Detección de comentarios tóxicos de contexto futbolístico en Twitter

Edgar Piña Cuentas [1], Fernando Andrés Gómez González [2]

[1] Universidad Tecnológica de Bolívar, Cartagena, Colombia, epina@utb.edu.co
[2] Universidad Tecnológica de Bolívar, Cartagena, Colombia, fegomez@utb.edu.co

Abstract: This project is focused on the processing and prediction of natural language, where the data set was extracted using the Twitter tool. The focus of our data is based on user comments about football and the goal is to model an ML model that determines if a tweet from a recurring person on the social network is generating hate.

Resumen: El presente proyecto está enfocado al procesamiento y predicción del lenguaje natural, donde el conjunto de datos fue extraído usando la herramienta de Twitter. El enfoque de nuestros datos se basa los comentarios de los usuarios sobre fútbol y el objetivo es modelar un modelo de ML que determina si un tweet de una persona recurrente en la red social es generador de odio.

1. Introducción

La red social de Twitter es muy usada para difundir noticias, opiniones e informaciones. Sin embargo, por la masividad y la cantidad de usuarios que hoy en día tiene se presta como medio de difusión de odio.

Este proyecto toma como tópico el fútbol, debido a la cantidad de opiniones que los usuarios generan cada día, entre estas críticas, comentarios y respuestas a distintos eventos que pueden suceder en el mundo de esta disciplina. Dicho deporte puede ser muy impredecible y lo que provocará el comentario de los usuarios que estén interesados, ya sea como aficionados o detractores. Se pretende ver si es posible crear un algoritmo capaz de identificar un comentario que pueda considerar toxico entre los millones de tweets que comparten las personas en la red social día a día.

1.1 Contexto del problema.

El fútbol como uno de los deportes más grandes y conocidos a nivel profesional, es uno también de los más criticados. Como es posible, debido a los resultados de un equipo y obviamente de la calidad de los jugadores, los aficionados al club tienden a apoyar o a censurar el rendimiento en el campo de juego. Todas estas respuestas a los estímulos presenciados por los escritores son fuente vital de información que permite la medición y clasificación de los sentimientos de los fanáticos, estimando así las intenciones con las que van dirigidas las misivas.

1.2 Descripción de la problemática.

La herramienta de Twitter es un medio de comunicación masivo que permite la interacción entre usuarios de manera abierta, donde cualquier persona puede expresar su libre opinión sin cautela. Los

discursos de odio suelen tener mayor presencia en los comentarios de publicaciones de equipos que no tengan un rendimiento gratificante, por lo tanto, las opiniones de detractores serán más numerosas. Se busca analizar si es posible identificar los comentarios de los usuarios, enfocados en contextos futbolísticos, que expresan textualmente quejas, críticas, insultos y/o juicios negativos.

1.3 Pregunta de investigación

¿Es posible identificar la toxicidad en las opiniones de los usuarios de Twitter en español sobre el futbol?

1.4 Hipótesis

Las opiniones toxicas en torno al futbol de los usuarios de Twitter en español suelen tener elementos característicos que hacen posible su identificación por medio de un modelo de Machine Learning

2. Trabajos relacionados

Como base teórica, está el documento de Análisis de Sentimiento en Tweets de Fútbol Argentino de Mario Ferreyra [1]. En este, se presenta la temática de minado de opiniones referentes al Fútbol Argentino con la finalidad de establecer la intencionalidad de las frases extraídas de la herramienta de Twitter, donde cada tweet pasó por un proceso de filtrado individual establecido por el grupo de trabajo del autor.

Mario Ferreyra implementó el uso de sistemas de predicción (pasando por limpieza y normalización de los textos) y el tipo de representación de texto “Bag of Words” para la experimentación y obtención de resultados. Como métricas de medición de precisión utilizó Matriz de Confusión y F1-Score.

Como conclusión del proyecto, tuvieron como resultado que los sistemas que desarrollaron clasifican de manera efectiva los tweets con polaridad Positiva y Negativa, sin embargo, los que presentan etiquetas Neutras y Ninguno no demuestran buen desempeño por el hecho de tener parecido a los anteriormente mencionados.

Como segundo fundamento teórico, tenemos el documento de Seguimiento De Emociones En Partidos de Fútbol que tiene como autor a Martí Rella Muñoz de 2015[2]. En este proyecto se enfoca en la detección de emociones expresadas como tweets cuando transcurren partidos de fútbol mediante procesamiento del lenguaje natural y sistemas de predicción. El autor clasificó los mensajes extraídos de Twitter para clasificarlos y etiquetarlos de dos formas, la primera sería por polaridad abarcando cuatro categorías como Positivo, Negativo, Neutro y Ninguno; por otro lado, se clasificaría por emoción o sentimiento abarcando siete categorías como Alegría, Enojo, Miedo, Repulsión, Sorpresa, Tristeza o Ninguno.

Como penúltimo apoyo, se presenta ELiRF-UPV en TASS 2015: Análisis De Sentimientos En Twitter por Lluís Hurtado, Ferran Pla y Davide Buscaldi [3]. Donde proponen una actualización de la 3ra edición del TASS, dándose la tarea de medir la intencionalidad de los mensajes en seis etiquetas y cuatro etiquetas.

Para el análisis de los datos tuvieron que hacer varios procesos de clasificación donde evaluaban diferencias entre P y P+, y N y N+ (Varían la intensidad de positividad y negatividad). Además, usaron varios corpus adicionales con el fin de abarcar mayor área en el entrenamiento.

Por último, se encuentra Análisis De Sentimiento En Twitter De Los Socios De Un Club De Futbol A Través De La Evaluación De Herramienta Que Manejan Gran Volumen De Información presentado por Ricardo Viteri Alvarado [4]. Tenía como propuesta el diseño de un modelo que analice los sentimientos de los tweets de alentadores de los clubes de fútbol en Ecuador y apuntando al propósito social de estudiar el comportamiento y puntos de mejora para alcanzar nuevos socios y mejorar su economía.

Este proyecto les dio como resultado un modelo que les permitió analizar la aceptación general de la hinchada, permitiéndoles así establecer los requerimientos mínimos necesarios para llegar a nuevos socios

3. Recursos

a. Conjunto de datos

El conjunto de datos a utilizar serán una serie de 400 tweets extraídos de la plataforma empaquetados en un archivo JSON y clasificados manualmente para su análisis.

3.1.1 Origen

Los datos serán extraídos por medio de una API construida para obtener una lista de tweets de la plataforma; esta utiliza la librería Python-Twitter para la extracción de los datos con las credenciales concedidas por la plataforma Twitter Developer.

3.1.2 Descripción de los datos

Después de realizar la extracción se seleccionaron las características consideras más relevantes para esta investigación. Se determinaron las siguientes características:

- id: Dato de tipo entero que determina un identificador único para un tweet.
- lang: Dato tipo string que determina el lenguaje en el que se escribió el tweet.
- text: Dato tipo string que contiene el tweet.
- user_mentions: Dato tipo entero que contiene el número de menciones en el tweet.

3.1.3 Descripción de la clase a predecir

Se definió la clase a predecir como “toxicidad”. Esta será una variable categórica que dispondrá de 2 categorías, definidas de la siguiente forma:

- Comentario “normal”: El comentario no genera toxicidad.
- Comentario “toxico”: El comentario genera toxicidad.

Para este caso se considerarán como comentarios tóxicos todos aquellos que expresen ira, quejas, criticas destructivas, insultos y/o juicios negativos dirigidos hacia clubs, selecciones nacionales, jugadores o entidades relacionadas con el deporte.

La categorización fue hecha en base a la “Guía de anotación para corpus de sentimientos en español” publicado por Edwin Alexander Puertas del Castillo.

3.1.4 Numero de instancias del conjunto de datos y pruebas

El número total de datos recolectados en el presente proyecto será de 400, extraídos mediante la API de Twitter. La distribución de instancias que se tendrá en cuenta será de un 80% del total de conjunto de datos destinados a los datos de entramiento y un 20% del total del conjunto de datos destinados para los datos de prueba.

b. Librerías Externas

En esta ocasión se utilizará una librería externa creada por el Doctor en Human Language Technology (Tecnología del lenguaje humano) Edwin Alexander Puertas del Castillo llamada “text_processing”, que ofrece una serie de herramientas que facilitan el tratamiento de los datos a nivel de limpieza, tokenizacion, entre otras funciones.

4. Descripción del sistema

El flujo del sistema será el siguiente:

- Inicialmente se seleccionará un modelo de ML con el cual se trabajará.
- Simultáneamente se extraerá un corpus inicial de la plataforma Twitter, de este corpus se seleccionarán las características necesarias y el resto se descartarán, además se clasificará cada tweet y se le agregará una característica adicional, la “polaridad” que será la variable objetivo.
- El dataset obtenido en la etapa anterior será divido en 2 data sets, uno de prueba y otro de entrenamiento, ambos balanceados, además a estos se les aplicará un tratamiento de datos.
- Se tomará el data set de entrenamiento obtenido en la etapa anterior y se aplicaran distintas técnicas de extracción de características para obtener las mejores.
- Se procede a entrenar el modelo.
- Una vez se entrene el modelo se utilizan los datos de prueba para calcular las distintas métricas y poder evaluar el modelo.
- Si al evaluar el modelo este cumple unos estándares mínimos, se procede a terminar el proceso de entrenamiento y el modelo es elegible para entrar en producción. Por otro lado si

este no cumple los requisitos mínimos es necesario reevaluarlo al momento de extraer las características, de no ser suficiente se retornara a la extracción de datos y se validara nuevamente el corpus, si con todo esto el modelo sigue resultando insuficiente se determinara que el modelo no es apropiado para este caso y se seleccionara otro modelo.

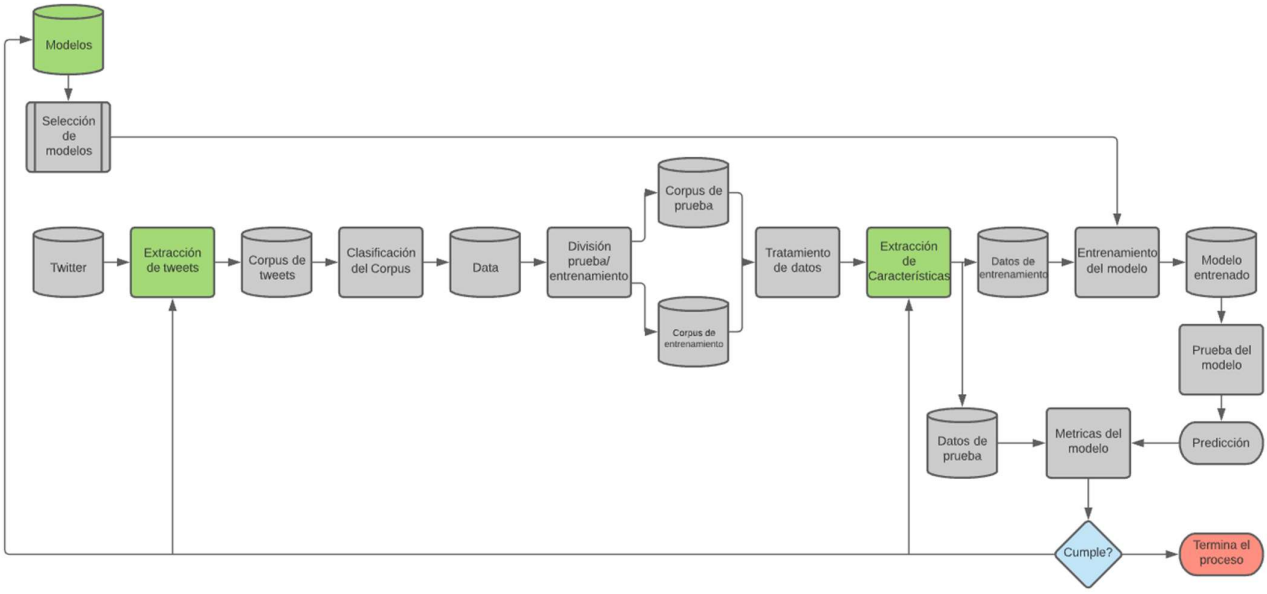

Ilustración 1 – Flujo del Sistema

5. Experimentos

5.1 Iteración Inicial

Inicialmente se definió el código base donde se para la selección de características se utilizan TF-IDF y Bag of Words, ambas con una configuración básica como se muestra a continuación:

```
FeatureUnion(transformer_list=[('tfidf_vector',TfidfVectorizer(max_df=0.9,min_df=10),
                               ('bow', CountVectorizer(max_df=0.9, min_df=10))])
```

Además, se decidió utilizar los modelos de Regresión Logistica (LR) y Super Vector Machine (SMV), con la siguiente configuración:

```
model_sklearn = [LogisticRegression(solver='liblinear'),
                 svm.SVC(kernel='linear', C=1, probability=True)]
```

Estas configuraciones ofrecieron los siguientes resultados iniciales:

```
LR
Desempeño básico: 0.65
SVM
Desempeño básico: 0.575
```

5.2 Primera Iteración

Para la primera iteración se decidió mejorar la calidad de las características, para esto se modificó el parámetro de min_df de 10 a 30 y aunque el modelo de LR bajo un poco en score el de SVM subió.

```
LR
Desempeño básico: 0.65
SVM
Desempeño básico: 0.6375
```

5.3 *Segunda Iteración*

En la configuración de TF-IDF se pasó de unigramas a unigramas y trigramas y genero una mejora para los 2 modelos:

```
LR
Desempeño básico: 0.6875
SVM
Desempeño básico: 0.7
```

5.4 Final

Después de la mejora lograda en la iteración anterior no se encontraron alternativas mejores, el resto de las iteraciones intentadas solo conllevaron a peores desempeños.

6. Análisis de resultados

6.1 Regresión Logística

El modelo de regresión logística generó un accuracy del 68.75%, un desempeño general considerado mediocre. Si nos remitimos a la matriz de confusión y al reporte de clasificación observamos que el algoritmo tiene más probabilidad de predecir correctamente si un tweet es toxico que si no lo es.

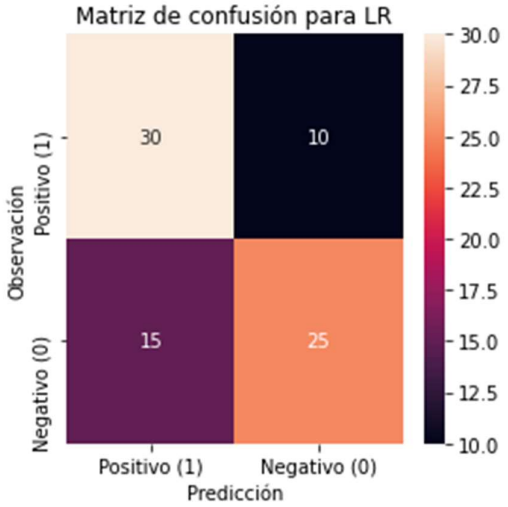

Ilustración 2-Matriz de confusión modelo de Regresión Logística

Tabla 1 – Reporte de Clasificación Regresión Logística

LR
Accuracy: 0.6875
F1-Score: 0.6862745098039216

	precision	recall	f1-score	support
0	0.67	0.75	0.71	40
1	0.71	0.62	0.67	40
accuracy			0.69	80
macro avg	0.69	0.69	0.69	80
weighted avg	0.69	0.69	0.69	80

6.2 Super Vector Machine

El modelo de SVM genero un accuracy del 70%, un resultado mediocre. Se observa que igual que LR este algoritmo identifica mejor los casos en los que un tweet es toxico, teniendo una precisión mayor para esta clase, esto lo comprobamos tanto en su matriz de confusión, como en su reporte de clasificación.

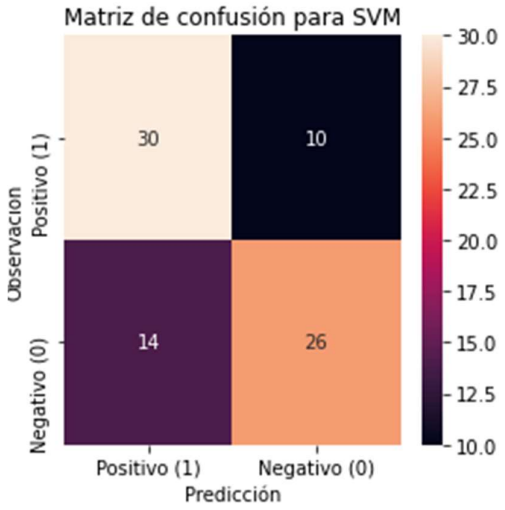

Ilustración 3-Matriz de confusión de SVM

Tabla 2 – Reporte de Clasificación SVM

SVM
Accuracy: 0.7
F1-Score: 0.6992481203007519

	precision	recall	f1-score	support
0	0.68	0.75	0.71	40
1	0.72	0.65	0.68	40
accuracy			0.70	80
macro avg	0.70	0.70	0.70	80
weighted avg	0.70	0.70	0.70	80

7. Conclusión

Para terminar, se puede concluir que en el proceso de entrenamiento y prueba ambos sistemas de predicción han presentado resultados de rendimiento mediocres, esto puede haberse dado por la cantidad de datos totales, donde si se hubiese hecho un proceso de extracción más amplio la calidad de los modelos presentados podría haber sido más alta. Por otra parte, la diferencia entre ambos modelos, aun siendo pequeña, SVM tuvo un mejor resultado general para el caso de aplicación.

Respecto a la hipótesis planteada, podemos decir que aun con un conjunto de datos relativamente pequeño ha sido posible identificar elementos característicos que establezcan si un comentario refleja

negativismo o toxicidad, por tanto, la hipótesis es correcta y es posible que si se utiliza un data set más grande el sistema propuesto podrá reconocer más características y por ende, los modelos presentarían mejores resultados.

8. Repositorio del proyecto

Repositorio de Github:
https://github.com/edgarpc6/Detecci-n-de-comentarios-t-xicos-de-contexto-futbol-stico-en-Twitter

9. Referencias

13.1 Websites

[1] Ferreyra, M. (s, f.). Análisis de Sentimiento en Tweets de Fútbol Argentino. Universidad Nacional de Córdoba.

[2] Rella, M. (2015). Seguimiento de emociones en partidos de fútbol. Universitat Pomeu Fabra Barcelona.

[3] Buscaldi, D. Hurtado, L & Pla, F. (2015). ELiRF-UPV en TASS 2015: Análisis de Sentimientos en Twitter. In TASS@ SEPLN (pp. 75-79).

[4] Viteri, R. (2016). Análisis de Sentimiento en Twitter de los Socios de un Club de Fútbol a través de la Evaluación de Herramienta que Manejan Gran Volumen de Información. Universidad de Guayaquil.