No hate” fueron clasificados de forma correcta como ”No hate” y 69 fueron clasificados de forma correcta como ”Hate”.

V. Conclusión y futuros trabajos

A. Conclusion

En conclusión, se observo que el modelo que nos brindó un mejor rendimiento fue el que utilizó la técnica de extracción de características TF. Esto se debe a que esta utiliza un conteo de la frecuencia de las palabras en el documento, pero no los pondera con la frecuencia en el conjunto de documentos. Esto, podría explicar como con el TF-IDF se tiene un menor accuracy, pues esta obtiene el peso de la palabra considerando la frecuencia en el conjunto de documentos.

Además, consideramos que como la cantidad de data es pequeña, no resulta tan beneficioso para la técnica word2vec, pues con una data mayor se tendría más palabras en su vocabulario y el entrenamiento de la red sería mucho mejor.

En general, consideramos que el análisis realizado en este trabajo proporciona una herramienta práctica en la detección de comentarios o discursos de odio dirigido a personas de la comunidad LGBT y que, adicionalmente, es un gran paso para comenzar a erradicar el uso de las redes sociales para este tipo de actos.

B. Futuros trabajos

Finalmente, para futuros trabajos recomendamos que se utilice una mayor cantidad de datos etiquetados, pues con ello se podría tener un mejor rendimiento del modelo.

Asimismo, para la parte del etiquetado de los datos, recomendamos que se utilice la ayuda de más personas pertenecientes la comunidad LGBT, pues muchas veces el criterio del investigador podría no ser el más adecuado.

References

[1] A. Cuncic, “Mental health effects of reading negative comments online.” *Very Well Mind*, 2021. [Online]. Available: https://www.verywellmind.com/mental-health-effects-of-reading-negative-comments-online-5090287

[2] M. Anderson, “About 1 in 5 victims of online harassment say it happened in the comments section,” 2014. [Online]. Available: https://www.pewresearch.org/fact-tank/2014/11/20/about-1-in-5-victims-of-online-harassment-say-it-happened-in-the-comments-section/

[3] “Hate crime hate speech,” *ILGA Europe*. [Online]. Available: https://www.ilga-europe.org/what-we-do/our-advocacy-work/hate-crime-hate-speech

[4] E. Hoeg, “One of four lgbt people experience hate speech,” *Science Norway*, 2019. [Online]. Available: https://sciencenorway.no/forskningno-gender-and-society-norway/one-of-four-lgbt-people-experience-hate-speech/1553837

[5] J. J. A. P. N. Ricardo Martins, Marco Gomes and P. Henriques, *Hate Speech Classification in Social Media Using Emotional Analysis.*, 2018.

[6] Z. A. S. K. G. M. Sindhu Abro, Sarang Shaikh and Z. H. Khand, *Automatic Hate Speech Detection using Machine Learning: A Comparative Study.*, 2020.

[7] S. Biere, *Hate Speech Detection Using Natural Language Processing Techniques.*, 2018. [Online]. Available: https://science.vu.nl/en/Images/werkstuk-biere$_tcm296-893877.pdf$