Universidad Autónoma de Nuevo León

Facultad de Ciencias Físico Matemáticas

Maestría en Ciencia de Datos

Tarea 3 - Clasificación de ciberbullying

Autor:
Lic. Leobardo García Reyes

Supervisado por:
Dr. Mayra Cristina Berrones Reyes

2 de junio del 2022

Clasificación de ciberbullying

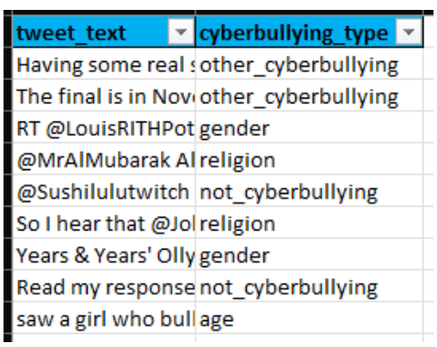
Datos

Las redes sociales tomaron un gran auge junto con la pandemia de Covid-19, trayendo consigo un gran aumento de ciberbullying, esto se puede combatir creando modelos de clasificación para detectar aquellos tweets que inciten al odio

El conjunto de datos que se seleccionó para aplicar preprocesamiento de texto y analizar, se obtuvo a través del sitio web "Kaggle"[2]. El conjunto de datos es una recopilación de más de 47,000 tweets etiquetados según la clase de ciberbullying: Edad (Age), Etnicidad (Ethnicity), Género (Gender), Religión (Religion), Otro tipo de ciberbullying (Other type of cyberbullying), No es ciberbullying (Not cyberbullying).

El conjunto de datos se ha equilibrado para contener maso menos 8,000 tweets de cada clase. Se proporciona de manera tabulada de la siguiente manera (Fig. 1):

- **tweet_text:** Texto que escribió el usuario.
- **cyberbullying_type:** Tipo de ciberbullying al que pertenece el tweet. Los cuales son: Edad (Age), Etnicidad (Ethnicity), Género (Gender), Religión (Religion), Otro tipo de ciberbullying (Other type of cyberbullying), No es ciberbullying (Not cyberbullying).



tweet_text	cyberbullying_type
Having some real s	other_cyberbullying
The final is in Nov	other_cyberbullying
RT @LouisRITHPot	gender
@MrAlMubarak Al	religion
@Sushilulutwitch	not_cyberbullying
So I hear that @Jol	religion
Years & Years' Olly	gender
Read my response	not_cyberbullying
saw a girl who bull	age

Figura 1: Conjunto de datos en disposición. Fuente: Imagen realizada por mi.

Como eran demasiados tweets, esto creaba un problema para realizar los entrenamiento del modelo. Entonces, se opto por elegir solo el 62.90 % de los datos de manera aleatoria, quedando de los 47,692 solo 30,000.

En la Fig. 2, se muestra la cantidad de tweets realizados por tipo de ciberbullying. Hay una cantidad equilibrada de cada clase (Cua. 1), lo que beneficiará en el entrenamiento de los modelos para que en las métricas se obtenga un mejor resultado.

Tipo de ciberbullying	**Número de tweets**
Age	5091
not_cyberbullying	5052
gender	5025
ethnicity	4971
religion	4931
other_cyberbullying	4930

Cuadro 1: Conteo de tweets realizados por tipo de ciberbullying.

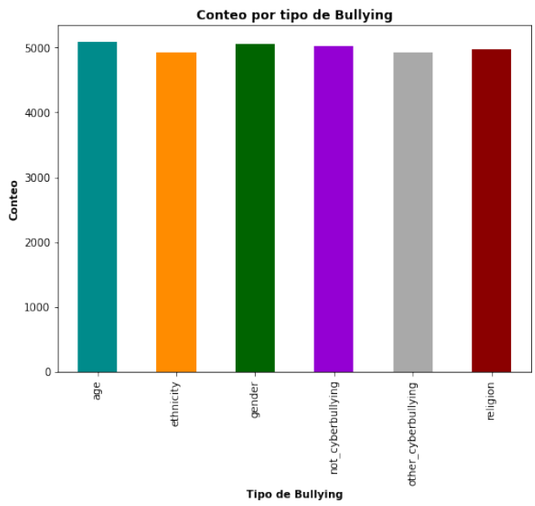
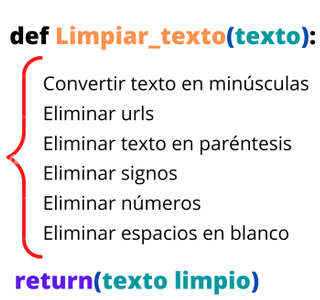


Figura 2: Conteo de tweets realizados por tipo de ciberbullying.

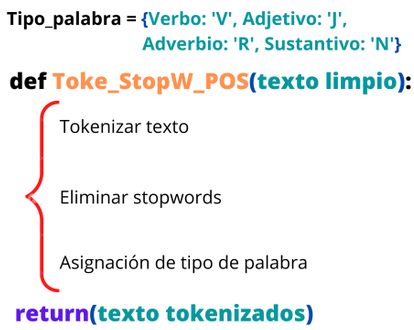
Metodología

Pre-procesamiento

En esta sección se describirá el preprocesamiento implementado en los tweets por ciberbullying. El preprocesamiento, es utilizado para remover texto que no sea relevante para el futuro análisis como son las urls, signos, números, espacios en blanco, etc. Para esto, en la Fig. 3a, se explica la función empleada para limpiar el texto como primera parte.



(a) Función utilizada para realizar una primera limpieza de texto.



(b) Función utilizada para tokenizar, remover stopwords y lematizar.

Figura 3: Funciones utilizadas para limpiar, tokenizar, remover stopwords y lematizar. Fuente: Imágenes realizadas por mi.

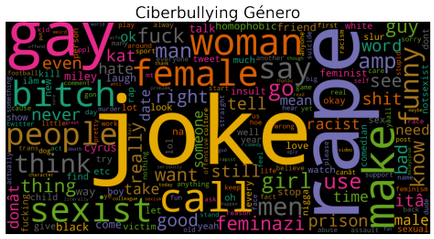
Una vez realizado una primera limpieza, es importante remover las palabras que no aporten información adicional como son las stopwords. También, es importante dejar las palabras a su lema y para esto, se empleó el método de Lematización. La Lemmatization, incorpora información sobre la parte del discurso del término utilizando diccionarios y un análisis morfológico mas complejo.

Para aprovechar el potencial del método Lematización a cada palabra tokenizada, se le asignará una letra dependiendo su contexto (Fig. 3b). Las cuales serían: V - Verbo, J - Adjetivo, N - Sustativo, R - Adverbio, esto sirve para transformar y agrupar las palabras raíz de una mejor manera.

Una vez realizado todo este proceso de limpieza de texto, se decidió de igual forma, implementar un función que elimine aquellas palabras que tengan menos de 2 caracteres. También, se añadió al listado de stopwrods nuevas palabras como rt, mkr, didn, bc, n, m, im, ll, y, ve, u, ur, don, t, s, amp y números del 0 al 10 escritos en texto.

Resultados

En la Fig. 4, se muestra los tipo de ciberbullying Género, Religión, Étnico y Edad, representados en un gráfico de nube de palabras. En cada una de esta se puede observar las palabras más utilizadas, en donde, cada una de estas parece haber una clara diferencia.



(a) Nube de palabras de ciberbullying género.

(b) Nube de palabras de ciberbullying religión.

(c) Nube de palabras de ciberbullying étnico.

(d) Nube de palabras de ciberbullying edad.

Figura 4: Nube de palabras de ciberbullying: Género, Religión, Étnico y Edad.

En el Cua. 2, se construyo a partir de la función chi2 de python, que mide la dependencia entre variables estocásticas, por lo que el uso de esta función ."elimina"las características que tienen más probabilidades de ser independientes de la clase y, por lo tanto, irrelevantes para la clasificación. Para cada uno de los tipos de ciberbullying se obtuvo las 3 primeras palabras más relevantes de una sola palabras y de dos palabras. Se observa como es que concuerda el Cua. 2 con la Fig. 4

Unigramas	**Bigramas**
Edad	
bully	bully high
high	school bully
school	high school
Étnico	
fuck	fuck obama
dumb	dumb fuck
nigger	dumb nigger
Género	
gay	joke gay
rape	gay joke
joke	rape joke
Religión	
christian	radical islamic
idiot	christian woman
muslim	islamic terrorism

Cuadro 2: Palabras más comunes por tipo de ciberbullying.

Para poder elegir un buen modelo que clasifique cada tweet con su correspondiente tipo de ciberbullying se corrieron 4 diferentes modelos: Random Forest, Linear SVC, Multinomial Naive Bayes, Logistic Regression. La métrica utilizada es Exactitud (Accuracy), ya que se tiene clases equilibradas. También se utilizó una validación cruzada de 10 para detectar el sobreajuste, es decir, en aquellos casos en los que no se logre generalizar un patrón.

En la Fig. 5, se observa cada modelo con su correspondiente diagrama de caja que indica su evaluación con la métrica Exactitud. El modelo Logistic Regression parece indicar que tiene un mejor rendimiento, siguiéndole Linear SVC. Para saber si existe una verdadera diferencia, se me ocurre realizar una prueba estadística para comparar las medias y saber si son iguales. Antes de eso, aplicaría también una prueba de varianza para saber si son iguales o diferentes, tratándose de métodos paramétricos. En métodos no paramétricos, usaría la prueba de Wilcoxon.

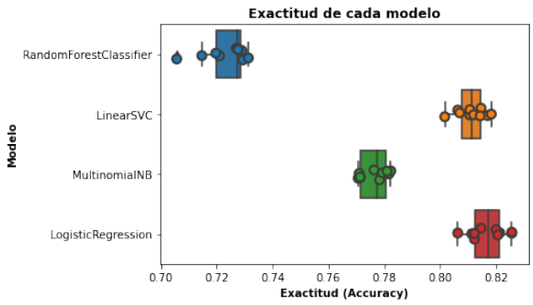


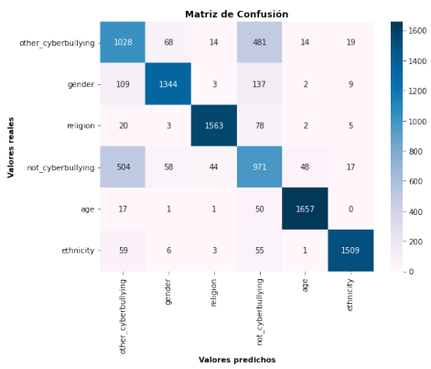
Figura 5: Evaluación de rendimiento de cada modelo con la métrica Exactitud (Accuracy).

En el Cua. 3, se muestra el promedio de Exactitud que tuvo cada modelo. Como se mencionó anteriormente, es posible que no exista una diferencia significativa entre Logistic Regression y Linear SVC.

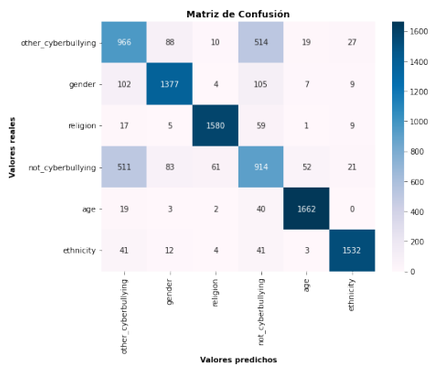
Modelo	**Exactitud**
Linear SVC	0.811267
Logistic Regression	0.817033
Multinomial Naive Bayes	0.776333
Random Forest	0.723300

Cuadro 3: Promedio de Exactitud de cada modelo.

En la Fig. 6, se muestra la matriz de confusión correspondiente de los dos mejores modelos. Se observa que son muy parecidos y tienen mayor confusión en las mismas clases, las cuales son other_cyberbullying y not_cyberbullying.



(a) Matriz de confusión con el modelo Logistic Regression.



(b) Matriz de confusión con el modelo Linear SVC.

Figura 6: Matrices de confusión con los modelos Logistic Regression y Linear SVC.

Una vez entrenado los modelos, nuevamente para cada uno de los tipos de ciberbullying se obtuvo las 3 primeras palabras más relevantes de una sola palabras y de dos palabras como se observa en el Cua. 4. Al compararlas, se puede apreciar que no existe mucha diferencia entre los modelos y esto puede ayudar a sostener a que no existe una diferencia significativa entre los modelos.

Logistic Regression	
Unigrama	**Bigrama**
Edad	
school	school bully
high	high school
bully	bully school
Étnico	
nigger	dumb nigger
dumb	dumb fuck
black	fuck dumb
Género	
rape	rape joke
gay	sexist woman
sexist	gay joke
Religión	
muslim	christian woman
christian	islamic terrorism
idiot	âchristianâ woman

Linear SVC	
Unigrama	**Bigrama**
Edad	
school	school bully
high	bully school
girl	bully high
Étnico	
nigger	dumb fuck
negro	dumb nigger
dumb	look stupid
Género	
rape	rape joke
gay	sexist woman
sexist	male friend
Religión	
muslim	christian woman
radical	islamic terrorism
christian	âchristianâ woman

Cuadro 4: Palabras más comunes por tipo de ciberbullying y por los modelos Logistic Regression y Linear SVC.

Se escogieron 5 tweets al azar para que sean clasificados de acuerdo a su contenido como se muestra en el Cua. 4. En los dos modelos se llegó al mismo resultado, los cuales fueron correctamente clasificados.

Tweets de prueba	Logistic Regression	Linear SVC
@LoLosWay I dont́ want to see a pic of your dumb nigger ass bitch fuck off'	Étnico	Étnico
Tweeting about the Muslim brotherhood and turkey and talking shit is good sport, you will have hoards of brainless idiots storming your feed	Religión	Religión
RT @OliveWahh: The scoring wasnt́ even done honestly #MKR	No Ciberbullying	No Ciberbullying
Tzuyu looked like a genius high school girl from a kdrama whoś always bullied but always saved by two handsome campus crushes	Edad	Edad
why on earth did i take this god awful fucking shift oh my god	Other Ciberbullying	Other Ciberbullying

Cuadro 5: Tweets para clasificar con los modelos Logistic Regression y Linear SVC.

En la Fig. 7, se muestra el informe de clasificación de métricas correspondiente a los modelos. También se puede observar que no hay mucha diferencia entre los 2 modelos.

	precision	recall	f1-score	support
0	0.59	0.63	0.61	1624
1	0.91	0.84	0.87	1604
2	0.96	0.94	0.95	1671
3	0.55	0.59	0.57	1642
4	0.96	0.96	0.96	1726
5	0.97	0.92	0.95	1633
accuracy			0.82	9900
macro avg	0.82	0.81	0.82	9900
weighted avg	0.82	0.82	0.82	9900

(a) Informe de clasificación de métricas con el modelo Logistic Regression.

	precision	recall	f1-score	support
0	0.58	0.59	0.59	1624
1	0.88	0.86	0.87	1604
2	0.95	0.95	0.95	1671
3	0.55	0.56	0.55	1642
4	0.95	0.96	0.96	1726
5	0.96	0.94	0.95	1633
accuracy			0.81	9900
macro avg	0.81	0.81	0.81	9900
weighted avg	0.81	0.81	0.81	9900

(b) Informe de clasificación de métricas con el modelo Linear SVC.

Figura 7: Informe de clasificación de métricas con los modelos Logistic Regression y Linear SVC.

Conclusiones

El experimentar con diferentes modelos, es de suma importancia, ya que así se puede dar cuenta que modelo tiene mejor rendimiento al momento de clasificar. Si solo se hubiera utilizado Random Forest, se hubiera tenido un pésimo rendimiento.

Es muy importante como es que la estadística puede ser una herramienta complementaría para el machine learning, porque así también se puede sustentar de manera precisa si dos modelos son iguales o no hay mucha diferencia significativa.

Algo importante a destacar, es la desventaja de usar un modelo para clasificar estos tipos de texto, y es que puede existir un tweet que no incite al odio o ataque a una persona y tan solo se este dando la opinión de uno de estos temas. Tal vez el modelo lo clasifique como malo, cuando realmente no es así.

Referencias

[1] GitHub https://github.com/Zarcklet/ProcesamientoClasificacionDatos

[2] Larxel. (2022). Cyberbullying Classification [Data set]. En Cyberbullying Classification. https://www.kaggle.com/datasets/andrewmvd/cyberbullying-classification.

[3] AWS (2022). Validación cruzada. https://docs.aws.amazon.com/es_es/machine-learning/latest/dg/cross-validation.html.

[4] Pedregosa et al. (2011). Sklearn.naive - bayes. MultinomialNB. Machine Learning in Python. Scikit-learn, 12(85). https://scikit-learn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html.

[5] Pedregosa et al. (2011). Sklearn.svm.LinearSVC. Machine Learning in Python. Scikit-learn, 12(85). https://scikit-learn.org/stable/modules/generated/sklearn.svm.LinearSVC.html.

[6] Waskom, M. (2021). Seaborn.boxplot. https://seaborn.pydata.org/generated/seaborn.boxplot.html.

[7] Waskom, M. (2021). Seaborn.stripplot. https://seaborn.pydata.org/generated/seaborn.stripplot.html.