Detección y análisis de toxicidad en Twitter

Lucía Inés Merlo y Nora Hafidi López
Universidad Politécnica de Valencia.

May 26, 2021

1 Objetivo

El objetivo de este informe consiste en la obtención de un modelo de Machine Learning que permita evaluar la polaridad de un sentimiento contenido en un tweet. Es decir, poder extraer a través de publicaciones acerca de un tema, producto, persona, o servicio el sentimiento contenido en el mismo.

Este análisis forma parte del campo de investigación llamado 'Procesamiento del Lenguaje Natural', y puede efectuarse en cualquier área dentro de la web. Su popularidad ha ido en aumento, puesto que es posible tanto para empresas como gobiernos, conocer opiniones e intenciones de los usuarios que hacen alusión a determinado tema de interés.

Dicho esto, en este trabajo nos centraremos en obtener un modelo de ML que cumpla una tarea de clasificación. De esta forma, será posible dado un nuevo tweet, catalogarlo apropiadamente según su contenido.

2 Introducción

Actualmente, internet ha tomado un papel primordial en la vida de las personas. Trabajo, estudio, comunicación, compras, o trámites; prácticamente todo puede realizarse si se cuenta con conexión a la red. Tanto es así, que muchas de estas personas, consideran internet como una extensión más que un mero complemento de su vida.

En el área de la comunicación, es de amplio conocimiento que las redes sociales son, de forma notable, las herramientas más utilizadas a la hora de intercambiar información, consumir contenido, o simplemente, hablar con otros usuarios.

Twitter es una de las principales redes sociales dentro del mundo virtual prácticamente desde su origen. Con sus característicos 140 caracteres, la publicación de un usuario se conoce como "tweet".

A partir de estos, se genera una gran cantidad de datos vinculados a cualquier acontecimiento a tiempo real, siendo un canal particularmente efectivo para que sus 322 millones de usuarios alrededor del mundo [2] y, concretamente, en España 4.1 millones de ellos [3] interactúen en directo.

Dado el amplio abanico de posibilidades dentro de la misma, no es extraño considerar la vasta cantidad de información disponible dentro de cada uno de estos intercambios, ya sea: contenido multimedia, artículos educativos, marketing, mensajes de usuarios a otros, e incluso dentro de los anteriormente mencionados, mensajes con connotaciones negativas.

Dicho esto, resulta interesante indagar acerca del comportamiento de los usuarios en esta red social tan popular. Concretamente, poder analizar y clasificar aquellos tweets en función de ciertas palabras, con el fin de determinar si son tóxicos o no, y si se cumple el primer caso, poder definir qué tanto lo son.

3 Metodología

3.1 Datos

El análisis se basa en datos contenidos en el dataset "DETOXIS" [1], los cuales han sido proporcionados por la plataforma de un concurso.

El dataset presenta aproximadamente 4357 tweets en español, recogidos en diferentes plataformas webs. Concretamente, han sido posteados en el periodo de Agosto del 2017 hasta Julio del 2020, donde el tema central de ellos es fundamentalmente la inmigración. Contiene comentarios que se encuentran clasificados según dos criterios. El primero y más genérico, consiste en etiquetarlo según si es tóxico o no: 'no tóxico' = 0 y 'tóxico' = 1. Posteriormente, se le asigna a cada comentario diferentes niveles de toxicidad: 'no tóxico' = 0, 'medianamente tóxico' = 1, 'tóxico' = 2, 'muy tóxico' = 3. A pesar de contener más atributos, estas variables serán las únicas utilizadas en este análisis.

El corpus de entrenamiento del dataset conforma el 80% del dataset y el 20% de los datos restantes, el etiquetado está oculto. Será el corpus de test de los organizadores del concurso. Para el entrenamiento adecuado de los modelos, se ha dividido el corpus de entrenamiento en 80/20 para entrenar y evaluar los modelos.

3.2 Preprocesamiento

Para la adecuada normalización del texto, se han efectuado los siguientes pasos:

- Eliminación de signos de puntuación y caracteres no fuesen alfanuméricos.
- Supresión de stopwords.

- Obtención de la raíz de las palabras.
- Obtención del lema de las palabras.
- Extracción de características.

Eliminación de signos de puntuación y ruido

Para eliminar los signos de puntuación, se ha utilizado una función de la librería string. Esta nos incluye en una lista todos los signos de puntuación existentes y basta con recorrer los tokens e ir eliminando los caracteres presentes en la lista. Para la eliminación de los caracteres no alfanuméricos, se ha utilizado la función isalpha().

Stopwords

Con nltk se han eliminado todos las stopwords presentes en el lenguaje español.

Raíz de las palabras

Con la función PorterStemmer(), se dispone a sustituir cada palabra por su raíz. Por lo tanto, palabras con la misma raíz, tendrán un significado similar.

Lema

Por último, se procede a lematizar el texto con lemma.lower()

Extracción de características

Para esto, se utilizan bolsas de palabras.

3.3 Modelos de Aprendizaje Automático

Para la tarea de aprendizaje, se han aplicado varios modelos de aprendizaje automático, con el fin de detectar y seleccionar aquel método que explica más adecuadamente los datos. Cabe destacar que se ha tomado en cuenta tanto modelos interpretables como no interpretables Estos últimos, mejor conocidos como "modelos de caja negra". Los mismos son:

Modelos interpretables	Modelos no interpretables
Árboles de decisión	Redes neuronales
Regresión Logística	SVM
K-vecinos	XGBoost
	Random Forest

Además de los modelos comentados previamente, otras técnicas fueron aplicadas. Específicamente, meta-algoritmos, tales como:

- AdaBoost, con un modelo base de Regresión Logística.

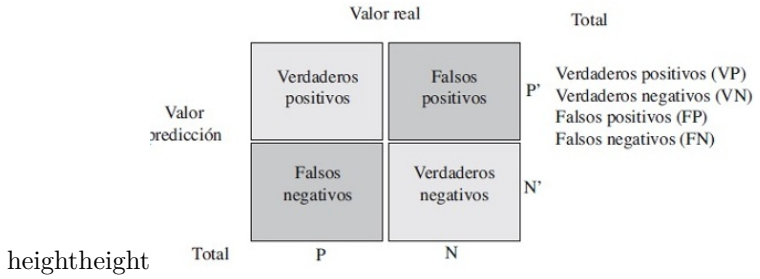

Figure 1: Matriz de confusión

- Bagging, con un modelo base de Regresión Logística.
- Stacking, con Redes Neuronales y Máquinas de Vectores Soporte como modelos base, y Árboles de decisión como segundo modelo base.

3.3.1 Métricas de evaluación

A partir de datos reales y predichos por un modelo de ML, es adecuado trazar una matriz de confusión con el fin de evaluar el desempeño del modelo aplicado.

En problemas de clasificación, el F1-score es una métrica que evalúa de alguna forma la exactitud de las predicciones. Específicamente, se calcula a partir de las medidas de precision y recall como su media armónica.

Precision se refiere al número de verdaderos positivos (VP) dividido por el número total de predichos positivos, hayan sido correctamente predichos o no. Por otro lado, el *recall* se calcula a partir del número de VP dividido por el número de todas las muestras que tendrían que haber sido predichas en la clase positiva, es decir, los falsos negativos, tal y como se observa en la figura.

$$F1 = 2 * \frac{precision * recall}{precision + recall}$$

El F1-score puede tomar valores entre [0-1]. Donde un valor igual a 1, indica una *precision* y *recall* perfectos. Por contraste, un valor igual a 0 indica que tanto la *precision* como el *recall* son 0. Es decir, puesto que en esta tarea resulta interesante evaluar tanto la precisión como el recall entre varias soluciones posibles, la métrica óptima será la comentada.

Esta métrica asume que ambos componentes son igual de relevantes, aunque evidentemente el F1-score será adecuado o no según el problema de interés.

4 Detección de toxicidad

La primera subtarea consiste en detectar la existencia o no de toxicidad en los mensajes contenidos en el dataset de entrenamiento. Cada comentario contiene una etiqueta que toma un valor de 0 o 1, donde 0 denota la ausencia de toxicidad y 1 la presencia de este.

4.1 Modelos de Aprendizaje Automático

Distintos algoritmos de aprendizaje automáticos han sido entrenados con el fin de predecir y evaluar la clasificación binaria de los datos de entrada, donde la partición de entrenamiento conforma el 80% del corpus de entrenamiento y la partición de test, un 20% del corpus de entrenamiento..

Tanto modelos interpretables, no interpretables como la combinación de los mismos fueron aplicados, estos modelos son:

Árboles de decisión, Regresión logística, SVM, Redes neuronales, K-vecinos, Bagging, AdaBoost, Random Forest, XGBoost y Stacking.

4.2 Métricas

Con el fin de cuantificar la calidad predictiva de los modelos seleccionados, se han obtenido valores de F1-score para cada tipo de modelo, tanto para la partición de entrenamiento como la de test, aplicando cross-validation con 5 particiones. Estos son los resultados:

Modelo	F1-score en entrenamiento	F1-score en test
Árboles de decisión	0.55	0.47
SVM	0.63	0.53
Regresión logística	0.67	0.62
Redes neuronales	0.63	0.59
K-vecinos	0.48	0.44
Bagging	0.65	0.60
AdaBoost	0.66	0.63
Random Forest	0.41	0.40
XGBoost	0.59	0.40
Stacking	0.59	0.55

Table 1: Métricas F1-score de los modelos utilizados .

4.3 Grid Search

El ajuste de hiperparámetros es realmente útil para extraer el máximo desempeño de los modelos aplicados dado el contexto de aplicación. La técnica de refinamiento que intenta hallar los valores óptimos de hiperparámetros de algoritmos de aprendizaje es conocida como Grid Search. Grid Search efectúa una búsqueda exhaustiva en un subconjunto de valores posibles de hiperparámetros especificado previamente, optimizando los resultados obtenidos en la tarea de predicción. Puesto que los resultados de la clasificación son ampliamente mejorables, se han ajustado los hiperparámetros de los modelos para obtener mejor calidad predictiva, aplicando cross-validation de 5 particiones. Estos son los resultados:

Modelo	F1-score en entrenamiento	F1-score en test
Árboles de decisión	0.56	0.47
SVM	0.62	0.58
Regresión logística	0.66	0.63
K-vecinos	0.45	0.43

Table 2: Métricas F1-score de los modelos utilizados.

4.4 Resultado

Grid Search ha mejorado notablemente el desempeño de los modelos utilizados, siendo notorio en los resultados. Este ha sido el caso del modelo de Regresión logística. El mismo, ha sido el algoritmo con mejor desempeño, presentando un valor de F1-score del 67% en el conjunto de entrenamiento y un valor del 62% en la partición de test.

Por otro lado, el algoritmo proveniente de la combinación de modelos que mejor resultado presenta es AdaBoost con 66% de F1-score en el entrenamiento, y 63% para el subconjunto de test.

5 Detección del nivel de toxicidad

La segunda subtarea consiste en detectar el nivel de toxicidad de los comentarios analizados en la primera subtarea. De forma similar que antes, cada comentario contiene una etiqueta que puede tomar valores de 0, 1, 2 o 3, denotando cada tweet de menos a más toxicidad en orden creciente.

5.1 Modelos de Aprendizaje Automático

Similarmente a la primera subtarea, distintos algoritmos de aprendizaje automáticos han sido entrenados con el fin de predecir y evaluar la clasificación multiclase de los datos de entrada, donde

la partición de entrenamiento conforma el 80% del corpus de entrenamiento y la partición de test un 20% del corpus de entrenamiento.

Tanto modelos interpretables, no interpretables, como una combinación de ellos fueron aplicados. Siendo estos: Árboles de decisión, Regresión logística, SVM, K-vecinos, Bagging, AdaBoost, Random Forest, XGBoost y Stacking.

5.2 Métricas

De forma similar a la primer tarea, para evaluar la calidad predictiva de los modelos, se han obtenido valores de F1-score para cada uno tanto en la partición de entrenamiento como en la de test. Para ello, se ha aplicado un cross-validation con 5 particiones. Los resultados son:

Modelo	F1-score en entrenamiento	F1-score en test
Árboles de decisión	0.23	0.24
SVM	0.27	0.22
Regresión logística	0.36	0.27
K-vecinos	0.24	0.22
Bagging	0.31	0.31
AdaBoost	0.28	0.23
Random Forest	0.20	0.19
XGBoost	0.30	0.27
Stacking	0.58	0.55

Table 3: Métricas F1-score de los modelos utilizados.

5.3 Grid Search

Al igual que en la subtarea anterior, en esta instancia igualmente se ha procedido a efectuar Grid Search con el fin de refinar hiperparámetros y mejorar los modelos de aprendizaje desarrollados. Y al igual que en los casos previos, aplicando cross-validation con 5 particiones, cuyos sus resultados se presentan en la tabla siguiente.

Modelo	F1-score en entrenamiento	F1-score en test
Árboles de decisión	0.27	0.17
SVM	0.27	0.23
Regresión logística	0.33	0.23
K-vecinos	0.24	0.21

Table 4: Métricas F1-score de los modelos utilizados.

5.4 Resultado

A pesar de haber utilizado Grid Search, en esta oportunidad no se han visto mejorados los resultados de los modelos en los que se aplicó.

Sin embargo, el algoritmo Stacking ha sobresalido del resto, con su resultado de f1-score en los datos de entrenamiento con un valor de 58%, mientras que para la parte de test consiguió llegar a un 55% de éxito.

6 Detección de errores

Resulta evidente que a pesar de efectuar una búsqueda exhaustiva de los mejores hiperparámetros de los modelos aplicados, la calidad de las predicciones aún contiene errores siendo mejorable.

La detección de errores consiste en detectar iterativamente los errores del aprendizaje para ir modificando aquellos aspectos en los que es posible. Es decir, se obtienen categorías de errores, donde se puede ir trabajando sobre ellos con el fin comentado anteriormente.

Sería interesante aplicarlo en un trabajo futuro, con el fin de refinar los algoritmos y sus consecuentes resultados.

7 Conclusiones

Con el objetivo de mostrar el poder y versatilidad que poseen los métodos de análisis morfosintáctico como los algoritmos de aprendizaje automático, se han efectuado tareas de clasificación de texto.

Asimismo, se ha demostrado que dependiendo del problema, un método puede ser más útil respecto a otro y viceversa. Por lo que en este tipo de tareas, es necesario extremar las precauciones en lo referido a establecimiento de hiperparámetros, validación, y en detectar adecuadamente y a tiempo la calidad de las predicciones.

En relación a los análisis desarrollados, las técnicas aplicadas han dado muy buenos resultados, ya que no obstante, y en nuestra opinión, en problemas del mundo real es extremadamente dificultoso alcanzar métricas sumamente elevadas.

De lo comentado previamente, se deduce la necesidad de seguir profundizando en este tipo de estudio, dedicando nuevas técnicas y enfoques con el fin de alcanzar en la medida de lo posible, el máximo provecho que los datos y de los métodos de análisis.

8 Aplicaciones

El volumen de datos creado en internet crece diariamente a pasos agigantados. En el caso de la red social de Twitter, esta situación es algo que también sucede.

Ya sea por política, hobbies, problemáticas sociales o cualquier otra razón variopinta, los usuarios de esta plataforma se valen del poder que le otorga una conexión web para emitir cualquier tipo de información u opinión.

Esto es algo que no sólo concierne a la comunidad twittera, si no que gobiernos, organismos e incluso muchas empresas, ven una fuente interminable de posibles beneficios a extraer. Tanto para detectar comportamientos que puedan afectar a la seguridad e integridad de cualquier persona, así como ver tendencias político-partidarias, las posibles aplicaciones son tan numerosas como personas hay en el mundo. Además, algunas utilidades no tan obvias, son explotadas por numerosas empresas. Detección de quejas de consumidores, segmentación de mercados o predicción de futuras ventas/usuarios son algunos ejemplos del provecho a obtener.

En resumen, el uso del procesamiento del lenguaje natural es extremadamente útil para cualquier tipo de contexto, en tanto en cuanto sea aplicado correctamente.

9 Concurso

Por último, se han enviado a un concurso organizado por [1], del cual como resultado para la primera subtarea se ha obtenido un F1-score de 45%, mientras que para la segunda tarea el valor de esta métrica ha sido de 62%.

Para obtener los resultados mencionados se ha utilizado el corpus de validación ya mencionado que conforma el 20% del total de los datos, y cuyas etiquetas únicamente eran conocidas por los organizadores.

Bibliografía

[1] *DETOXIS-IberLEF 2021*. 2021. URL: https://detoxisiberlef.wixsite.com/website.

[2] Rosa Fernández. *Número de usuarios mensuales activos de Twitter en el mundo 2018-2024*. 2021. URL: https://es.statista.com/estadisticas/636174/numero-de-usuarios-mensuales-activos-de-twitter-en-el-mundo/ (visited on 03/21/2021).

[3] Rosa Fernández. *Twitter en España - Datos estadísticos*. 2021. URL: https://es.statista.com/temas/3595/twitter-en-espana/ (visited on 03/30/2021).