Construcción de un sistema para la identificación de sexismo en las redes sociales (EXIST 2021)

Iria Durán Lusquiños y Daniel Goig Martinez

Lenguaje Natural y Recuperación de la Información
Grado en Ciencia de datos, Universitat Politécnica de Valencia

Abstract: En el presente informe se realiza una descripción de los sistemas de clasificación de textos desarrollados para la tarea " Sexism Identification in Social Networks (EXIST 2021)" organizada por IberLEF. Dicha tarea se centra en la identificación de mensajes sexistas procedentes de Twitter, red social que posee una política de censura y de Gab.com, medio donde no se filtran los textos. Se encuentran escritos tanto en inglés como en castellano. A partir de los vectores resultantes del preprocesamiento previo de los textos y de aplicar el método TF-IDF de extracción de características, entrenamos diferentes modelos de clasificación buscando maximizar las métricas de evaluación utilizadas en la competición: Accuracy y F1. Conseguimos los mejores resultados con los modelos de regresión logística, el modelo de vectores soporte (SVM) y la combinación de éstos con técnicas como bagging o boosting.

Palabras clave: TF-IDF, SVM, Regresión Logística, Sexismo, Redes sociales, Codificador de frases, Twitter.

1. Introducción

Desde finales del siglo XIX se ha cuestionado el papel de la mujer y se ha luchado por la igualdad de ambos géneros. Se trata de uno de los temas que más concierne a la sociedad actual que busca la eliminación de toda muestra de descriminación e injusticia por motivos de sexualidad , en especial, aquellos referentes a la mujer. No obstante, el extendido uso de las redes sociales dificulta dicha tarea al generarse grandes volúmenes de textos de naturaleza muy dispersa. En especial, Twitter destaca como medio de difusión de opiniones. La creación de un sistema de detección del sexismo es fundamental para mejorar la convivencia dentro de la comunidad de usuarios de las redes sociales así como poner un granito de arena en la eliminación de estos comportamientos. Se han realizado algunas investigaciones al respecto, tanto en documentos públicos [1] como en académicos[2], así como con tweets [3, 4].

El gran reto de identificar aquellos mensajes con prejuicios, estereotipos o discriminantes reside en la diferenciación de aquellos mensajes con uso de lenguaje figurativo , en especial, aquellos que puedan sonar de primeras "amigables" y "gracioso" pero son ofensivos y viceversa. El presente estudio tiene como objetivo la identificación automática de este amplio espectro de posibilidades de textos sexistas mediante el uso de técnicas y modelos de machine learning. Para poder alcanzarlo, se han experimentado con varias técnicas de aprendizaje automatizado (Regresión Logística, Máquinas de Vectores Soporte Support (SVM), Random Forest, Árboles de Decisión, Naive Bayes, Perceptrón Multicapa y BERT -Representaciones de codificadores bidireccionales a partir de transformadores como BERT-) , comparado el resultado en la tarea.

En este trabajo, proponemos varios sistemas formado por la combinación de cuatro estrategias diferentes para completar , por cada idioma, las dos tareas propuestas por la organización. Por un lado, desde un

enfoque más genérico , realizar una clasificación binaria para decidir si un tweet o un gab es sexista, es decir, es sexista en sí mismo, describe una situación sexista o critica un comportamiento sexista[5]. Por otro lado, el sistema debe categorizar aquellos mensajes sexista según el tipo de contenido en una de las cinco categorías siguientes: Desigualdad-Ideológica, Dominación-Estereotipos, Objetivación, Violencia-sexual y Misoginia-No-Violencia-Sexual.

2. Extracción de los datos

Como se describe en las bases de la tarea [5], se recogieron 6977 tweets para entrenamiento dentro del período del 1 al 31 de diciembre de 2020 y 3.386 tweets para test escritos del 1 al 28 de febrero de 2021. De ambos conjuntos se han seleccionado al azar respectivamente 9.000 y 4.000 conjuntos etiquetados. De esta forma se ha garantizado el equilibrio de las clases así como la separación temporal de ambos. Además, se han recopilado 492 "gabs" en inglés y 490 en español siguiendo un procedimiento similar al descrito anteriormente. No obstante, este conjunto se incluye en el test para así poder medir la diferencia entre las redes sociales con y sin control de contenido.

Para realizar el análisis, es necesario diferenciar entre ambos idiomas ya que a pesar de que es posible que tengan la misma estructura general, tanto el preprocesamiento como los modelos creados pueden diferir al no tratarse de lenguas con la misma morfología y sintaxis. Por este motivo, trabajamos inicialmente con los textos español y más tarde con los mensajes en inglés.

3. Preprocesamiento y extracción de características

Se ha realizado el siguiente preprocesamiento de los mensajes puesto que no es posible entrenar los sistema de aprendizaje tradicionales con textos en crudo.

En primer lugar, se ha realizado una **normalización** del texto. Para ello, se planteó inicialmente la división de los **hashtags** (que normalmente están formados por varias palabras unidas) en base a que la segunda palabra estuviera escrita la primera letra en mayúscula (por ello, se analizará previamente a la conversión a minúsculas). No obstante, observamos que no siguen dicho patrón por lo que optamos por eliminar el símbolo # y mantener el contenido al poder contener información relevante y distintiva. Una vez descartada dicha idea, se siguieron los siguientes pasos:

- **Conversión** de todas las letras a minúsculas.
- **Eliminación del ruido**. Para ello, se han eliminado, utilizando expresiones regulares, las fechas puesto que consideramos que para el tema a tratar no son relevantes (se ha eliminado también aquellas escritas en letra como por ejemplo 15 de marzo de 2021). Por otro lado se han suprimido los signos de puntuación, los posibles emails, las urls y el símbolo # por el motivo mencionado previamente. Utilizando la librería tweet- preprocessor se han descartado los emoticonos (smiley), los emojis así como las palabras reservadas de Twitter y los números (lo que incluye las horas y las fechas escritas en este formato). Cabe destacar que se ha realizado una conversión de todas aquellas menciones de usuarios por la palabra "user" para así mantener en el análisis la información de que se ha llamado a un usuario (no es relevante quien).
- **Tokenización**. Dividiendo el texto en palabras.
- **Eliminación de palabras irrelevantes (stopwords)**, es decir, aquellas palabras presentes en todos los textos y que no aportan ningún significado adicional o distintivo del tweet.

Así pues, a la vez que se eliminan las palabras irrelevantes (lo que permite una mayor eficiencia) se realiza una **estemización** de la palabra, es decir, de aquellas palabras que son importantes se obtiene su raíz.

Hasta este momento se ha obtenido una serie de cadenas de texto con aquella información que se ha considerado relevante. No obstante, estas secuencias de palabras no pueden alimentar directamente a los algoritmos puesto que la mayoría de ellos esperan vectores de características numéricas con un tamaño fijo. Por este motivo, se transformaron los textos libres que poseen una longitud variable en vectores de características numéricas de tamaño fijo siguiendo la estrategia **Codificación TF-IDF.: (Term Frequency - Inverse Document Frequency)**. Esta técnica consiste en una medida numérica que expresa cuán relevante es una palabra para un texto en un corpus. Busca reducir el peso de los términos en proporción al número de textos en los que aparecen. De esta forma, el valor de un término aumenta proporcionalmente con el número de veces que aparece en el texto, siendo compensado por su frecuencia en el corpus. Para realizar esta transformación utilizamos la función **Tf idf Transformer** y fijamos el parámetro `min_df = 2` para así ignorar los términos de nuestros datos cuya frecuencia fuera inferior a 2. De esta forma, no tenemos en cuenta aquellas palabras que aparecen solo una vez ya que consideramos que no son relevantes para la descriminación de textos.

4. Modelos y Evaluación

Una vez obtenido el formato correcto de los textos de entrada para los modelos, se ha procedido a la selección del mejor sistema de clasificación. Hemos probado diferentes **modelos** con sus diferentes parámetros buscando maximizar las dos métricas escogidas por los organizadores: Accuracy para la Identificación de sexismo y el F1 Score para la Categorización de los tweets y mensajes Gab según su tipo de sexismo. Entre los modelos tradicionales de la librería sklearn que hemos utilizado se encuentran *Naive Bayes, SVM, Regresión Logística, Árboles de decisión* y *Perceptrón multicapa*. Además, probamos con *BERT* y la combinación de los diferentes modelos mediante técnicas como *Bagging* y *Boosting*. Así pues, en Bagging se ajustan múltiples modelos cada uno con un subconjunto distinto de los datos de entrenamiento mientras que en Boosting se entrenan secuencialmente múltiples modelos sencillos de forma que cada modelo aprende de los errores del anterior.

En cuanto a la estrategia de validación, hemos utilizado la técnica de **validación cruzada** para evaluar los resultados obtenidos por cada modelo. Esta técnica consiste en dividir el dataset train en subconjuntos de entrenamiento y validación lo que nos permite calcular una probabilidad de ser o no sexista (o la categoría sexista del texto) y por tanto una clasificación en dichos grupo más ajustada a la realidad. En esta etapa se ha decidido limitar el k-fold a 10.

Para el ajuste de parámetros hemos utilizado Grid Search. Esta técnica considera exhaustivamente todas las combinaciones de hiperparámetros de un modelo que le pases a la función **GridSearchCV** de la librería sklearn. Nos devuelve, por tanto, la combinación de hiperparámetros que maximiza la métrica de evaluación.

5. Resultados

5.1 Resultados TAREA 1 Identificación del Sexismo.

Para los textos en español, los mejores modelos fueron AdaBoost Regresión Logística, Regresión Logística, BERT y SVM.

MODELO	ACCURACY	F1
AdaBoost Regresión Logística	0.73144	0.42031
Regresión Logística	0.72859	0.6807
BERT	0.728	-
SVM	0.72407	0.6755

Tabla 1: Resultados Identificación Sexismo Español

Los mejores parámetros para el modelo **AdaBoost Regresión Logística** fueron class_weight = 'balanced', C= 0.8, max_iter = 100, solver = 'newton-cg' y n_estimators=100. Para **Regresión Logística** class_weight = 'balanced',C= 0.8, max_iter = 100, solver = 'newton-cg' y n_estimators=100. En cuanto a **BERT** bert-base-multilingual-uncased, max_seq_length = 70 y epochs' = 4. Y por último, para **SVM** los mejores resultados los obtuvimos con los parámetros class_weight = 'balanced', gamma = 'scale', C = 0.75 y kernel= 'linear'.

Para el inglés, los mejores modelos fueron dos que también funcionaron bien en el español: SVM y Regresión Logística.

MODELO	ACCURACY	F1
Regresión Logística	0.73225	0.72809
SVM	0.72439	0.71904

Tabla 2: Resultados Identificación Sexismo Inglés

Los mejores parámetros para la **Regresión Logística** fueron class_weight = 'balanced', C = 0.6, max_iter = 100 y solver = 'liblinear'. Para el **SVM** fueron class_weight = 'balanced', gamma = 'scale', C = 1, kernel = 'rbf'

5.2 Resultados TAREA 2 Categorización del Sexismo

En cuanto a la TAREA 2 Categorización del Sexismo los modelos que dieron mejores resultados fueron similares a los de la tarea anterior.

Para el español fueron Regresión Logística, Bagging Regresión Logística, SVM y Bagging SVM. Para el inglés los modelos eran los mismos: SVM y Regresión Logística.

MODELO	ACCURACY	F1
Regresión Logística	0.64047	0.52221
Bagging Regresión Logística	0.63849	0.51932
SVM	0.62974	0.51237
Bagging SVM	0.63481	0.50955

Tabla 3: Resultados Categorización Sexismo Español

Los mejores parámetros para la Regresión Logística fueron class_weight = 'balanced', C= 1, max_iter = 100 y solver = 'sag'. Para **Bagging Regresión Logística** class_weight = 'balanced',C= 1, max_iter = 100, solver = 'sag' y n_estimators=10. En cuanto a **SVM** C = 0.75, kernel = 'linear', class_weight = 'balanced' y gamma = 'scale'. Y por último, para **Bagging SVM** los mejores resultados los obtuvimos con los parámetros class_weight = 'balanced',gamma = 'scale', C = 0.75, kernel= 'linear' y n_estimators=10.

MODELO	ACCURACY	F1
Regresión Logística	0.63068	0.48348
SVM	0.61583	0.45990

Tabla 4: Resultados Categorización Sexismo Inglés

Los mejores parámetros para la **Regresión Logística** fueron class_weight = 'balanced', C = 0.6, max_iter = 100 y solver = 'liblinear' y para **SVM** class_weight = 'balanced', gamma = 'scale', C = 1 y kernel = 'rbf'.

6. RUNS competición EXIST 2021

En ambas tareas enviamos los modelos SVM y Regresión Logística tanto para el español como para el inglés. El tercer run fue una combinación del modelo Regresión Logística para el inglés junto a Adaboost Regresión Logística en la tarea 1 y Bagging Regresión Logística para la tarea 2 en español. Los resultados se muestran en las siguientes dos tablas.

Posición	Modelo Español	Modelo Inglés	Accuracy	F1	Nombre Archivo
42	SVM	SVM	0.7072	0.7068	task1_Nerin_1
47	Regresión Logística	Regresión Logística	0.7022	0.7016	task1_Nerin_2
50	Adaboost Regresión Logística	Regresión Logística	0.691	0.690	task1_Nerin_3

Tabla 5: Resultados Competición Identificación Sexismo

Posición	Modelo Español	Modelo Inglés	Accuracy	F1	Nombre Archivo
35	Bagging Regresión Logística	Regresión Logística	0.604	0.481	task2_Nerin_3
36	Regresión Logística	Regresión Logística	0.60	0.471	task2_Nerin_2
48	SVM	SVM	0.582	0.423	task2_Nerin_1

Tabla 6: Resultados Competición Categorización Sexismo

El mejor puesto en la tarea de identificación del sexismo lo obtuvimos con el modelo **SVM** tanto para el inglés como para el español. Quedamos en la posición 42. Por otro lado, el mejor puesto en la tarea de categorización del sexismo lo obtuvimos con el modelo **Bagging Regresión Logística** para el **español y Regresión Logística** para el **inglés**. En este caso logramos la posición 35.

El modelo SVM nos funcionó muy bien en la primera tarea para clasificar solo entre dos etiquetas posibles: sexismo o no sexismo. Sin embargo, para clasificar los textos entre las 5 posibles categorías de sexismo este fue el modelo que peor funcionó. La Regresión Logística obtuvo buenos resultados en ambas tareas. Y por último, para el tercer "run" los resultados fueron diferentes para ambas tareas. Mientras que en la primera la combinación de Regresión Logística junto a Adaboost Regresión Logística fue el peor de los 3, para la segunda la combinación de Regresión Logística con Bagging Regresión Logística fue la que mejor funcionó.

7. Conclusión

Llegados a este punto podemos concluir que hemos alcanzado nuestro objetivo de crear un sistema de identificación de sexismo en las redes sociales.

La capacidad computacional fue una de las grandes limitaciones que afrontamos durante el presente estudio. Tuvimos grandes dificultades a la hora de entrenar los modelos y encontrar la mejor combinación de parámetros debido al tiempo que suponía la ejecución de cada función. Asimismo, nos fue imposible implementar algún modelo de transformación de Huggingface [6] ya que cada modificación nos suponía horas de trabajo. Por este motivo, solo pudimos entrenar modelos básicos de BERT de la librería de sklearn. Por tanto, en trabajos futuros con equipos informáticos más potentes, mejoraremos nuestros sistemas aplicando y utilizando otras variantes de BERT y seleccionando nosotros los diferentes parámetros que influyen en estos modelos .

Finalmente , destacar que todo el proceso realizado junto a los runs enviados a la competición están en la carpeta Proyecto EXIST 2021.

8. Referencias

[1] Ramos Talens, P., 2018. *Detección de lenguaje sexista en documentos*. [online] Riunet.upv.es. Available at: <https://riunet.upv.es/bitstream/handle/10251/93786/RAMOS%20-%20Detecci%c3%b3n%20de%20lenguaje%20sexista%20en%20documentos.pdf?sequence=1&isAllowed=y>

[2]Pedro Orgeira-Crespo, 2020 Decision Algorithm for the Automatic Determination of the Use of Non-Inclusive Terms in Academic Texts <https://www.mdpi.com/2304-6775/8/3/41/htm >

[3] Rodríguez Sánchez, F., Carrillo de Albornoz, J. and Plaza, L., 2020. *Automatic Classification of Sexism in Social Networks: An Empirical Study on Twitter Data*. [online] Ieeexplore.ieee.org. Available at: <https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9281090>

[4]Dylan Grosz, Patricia Conde-Cespedes. Automatic Detection of Sexist Statements Commonly Used at the Workplace. Pacific Asian Conference on Knowledge Discovery and Data Mining (PAKDD), Wokshop (Learning Data Representation for Clustering) LDRC, May 2020, Singapour, Singapore. ffhal-02573576f

[5] EXIST. 2021. *EXIST*. [online] Available at: <http://nlp.uned.es/exist2021/>

[6] GitHub. 2021. *Hugging Face*. [online] Available at: <https://github.com/huggingface>