UNIVERSIDAD NACIONAL DE COLOMBIA

Proyecto Análisis de Bases de Datos
Análisis de sentimientos en noticias y clasificación según sesgo y confiabilidad

Juan Diego Gonzalez Layton, Santiago Botero Daza, David Leonardo Ortiz Uribe, David José Daza Jaimes

Prof. Humberto Sarria Zapata

1. Introducción

En la era digital actual, el análisis de sentimientos se ha convertido en una herramienta clave para entender las opiniones y emociones que subyacen en los contenidos generados por los usuarios o en las publicaciones de los medios de comunicación. El análisis de noticias es particularmente relevante en contextos donde la información es constante y puede tener un impacto significativo en la percepción pública y la toma de decisiones. En este proyecto, se busca emplear diferentes algoritmos de análisis de sentimientos para examinar artículos de cuatro portales de noticias colombianos: El Tiempo, El Colombiano, La Silla Vacía, El Espectador y Semana.

A través de la implementación de estos algoritmos, se pretende evaluar cómo las emociones y opiniones se reflejan en las noticias cubiertas por estos medios, proporcionando una visión más profunda de los sentimientos predominantes en las narrativas informativas. Posteriormente, se utilizarán los resultados de este análisis para generar un embedding, lo que permitirá la representación numérica de los sentimientos detectados. Esta representación se someterá a un proceso de agrupamiento utilizando el algoritmo K-means, con el fin de identificar patrones y tendencias recurrentes en las noticias.

El objetivo final de este proyecto es explorar cómo los sentimientos influyen en la cobertura mediática y si existen agrupaciones o tendencias emocionales comunes en los distintos portales de noticias, proporcionando un análisis valioso para estudios sobre la dinámica de los medios y la percepción pública en Colombia.

1.1. Objetivo General

Identificar el sesgo y el sentimiento general hacia un tema específico, además de intentar encontrar tendencias hacia temas de actualidad relevantes en cuatro de los principales medios de comunicación de Colombia, mediante el análisis de sentimientos aplicado a noticias de El Tiempo, El Colombiano, La Silla Vacía, El Espectador y Semana

1.2. Objetivos Específicos

- Evaluar los resultados obtenidos de diferentes algoritmos para cada uno de los medios
- Encontrar posibles patrones en los medios usando algoritmos de agrupación cómo k-means
- Decidir si existe o no sesgo en cada uno de los medios
- Mostrar estos resultados de manera clara con el fin que el consumidor tome una decisión informada sobre que medio consumir

2. Metodología

Primero se realizó el preprocesamiento de los datos de noticias, eliminando duplicados y aplicando técnicas básicas de limpieza utilizando librerías como `re`, `pandas` y `nltk`, considerando enfoques con y sin lematización. Con ello, se implementó un preprocesamiento avanzado mediante `spaCy`, usando el modelo en español (`es_core_news_sm`) para normalizar el texto y transformar las palabras a su forma canónica, lo que reduce la variabilidad léxica. Con los textos limpios se extrajeron características relevantes aplicando TF-IDF, configurado para extraer n-gramas y limitar el número de términos, lo que permitió identificar los elementos clave

del contenido. Posteriormente se procedió al análisis de sentimientos utilizando `TextBlob`, evaluando la polaridad emocional tanto en los textos sin lematización como en los preprocesados con lematización. Más adelante, para capturar la semántica de los textos se utilizaron técnicas basadas en embeddings, primero con `Sentence Transformers` y luego mediante modelos basados en BERT a través de la librería `transformers`, generando representaciones vectoriales que permiten una comprensión más profunda del contenido. Finalmente, se aplicó clustering usando el algoritmo K-means sobre los embeddings obtenidos, y se redujo la dimensionalidad de los datos con técnicas como TSNE y PCA, lo que facilitó la visualización y la interpretación de la distribución de las noticias en distintos clusters. Todo el proceso se complementó con la generación de visualizaciones mediante `Matplotlib` y `Seaborn` para presentar de forma clara los resultados del análisis.

3. Estado del arte

El estado del arte en algoritmos de análisis de sentimientos ha experimentado una notable evolución, pasando de enfoques basados en léxicos y reglas a métodos que aprovechan técnicas de machine learning y deep learning. Inicialmente, se utilizaron recursos como *SentiWordNet* [1] y otros métodos léxicos que permitían evaluar el sentimiento de forma rudimentaria. Más adelante, se introdujeron técnicas de aprendizaje supervisado –por ejemplo, utilizando algoritmos como SVM y regresión logística junto con representaciones de texto basadas en TF-IDF–, lo que mejoró la precisión en la detección de polaridades.

La revolución en el campo se produjo con la llegada del deep learning. Modelos basados en redes neuronales recurrentes (RNN), LSTM y CNN permitieron capturar dependencias y matices en el lenguaje, pero fue la aparición de los transformers lo que marcó un antes y un después. En particular, BERT [2] y sus variantes han establecido nuevos estándares en la detección de sentimientos al aprovechar técnicas de preentrenamiento y *fine-tuning* que capturan contextos complejos en el lenguaje. Además, estudios como el de Socher et al. (2013) [3] han demostrado la efectividad de los modelos de árboles recursivos en el análisis composicional del sentimiento.

En el análisis de noticias, estos avances permiten abordar desafíos específicos, como la detección de ironía, sarcasmo y sesgos mediáticos. Un ejemplo destacado es la plataforma *Ground News*, que aplica técnicas avanzadas de análisis de sentimientos y NLP para evaluar la polaridad y el sesgo de las fuentes noticiosas. Ground News integra diversas metodologías para comparar cómo distintos medios presentan la misma noticia, ayudando a identificar tendencias y posibles sesgos en la información.

Referencias

[1] Esuli, A., & Sebastiani, F. (2006). SentiWordNet: A publicly available lexical resource for opinion mining. *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06)*.

[2] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805*.

[3] Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A. Y., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. *Proceedings of the 2013 conference on empirical methods in natural language processing (EMNLP)*.