Identificación de lenguaje ofensivo en mensajes de texto,

utilizando técnicas de aprendizaje automático.

Kevin Steven Ocampo Morales
Pontificia Universidad Javeriana Cali
kestoc@javerianacali.edu.co

Juan Sebastian Arango Salazar
Pontificia Universidad Javeriana Cali
juan.sebastian.7@hotmail.com

31 de enero del 2024

Resumen

Este proyecto de investigación se centró en el estudio y desarrollo de modelos de aprendizaje automático supervisado, incluyendo variantes de Naive Bayes, máquinas de soporte vectorial y redes neuronales convolucionales, con el propósito de identificar y clasificar tweets como ofensivos o no ofensivos. A lo largo de esta investigación, se siguieron varios pasos fundamentales que desempeñaron un papel importante en la creación de los modelos finales. Los diversos procesos experimentales desarrollados a lo largo de la investigación arrojaron resultados de relevancia. Inicialmente, se implementaron modelos base predeterminados disponibles en las librerías. A medida que avanzábamos e iteramos, además de la constante incorporación de métodos y técnicas más avanzadas que permitían enriquecer y perfeccionar los modelos. Al concluir la investigación tanto los modelos de Naive Bayes, junto con el modelo de máquinas de soporte vectorial, arrojaron resultados excelentes durante las fases de entrenamiento, pero al momento de comprobar con la fase de prueba los resultados fueron deficientes. A pesar de implementar diversas estrategias, métodos y técnicas para mejorar su eficacia en el proceso de la clasificación de tweets, no se logró un desempeño satisfactorio debido a problemas de sobreajuste. Además, el modelo de redes neuronales, junto con las técnicas implementadas para optimizar su rendimiento, demostró ser efectivo al proporcionar resultados satisfactorios. En resumen, este estudio facilitó la exploración de diversos métodos y técnicas en el desarrollo de modelos de clasificación, destacando la relevancia de la iteración continua para el constante perfeccionamiento de la investigación.

Introducción

En los últimos años, el uso de foros en línea y sitios web ha venido en un incremento constante y potencialmente positivo tanto para las empresas como para la sociedad actual, esto debido a que la tecnología está cada vez más presente en la vida cotidiana. Twitter, una red social muy conocida
y relevante a nivel mundial en la actualidad, cuenta con millones de usuarios y por ende el tráfico de tweets es realmente alto. Estos medios digitales permiten que las personas estén conectadas a través de internet, permitiéndoles expresar de manera abierta y pública sus opiniones e ideas sobre temas presentes o de gran controversia, a la vez que se informan o comunican a otros sobre lo que acontece en el día a nivel mundial o personal.

Es por este amplio grado de libertad que tiene la sociedad en estos medios que, en muchos casos, debido al gran volumen de datos y el contenido ofensivo que puede existir en este, ocasiona que la supervisión y control llegue a ser mínimo. Adicional a esto, la máscara del anonimato que las redes sociales ofrecen, incentiva a los individuos a llevar a cabo este tipo de comportamientos y comentarios con el fin de ofender, lastimar o discriminar, afectando la autoestima de las personas y su estado mental, así mismo generando un impacto negativo en la sociedad. Debido a que los usuarios en las redes sociales producen grandes volúmenes de información, es imposible monitorear de forma manual este tipo de contenido; por esto es por lo que se han desarrollado modelos de aprendizaje automático que, de manera autónoma, detectan este contenido.
En consecuencia, la identificación oportuna del lenguaje ofensivo se convierte en un reto importante por tratar. Es por esto por lo que nuestro proyecto busca entrenar modelos de aprendizaje automático supervisado que identifiquen y clasifiquen de manera apropiada los tweets entre ofensivos y no ofensivos, además de evaluar el desempeño de estos y analizar los resultados obtenidos.

Fundamentación teórica

Preprocesamiento

El preprocesamiento de datos es un paso fundamental para el desarrollo de modelos de aprendizaje automático. Este proceso nos permite garantizar la calidad y fiabilidad de los resultados obtenidos al entrenar y evaluar modelos. Analizando métodos, herramientas,procedimientos y/o técnicas que permiten manipular, seleccionar, transformar, reducir y limpiar el conjunto de datos con el que trabajaron los modelos. Para este proyecto se hizo uso de variedad de librerías y/o herramientas de procesamiento de lenguaje natural y aprendizaje automático. Cada una suministra un conjunto de utilidades, funciones o métodos esenciales para la manipulación de los datos y el desarrollo de los modelos, tales como: Pandas, NLTK, Sklearn, Wordninja, Keras, etc. Por otro lado, una debida y óptima limpieza y normalización del texto, permiten tener homogeneidad en los datos y manejar un estándar en ellos y por ello se realizaron pasos como: Eliminación de caracteres inválidos, conversión a minúsculas, manejo de stopwords, etc.

El siguiente paso es la tokenización el cual es un proceso que divide el texto en partes más pequeñas (tokens), permitiendo representar de forma vectorial los datos, haciendo que sea más fácil el análisis por parte del algoritmo. Luego tenemos la ingeniería de características, la cual desempeña un papel crucial en el preprocesamiento de datos en tareas relacionadas con el procesamiento de lenguaje natural. Esta práctica tiene como finalidad extraer información relevante y obtener representaciones adecuadas del texto a procesar, permitiendo a los modelos comprender de manera más efectiva el lenguaje humano y es por ello que las técnicas utilizadas fueron: Bolsa de palabras, TF-IDF y Word embedding. Para finalizar se tiene la separación de datos y el balanceo de clases, donde se escogió un 80% de los datos disponibles para entrenar los modelos y el 20% restante para el proceso de validación, además de aplicar técnicas de balanceo de clases como submuestreo y sobremuestreo para manejar este problema.

Modelos entrenados

Las tres técnicas de aprendizaje automático que se utilizan en el proyecto son: Support Vector Machines (SVM), Redes Neuronales (NN) y Naive Bayes. Las máquinas de soporte vectorial (SVM) son un algoritmo de aprendizaje supervisado que se usa en gran medida para tareas de clasificación y regresión de datos, se caracteriza por ser altamente efectivo en la clasificación de datos y tiene bastante uso en el análisis de opiniones y sentimientos. En particular para este proyecto se utilizó la implementación "Classifier" (SVC) de sklearn con el parámetro de kernel, es por ello que resulta fundamental considerar la elección del kernel en el algoritmo, ya que esta elección puede afectar de manera importante en la complejidad temporal de las máquinas de soporte vectorial (SVM).

Por otro lado tenemos las redes neuronales, que son un campo de estudio de la inteligencia artificial y el aprendizaje automático que se basa en el funcionamiento y trabajo del cerebro humano, se encuentran compuestas con conjuntos de unidades de procesamiento interconectadas que simulan las neuronas, estas neuronas trabajan en conjunto para realizar tareas de procesamiento de datos. En particular para este proyecto, las redes neuronales convolucionales fueron la elección ya que demuestran gran eficacia en el procesamiento de datos estructurados, como imágenes, señales o texto. Estas redes neuronales usan capas de convolución en lugar de las capas tradicionales totalmente conectadas, esto permite que se puedan aplicar filtros a ciertas regiones de las entradas y extraer de esta manera características locales relevantes que fluyen a través de la red y se refinan para tener una mejor precisión.

Por último pero no menos importante, tenemos el algoritmo de Naive Bayes, el cual es un algoritmo probabilístico conocido y muy utilizado a causa de su simplicidad y eficacia en tareas de clasificación. Para este proyecto, se eligieron utilizar dos variantes del algoritmo principal, estos son: Naive Bayes Multinomial y Naive Bayes Complementario. El multinomial es un modelo de lenguaje uni grama basado en que las características son representadas como un conteo discreto, esto se refiere al conteo de aparición de una palabra o la frecuencia de aparición de un término en un documento. Mientras que el complementario es adecuado para conjunto de datos desequilibrados, esto es importante a tener en cuenta, ya que el algoritmo estándar de Naive Bayes supone que las clases del conjunto de datos son independientes y bien distribuidos.

Hiperparámetros

Los hiperparámetros son parámetros que se configuran antes de entrenar el modelo. Estos parámetros controlan el funcionamiento del modelo, pero no se aprenden a partir de los datos de entrenamiento. La búsqueda y optimización de estos es el proceso de encontrar los valores de los hiperparámetros que optimizan el rendimiento del modelo y en este caso se realizó con la técnica de búsqueda exhaustiva llamada Grid Search. El objetivo de esta búsqueda y optimización es el mejorar la métrica llamada Recall, que indica la capacidad del modelo para identificar la mayor cantidad posible de tweets ofensivos en el conjunto de datos y reducir de esta manera la cantidad de falsos negativos que generaban los modelos al realizar la tarea de clasificación. La justificación radica en el contexto específico del proyecto, donde se valora más la identificación precisa de tweets ofensivos para evitar pasar por alto aquellos que, aunque no sean explícitamente ofensivos, podrían llegar a serlo. Es por ello que los parámetros a optimizar para cada modelo y

el rango de valores a probar, estuvo determinado por una exhaustiva investigación y documentación de la influencia que dichos parámetros ejercían sobre los modelos y la convergencia de los valores dado un rango y su delimitación tras un proceso de prueba y error.

Resultados

	Predicción Negativa	Predicción Positiva
Actual Negativo	151	89
Actual Positivo	124	116

Modelo	Precision	Recall	F1-score	Accuracy
Submuestreo - Data 2	0.57	0.48	0.52	0.56

Tabla 1: Naïve Bayes complementario

	Predicción Negativa	Predicción Positiva
Actual Negativo	508	112
Actual Positivo	185	55

Modelo	Precision	Recall	F1-score	Accuracy
Submuestreo - Data 1	0.33	0.23	0.27	0.65

Tabla 2: Maquinas de soporte vectorial

Con bolsa de palabras

	Predicción Negativa	Predicción Positiva
Actual Negativo	493	127
Actual Positivo	86	154

Modelo	Precision	Recall	F1-score	Accuracy
Base - Data 1	0.55	0.64	0.59	0.75

Tabla 3: Redes neuronales convolucionales modelo base

Después de analizar los resultados obtenidos al pasar los datos de prueba por los modelos resultantes de Naive Bayes complementario, la máquina de soporte vectorial y la red neuronal convolucional, se logra observar que los dos primeros mostraron un rendimiento deficiente. No

lograron superar el umbral del 50% de efectividad al clasificar los tweets como ofensivos o no ofensivos, lo que sugiere un posible problema de sobreajuste en ambos modelos.

En contraste, la red neuronal convolucional destacó entre estos modelos. Entrenada con el conjunto de datos "Data 1" y evaluada con los datos de prueba, alcanzó un valor de 64% en la métrica de "Recall". Aunque estos resultados no son excelentes, en comparación con los demás modelos, representan una notable mejora y marca una diferencia significativa, esto en gran medida al poder de las redes neuronales y la representación de los datos.

Discusión y conclusiones

En el transcurso de este proyecto, logramos desarrollar 3 modelos para la clasificación, observamos la importancia de diferentes medidas y pasos para el correcto desarrollo y alcance de nuestro objetivo. La clasificación de tweets se reveló como una tarea crítica debido a sus implicaciones tanto en el ámbito físico como en el digital, especialmente en un mundo globalizado e interconectado.

La fase central del proyecto se enfocó en el preprocesamiento de tweets, abordando tareas como limpieza de datos, tokenización y corrección ortográfica. A pesar de los desafíos debido al conflicto entre librerías, se superaron mediante la implementación de funciones y aplicación de pasos adicionales, como la eliminación de palabras vacías y la lematización, con el objetivo de mejorar la calidad semántica. Durante la etapa de ingeniería de características, se utilizaron técnicas como Bolsa de Palabras, tf-idf y word embedding para proporcionar representaciones adecuadas, considerando cuidadosamente la representación del conjunto de datos. En este aspecto, invitamos a futuras investigaciones a explorar diferentes pasos o configuraciones en la etapa de preprocesamiento, ya que se identificó que la variabilidad o la aplicación de diversas técnicas pueden influir en la mejora o empeoramiento de los modelos en el proceso de clasificación.

A pesar de los desafíos, se destaca el rendimiento constante de la red neuronal convolucional, atribuido a la ingeniería de características y, especialmente, al word embedding. Esta técnica, al transformar palabras en vectores de alta dimensionalidad, logró capturar las complejidades lingüísticas y preservar relaciones semánticas, contribuyendo a un análisis más eficaz y preciso en la clasificación de tweets.

El desbalance entre las clases de tweets ofensivos y no ofensivos se identificó como un factor crítico que afectó el rendimiento de los modelos. A pesar de las estrategias aplicadas para contrarrestarlo, los resultados no alcanzaron las expectativas. También se observaron desafíos relacionados con la calidad de los datos y la configuración de pasos en la etapa de preprocesamiento, especialmente en la lematización y corrección ortográfica.

La importancia de este trabajo remarca en gran medida la exploración de nuevas técnicas y diferentes formas de aplicar estos conocimientos para el mejoramiento de los modelos como

también identificar en qué casos ciertas configuraciones ocasionan rendimientos pésimos en estos.

Referencias

- Cortes, C., Vapnik, V. (1995). Support-vector networks. Machine learning, 20(3), 273-297.
- G. Webb, Naïve Bayes. Encyclopedia of Machine Learning and Data Mining, 2017.
- Pedregosa, et al., Scikit-learn: Machine Learning in Python', JMLR 12, pp. 2825-2830, 2011.
- A. McCallum, K. Nigam, A Comparison of Event Models for Naive Bayes Text Classification, AAAI/ICML-98 Workshop on Learning for Text Categorization, 1998.
- J. Schmidhuber, Deep Learning in Neural Networks: An Overview, Arxiv. Cornell University, 2014.
- Y. Lecun, et al., Deep learning, Nature Vol 521, Pag(436-44), 2015.
- M. Zampieri, et al., Predicting the Type and Target of Offensive Posts in Social Media, Proceedings of NAACL, 2019.
- M. pertegal, et al., Revisión sistemática del panorama de la investigación sobre redes sociales:Taxonomía sobre experiencias de uso, Revista científica de educomunicación, vol. 27, 2019.
- G. Gunatilleke, Justifying Limitations on the Freedom of Expression, Hum Rights Rev 22, pp. 91–108, 2021.
- Liddy E, Natural Language Processing, In Encyclopedia of Library and Information Science, 2001.
- Y. Lecun, et al., Deep learning, Nature Vol 521, Pag(436-44), 2015.
- P. Badjatiya, et al., Deep Learning for Hate Speech Detection in Tweets. Proceedings of the 26th International Conference on World Wide Web Companion, 2017.
- TensorFlow. Tokenization and Text Data Preparation with TensorFlow, TensorFlow documentation, 2021.