Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

Tipo de artículo: Artículos originales
Temática: Inteligencia artificial
Recibido: 10/07/2022 | Aceptado: 28/08/2022 | Publicado: 30/09/2022

Identificadores persistentes:
ARK: ark:/42411/s9/a65
PURL: 42411/s9/a65

Uso de una herramienta de NLP aplicada a la detección del ciberacoso en Twitter

Use of an NLP tool applied to the detection of cyberbullying on Twitter

Jonathan Aguirre Soto[1], Hector Ávila Gonzales[2], Valeria Bravo Saines[3]

[1] Universidad La Salle. Arequipa, Perú. jaguirres@ulasalle.edu.pe
[2] Universidad La Salle. Arequipa, Perú. havilag@ulasalle.edu.pe
[3] Universidad La Salle. Arequipa, Perú. vbravos@ulasalle.edu.pe

[*] Autor para correspondencia: jaguirres@ulasalle.edu.pe

Resumen

En este documento se dará un breve resumen de como en la actualidad el constante desarrollo de la información y las tecnologías de comunicación (TICs) ha cambiado la interacción entre las personas hoy en día, por lo que las experiencias reales se han trasladado a un método virtualizado en este caso internet. Aunque las barreras de espacio-tiempo de la comunicación tradicional se han fragmentado, las relaciones sociales se han vuelto más fuertes, pero surgen nuevos problemas relacionados con diferentes conductas. El acoso, se define como un acto que amenaza el bienestar de una persona, y se convierte en ciberacoso cuando es realizado a través de internet generando a gran escala problemas de ansiedad, depresión e incluso el acto de suicidio y por lo cual es fundamental detectar a tiempo estos comportamientos malignos. Haremos uso de una herramienta de Procesamiento de Lenguaje Natural (NLP) utilizando Twitter como base para la extracción de las bases de conocimiento.

Palabras clave: Twitter, NLP, procesamiento de lenguaje natural, ciber-acoso, TICs, tecnologías de la información.

Abstract

This paper will briefly overview how the constant development of information and communication technologies (ICTs) has changed the interaction between people today. Real experiences have been transferred to a virtualized method, in this case, the internet. Although the space-time barriers of traditional communication have been fragmented, social relations have become more assertive, but new problems related to different behaviors arise. Bullying is defined as an act that threatens the well-being of a person and becomes cyberbullying when it is carried out over the internet, generating large-scale problems of anxiety, depression, and even the act of suicide, which is why it is essential to detect these malicious behaviors in time. We will use a Natural Language Processing (NLP) tool using Twitter as the basis for the extraction of knowledge bases.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

Keywords: *Twitter, NLP, natural language processing, cyber-bullying, TICs, information technologies.*

Introducción

El ciberacoso es un acto reiterado que acosa, humilla o amenaza a las personas a través de sus ordenadores, teléfonos celulares, laptops, tabletas y otros dispositivos electrónicos, por otro lado también se incluyen sitios sitios web que mantienen activas las redes sociales. Ciberacoso a través de internet utilizando las redes sociales, se ha vuelto más peligroso que el acoso tradicional porque tiene la capacidad de amplificar el daño y humillación a un grupo de personas que están conectadas en línea.

Muchas víctimas del acoso cibernético se sienten tristes, deprimidas, frustradas, incluso tienen pensamientos suicidas. Una encuesta realizada por UNICEF y el Ministerio de Comunicación e Información, en el Perú existe el cyberbullying y se encuentra en tasas de hasta un 40% entre los 13-15 años existe el bullying tradicional y de 11 a 14 años hay una tasa de incidentes de ciberacoso. Algunos de ellos podrían ser los acosadores; sin embargo, reconocen el peligro y los efectos negativos que llega a generar.

Por lo tanto, debería haber una forma de detectar a los principales actores del ciberacoso antes de que se den cuenta. reportar los daños que han causado en tiempo real. Por eso debe haber reglas para poder categorizar textos de ciberbullying. Por ello analizaremos la detección del ciberacoso mediante una herramienta de PNL usando un clasificador de texto.

Materiales y métodos o Metodología computacional

Motivación

El objetivo de este trabajo será desarrollar una herramienta para poder detectar las agresiones que se producen durante el uso de la aplicación utilizando lenguaje de procesamiento natural y un clasificador de texto a partir de datos de Twitter que nos dice que si es un comentario sospechoso o es un comentario correcto mostrar a través de una clasificación binaria. Las preguntas que nuestro proyecto trata de responder son:

- ¿En qué dominio del conocimiento estamos trabajando?
- ¿Quiénes son los usuarios objetivo?
- ¿Por qué es interesante el tema propuesto?

- ¿Cuáles son las preguntas que nuestro proyecto de PNL trata de responder?

Trabajos Relacionados

1. K-Nearest Neighbor (KNN): el algoritmo K-Nearest Neighbor es el algoritmo de clasificación basado en instancias. Este algoritmo calcula la distancia (por ejemplo, la distancia euclidiana) o la distancia de similitud (por ejemplo, la similitud del coseno) entre el entrenamiento de datos y la prueba de datos.
Los tweets de datos que han sido clasificados se transforman en un formato de grafico dirigido, con los usuarios de Twitter como nodos y las menciones en los tweets como bordes, como se muestra en la Figura [1]. El borde tiene un peso que representa cuantos tweets env´ıa un usuario X, a otros usuarios Y o Z en un periodo de tiempo t. Además, contiene la lista de tweets clasificados del usuario X al usuario Y o Z.

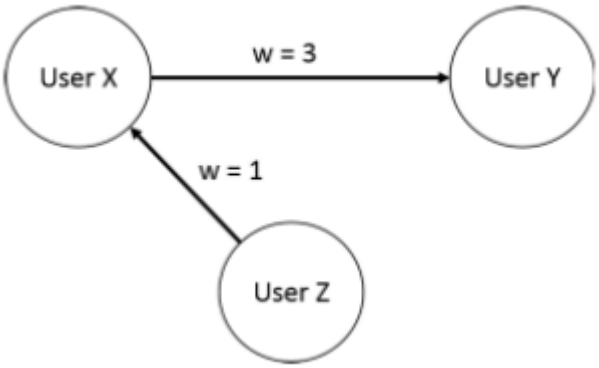

Figura 1. Visualización de datos gráficos de Twitter

2. Ocho reglas generales para la extracción de características usando Sarna

Estas reglas se pueden observar en la Figura [2] así como su definición a continuación:
a) El número de malas palabras en el tuit.
b) El número de palabras que muestran emociones negativas.
c) El número de palabras que muestran emoción positiva.
d) Combinación de un pronombre de primera persona, emoción negativa y combinación de un pronombre de segunda persona para capturar el ciberacoso.
e) Combinación de segundo pronombre con malas palabras para captar el ciberacoso.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

f) Combinación de pronombres en primera persona, palabras que expresan emociones negativas y pronombres de tercera persona o nombres propios para capturar el ciberacoso.

g) Combinación de un pronombre de tercera persona o nombre propio con una mala palabra para capturar el acoso cibernético.

h) La combinación de enlace, blasfemia y pronombres también se utiliza para capturar el acoso cibernético.

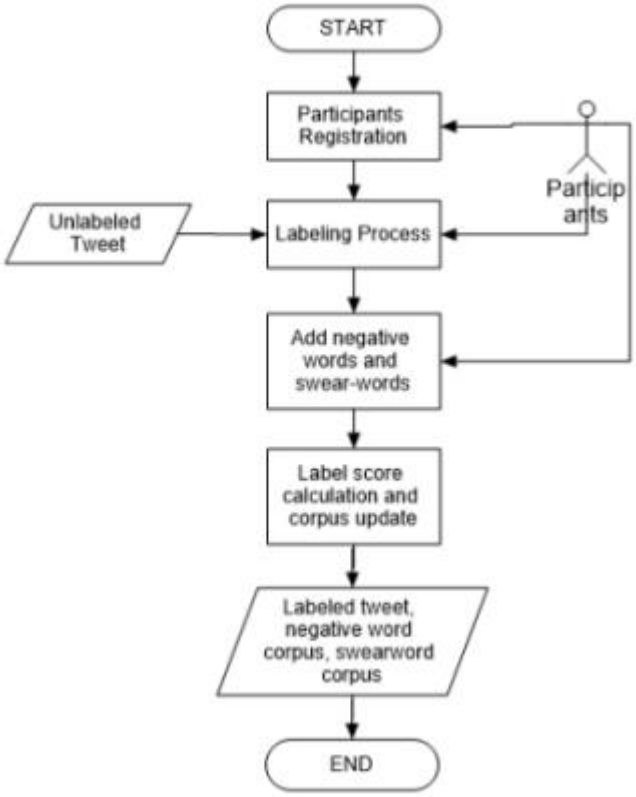

Figura 2. Sistema de etiquetado de datos

3. spaCy: Cada vez es más reconocido por procesar y examinar datos en PNL. Los datos textuales no estructurados se generan a gran escala y es fundamental para procesar y obtener información de datos no estructurados.

4. Modelo BERT: Actualmente es muy conocido puesto que Google está utilizando este modelo de lenguaje de procesamiento natural para mejorar sus búsquedas; En su lugar, lo hemos aplicado como un clasificador de texto para realizar el análisis de sentimiento.

5. Tokenización BERT: El comentario o Tweets seleccionados los empieza a tokenizar y nosotros lo hemos usado para obtener el conjunto de datos que contiene cada Tweet.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

Propuesta

Para este trabajo, estamos utilizando un conjunto de datos de tweets relacionados con el ciberacoso, que habría que analizarlos en profundidad mediante la técnica de Sarna para poder clasificarlos de forma booleano; es decir, los que contengan 0 serán tuits de ciberacoso o de los que se pueda sospechar contienen cyberbullying, los que contengan 1 serán los tweets normales por lo que debemos realizar un sistema de etiquetado que podría utilizar diferentes personas que quisieran participar para etiquetar cada pieza de información, en cada uno de nuestros tuits en particular. Las personas pueden elegir una de las cuatro opciones. Opciones ofrecidas para un tweet. El propósito de usar cuatro opciones es permitir que las personas elijan una etiqueta para clasificar el tweet según su confiabilidad en un rango de 1 a 4 como se puede ver en la figura [3]. Esto se debe al nivel de confianza a la hora de elegir si el Tweet es ciberacoso o no.

Options	Weights	Descriptions
1	−2	Very non-cyberbullying
2	−1	Non-Bullying
3	1	Bullying
4	2	Very Cyberbullying

Figura 3. Opciones de etiquetado

1. Etiquetado: Las opciones 1 y 2 se utilizan para calcular la puntuación total sin ciberacoso. Para su parte, se utilizan las opciones 3 y 4 se utilizan para calcular la puntuación total de ciberacoso. Para diferenciar el grupo que acosa cibernéticamente y grupo que no acosa cibernéticamente, hay un signo menos en el peso del grupo que no acosa cibernéticamente.

Finalmente, se obtuvieron los estadios de detección del ciberacoso. Nuestro principal objetivo es detectar cualquier situación de ciberacoso en Twitter; por lo que su funcionamiento parte de la adecuada formación de un algoritmo de aprendizaje supervisado. De ahí la importancia de proporcionar una base de conocimientos a partir de un análisis semántico y determinación de sentimientos para alcanzar la máxima tasa de precisión. Así, para llevar a cabo estas tres etapas ha sido necesario: (i) obtener la base de conocimientos, (ii) capacitación de modelos de aprendizaje supervisado, y (iii) implementación de estudios de casos.

1. Obtención de la base de conocimientos: Se debe establecer una base de conocimientos adecuada, comúnmente conocido como corpus, que interviene en la detección presuntiva del ciberacoso en lengua española como se puede ver

en la figura [4]. Un corpus es un conjunto de palabras o frases que tienen previamente clasificados según diferentes intereses a través de etiquetas.

Pejorative word or insult	Synonym Ecuador
Animal, bestia	Huev*n
Mujer interesada	Grilla
Sinvergüenza	Caretuco
Imbécil	Careverg*
Bastardo	Hijo de put*
Temeroso	Ahuev*ado
Enojado	Arrech*
Fea/feo	Bagre
Feo/bajo status social	Batracio

Figura 4. Diez ejemplos de palabras insultantes y sus sinónimos en Ecuador.

2. Entrenamiento de modelos de aprendizaje supervisado: Se utilizarán tres métricas específicas de precisión.

a) Puntuación de precisión, calculará la precisión del modelo de aprendizaje contando el número de muestras de ocurrencias que coinciden con el conjunto predeterminado de valores.

b) Puntuación de precisión promedio, resumida en una curva de recuperación de precisión como la media de las precisiones alcanzadas en cada umbral.

c) puntuación F1, se puede interpretar como un promedio ponderado de precisión y recuperación, donde una puntuación F1 alcanza su mejor valor en 1 y su peor puntuación en 0.

3. Implementación de estudios de casos: Existen tres tipos de análisis para la detección presuntiva de acoso cibernético.

a) Análisis de oraciones y/o palabras, reflejará si el contenido de una oración o texto presumiblemente representa algún tipo de acoso.

b) Análisis del perfil de usuario, presumiblemente determinando un porcentaje de acoso en base a un número predefinido de tweets históricos de un usuario específico. Para ello, introduzca el perfil en Formato de Twitter: @profildeuser.

Resultados y discusión

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

Análisis del modelo

El modelo utilizado fue BERT el cual nos ayudará a determinar si un Tweet es clasificado como cyberbullying (comentario agresivo, comentario insultante o comentario tóxico) o neutral (comentario que no es tóxico). La clasificación se desarrolla a partir de la propuesta realizada, ya que actualmente estamos clasificando los comentarios o Tweets en opciones de 1 (que es un comentario que contiene cyberbullying) y 0 (que es es un comentario neutral), esto se puede ver en las Figuras [5] y [6] para que se vean las gráficas de clasificación.

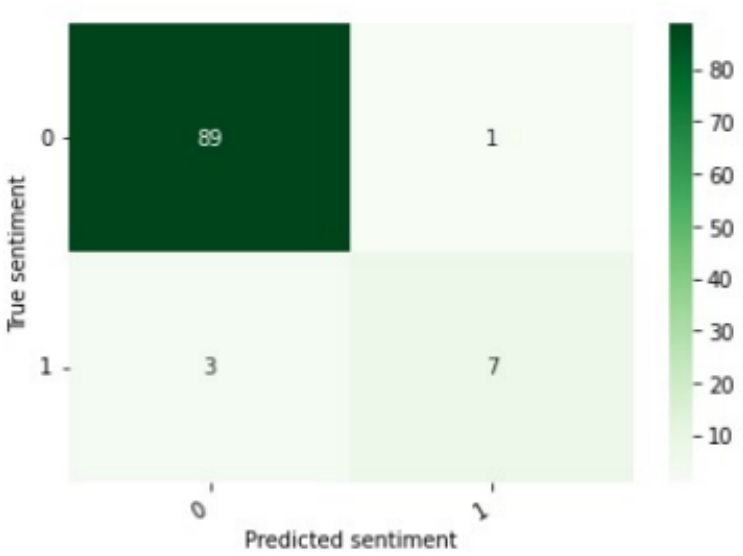

Figura 5. Clasificación de Tweets en opciones de 1 y 0.

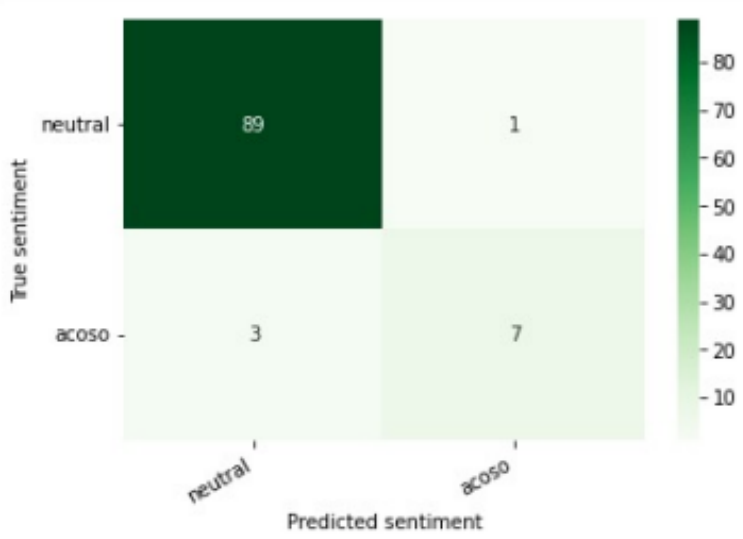

Figura 6. Clasificando los Tweets de acoso y los Tweets neutrales.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

Análisis de clasificación

Sin embargo, para que BERT pudiera clasificar los Tweets, tuvo que pasar por un proceso de entrenamiento, pero ¿por qué entrenarlo? Es necesario hacerlo para que tenga un procesamiento más fluido, lo que hará de él un modelo más eficaz y eficiente por lo que los sentimientos que tiene que predecir serán más fácil de identificar, este entrenamiento se puede ver en la Figura [7], tuvieron una duración de 20 aproximadamente 40 minutos haciendo que la computadora esté activa durante al menos 2 horas para que pueda lograr el objetivo básico.

```
Epoch 1 de 5
------------------
/usr/local/lib/python3.7/dist-packages/transformers/tokenization_utils_base.py:2307: FutureWarning: The `pad_to_max_length`
  FutureWarning,
Entrenamiento: Loss: 0.3049762084893882, accuracy: 0.89875
Validación: Loss: 0.16842625822339738, accuracy: 0.95

Epoch 2 de 5
------------------
Entrenamiento: Loss: 0.2376163786649704, accuracy: 0.925
Validación: Loss: 0.29977081935586675, accuracy: 0.95

Epoch 3 de 5
------------------
Entrenamiento: Loss: 0.15388401919975878, accuracy: 0.95625
Validación: Loss: 0.2482876716447728, accuracy: 0.96

Epoch 4 de 5
------------------
Entrenamiento: Loss: 0.07530190275982022, accuracy: 0.975
Validación: Loss: 0.24376138745407974, accuracy: 0.95

Epoch 5 de 5
------------------
Entrenamiento: Loss: 0.04019912391435355, accuracy: 0.98875
Validación: Loss: 0.2275712515693158, accuracy: 0.96
```

Figura 7. Entrenamiento del Modelo Bert para clasificar los Tweets.

Resultados de entrenamiento

Después de entrenar al modelo, nos muestra los resultados de cómo realizó la clasificación tan pronto como con precisión n; es decir, qué tuits son realmente de acoso (etiqueta 1) y cuáles son tuits neutros (etiqueta 0), esto se puede observar en las Figuras [8] y [9] para dar una vista detallada de los resultados obtenidos durante el proceso y destacar que cada entrenamiento contiene a una parte de nuestro conjunto de datos inicial.

Revista Innovación y Software
Vol. 3, No. 2, Mes Septiembre-Febrero, 2022
ISSN: 2708-0935
Pág. 81-90
https://revistas.ulasalle.edu.pe/innosoft

	precision	recall	f1-score	support
0	0.97	0.99	0.98	90
1	0.88	0.70	0.78	10
accuracy			0.96	100
macro avg	0.92	0.84	0.88	100
weighted avg	0.96	0.96	0.96	100

Figura 8. Resultados del entrenamiento del Modelo BERT con etiquetas de 1 y 0.

	precision	recall	f1-score	support
neutral	0.97	0.99	0.98	90
acoso	0.88	0.70	0.78	10
accuracy			0.96	100
macro avg	0.92	0.84	0.88	100
weighted avg	0.96	0.96	0.96	100

Figura 9. Resultados del entrenamiento del Modelo BERT con etiquetas de acoso y neutral.

Conclusiones

Para finalizar, es importante mencionar que el ciberacoso es un problema actual que debemos solucionar inmediatamente, ya que muchas personas terminan perjudicadas por estos comentarios negativos en sus vidas, así que para llegar a su solución aplicaremos el procesamiento de lenguaje natural, el cual es una disciplina importante en nuestra actualidad porque gracias a su comprensión es posible identificar y analizar diferentes textos. Gracias a ello se puede aplicar en diferentes áreas para mayor beneficio, en este caso, el área que nos ocupa es ciberacoso en la aplicación de Twitter. Como parte fundamental, este clasificador de texto nos ayudará a cumplir con el objetivo principal que es frenar el ciberacoso, implementando la aplicación del modelo BERT para lograr un resultado favorable y así poder frenar este problema denominado “ciberacoso”.

Referencias

[1] Gabriel A. Leon-Paredes, Wilson F. Palomeque-Leon, 2019. *Presumptive Detection of Cyberbullying on Twitter through Natural Language Processing and Machine Learning in the Spanish Language*. https://ieeexplore.ieee.org/abstract/document/8987684/

[2] Hani N., Dade N. *Indonesian Twitter Cyberbullying Detection using Text Classification and User Credibility*, 2018. https://ieeexplore.ieee.org/abstract/document/8350758/

[3] Monirah A. Al-Ajlan, Mourad Y. *Optimized Twitter Cyberbullying Detection based on Deep Learning*, 2018. https://ieeexplore.ieee.org/abstract/document/8593146

[4] Hoy interessa. *Cyberbullying in Peru stand at rates of up to forty per cent* https://gestion.pe/tendencias/ciberbullying-peru-situa-tasas-40-140489-noticia/

[5] Orhan G. Yalcin. *Sentiment Analysis in 10 Minutes with BERT and TensorFlow* https://towardsdatascience.com/sentiment-analysis-in-10-minutes-with-bert-and-hugging-face-294e8a04b6

[6] TensorFlow. *Classify text with BERT* https://www.tensorflow.org/text/tutorials/classify_text_with_bert

[7] Kaggle. Classified Tweets https://www.kaggle.com/datasets/munkialbright/classified-tweets

[8] GitHub. Cyberbullying Detection in Tweets https://github.com/apeksha104/ Cyberbullying-Detection-in-Tweets

[9] V Krithika, V Priya. A Detailed Survey On Cyberbullying in Social Networks https://ieeexplore.ieee.org/abstract/document/9077794

[10] Saima Sadiq, Arif Mehmood, Saleem Ullah, Maqsood Ahmad, Gyu S. Choi, Byung-Won On Aggression detection through deep neural model on Twitter https://www.sciencedirect.com/science/ article/abs/pii/S0167739X19330717

[11] Bandeh A. Talpur, Declan Oaˆ€™Sullivan Multi-Class Imbalance in Text Classification: A Feature Engineering Approach to Detect Cyberbullying in Twitter https://www.mdpi.com/2227-9709/7/4/52

Roles de Autoría

Jonathan Matwey Aguirre Soto: Conceptualización, Curación de datos, Análisis formal, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Hector Ávila Gonzales:** Curación de datos, Investigación, Metodología, Software, Validación, Redacción - borrador original. **Valeria Bravo Saines:** Conceptualización, Análisis formal, Investigación, Metodología, Software, Redacción - borrador original.