Sentiment Analisis Wellnes Tool

Jesús Aldair Alfonso Pérez, Jose Leyva, Eduardo Maleta

Abstract

Este artículo presenta un estudio comprensivo sobre el análisis de sentimientos en redes sociales utilizando técnicas de aprendizaje automático. Exploramos diversas metodologías para la extracción de características, incluyendo el uso del léxico Empath para cuantificar los tonos emocionales en contenido generado por usuarios. Nuestro análisis emplea clustering KMeans para aumentar los datos y mejorar la eficacia del entrenamiento del modelo. Evaluamos el rendimiento de diferentes modelos de aprendizaje automático a través de validación cruzada estratificada, proporcionando información sobre sus capacidades predictivas. Los resultados indican correlaciones significativas entre las características emocionales y las etiquetas de sentimiento, destacando el potencial para una mejor comprensión de la dinámica de la opinión pública en plataformas de redes sociales.

Palabras clave: Análisis de Sentimientos, Aprendizaje Automático, Redes Sociales, Extracción de Características, Clustering KMeans

1 Introducción

El análisis de sentimientos ha emergido como una herramienta fundamental en la era digital, donde millones de usuarios comparten sus opiniones y emociones a través de plataformas sociales. Con el crecimiento exponencial del uso de redes sociales como Twitter, Facebook e Instagram, se ha generado una cantidad masiva de datos que pueden ser analizados para comprender mejor las percepciones públicas sobre diversos temas, desde productos hasta eventos políticos.

Sin embargo, el análisis de sentimientos presenta varios desafíos. La ambigüedad del lenguaje natural, el uso del sarcasmo y las diferencias culturales en la expresión emocional complican la tarea de clasificar correctamente los sentimientos expresados en los textos. A pesar de estos desafíos, la capacidad para extraer información valiosa a partir de estos datos no estructurados puede influir significativamente en decisiones empresariales y políticas.

Este estudio tiene como objetivo explorar y evaluar la efectividad de diversas técnicas de aprendizaje automático para el análisis de sentimientos en datos extraídos de redes sociales. Se busca determinar qué modelos ofrecen mejores resultados y cómo las características emocionales pueden ser extraídas utilizando herramientas como el léxico Empath. Además, se investigarán patrones en los datos sentimentales mediante técnicas de clustering.

Las preguntas que guiarán esta investigación incluyen: ¿Qué tan precisos son los modelos de aprendizaje automático al clasificar sentimientos en publicaciones de redes sociales? ¿Cómo afecta el preprocesamiento de datos a los resultados del análisis?

El artículo está organizado como sigue: En la sección 2 se presenta una revisión de la literatura relevante; en la sección 3 se detalla la metodología utilizada; la sección 4 presenta los resultados obtenidos; la sección 5 discute las implicaciones y limitaciones del estudio; finalmente, se concluye con reflexiones sobre los hallazgos.

2 Desarrollo

2.1 Datos y preprocesamiento

Los datos utilizados en este estudio provienen de un conjunto de datos disponible en línea, que está catalogado en siete clases diferentes. Cada entrada del conjunto de datos está estructurada en dos columnas: una que contiene el texto y otra que proporciona la etiqueta correspondiente que describe el sentimiento del texto. Este formato permite una fácil manipulación y análisis mediante herramientas de aprendizaje automático.

Los datos recopilados fueron sometidos a un proceso de limpieza que incluyó la eliminación de duplicados y entradas irrelevantes. En una primera instancia, se utilizó una biblioteca especializada para transformar el texto original en temas (topics) que describen la temática general del contenido. Este proceso redujo la cantidad de características a un total de 190. Posteriormente, se analizó la correlación entre características para eliminar aquellas que estaban altamente relacionadas, lo que ayudó a simplificar el conjunto de datos y mejorar la eficiencia del modelo.

Se utilizó el léxico Empath para extraer características emocionales relevantes. Cada texto fue analizado para cuantificar emociones como alegría, tristeza y enojo, además de los temas identificados anteriormente. Los resultados se representaron en una matriz donde cada fila correspondía a un texto y cada columna a una característica emocional o tema.

2.2 Modelos

Se seleccionaron varios modelos para la clasificación de sentimientos, incluyendo Redes Neuronales, KMeans y Redes Neuronales Convolucionales (CNN). Las Redes Neuronales se utilizaron para aprender patrones complejos en los datos textuales, mientras que KMeans se empleó para realizar clustering y agrupar textos similares. Las CNN se aplicaron para capturar características espaciales en los datos textuales, mejorando así la precisión del análisis.

2.2.1 Redes Neuronales

Las Redes Neuronales son modelos inspirados en el funcionamiento del cerebro humano que son capaces de aprender patrones complejos en los datos. Se utilizan en este estudio para la clasificación de sentimientos debido a su capacidad para manejar grandes volúmenes de datos y aprender representaciones no lineales.

En este caso, se implementó una red neuronal densa (fully connected) que toma como entrada la matriz de características extraídas. La red consta de varias capas ocultas que permiten aprender patrones complejos en las emociones expresadas en los textos. Este enfoque es particularmente útil dado que los datos textuales pueden contener relaciones intrincadas entre las palabras y las emociones.

2.2.2 KMeans

KMeans es un algoritmo de clustering que se utiliza para agrupar datos similares en conjuntos. En el contexto del análisis de sentimientos, KMeans se emplea

para identificar grupos de textos que comparten características emocionales similares.

La elección de KMeans se justifica por su simplicidad y eficacia en la identificación de patrones en datos no etiquetados. Al aplicar KMeans, se puede explorar la estructura subyacente del conjunto de datos, lo que proporciona información valiosa sobre cómo se distribuyen los sentimientos dentro del corpus analizado.

2.2.3 Redes Neuronales Convolucionales (CNN)

Las Redes Neuronales Convolucionales (CNN) son especialmente efectivas para aprender patrones a partir de conjuntos de datos al considerar múltiples entradas simultáneamente. En este estudio, las CNN se utilizan para analizar varios textos como un conjunto, permitiendo que el modelo identifique patrones comunes y relaciones contextuales entre ellos.

El enfoque de las CNN se basa en la aplicación de capas convolucionales que extraen características locales y globales del texto. Esto permite al modelo aprender no solo a partir de palabras individuales, sino también a entender cómo estas palabras interactúan dentro del contexto más amplio del texto. Al agrupar múltiples textos, las CNN pueden captar patrones recurrentes en el lenguaje que indican sentimientos específicos.

Este enfoque es particularmente útil en el análisis de sentimientos porque permite a las CNN generalizar mejor a partir de ejemplos diversos, mejorando así la precisión del modelo al clasificar emociones. Al final, este proceso contribuye a una comprensión más rica y matizada de los sentimientos expresados en los datos textuales.

En resumen, la combinación de Redes Neuronales, KMeans y CNN permite abordar el problema del análisis de sentimientos desde diferentes ángulos, aprovechando las fortalezas únicas de cada modelo. Esta diversidad en los enfoques contribuye a una evaluación más robusta y completa del conjunto de datos.

3 Conclusiones

En esta investigación, se llevó a cabo un análisis utilizando el algoritmo KMeans para agrupar datos basados en dos características: Feature2 y Feature3. Los resultados obtenidos con KMeans no fueron tan satisfactorios en comparación con los resultados de otros algoritmos de clasificación, como Naive Bayes (NB), Árboles de Decisión y Máquinas de Vectores de Soporte (SVM).

Además, se implementaron Redes Neuronales Convolucionales (CNN) y Redes Neuronales (NN), pero los resultados obtenidos con estos métodos tampoco fueron satisfactorios. A pesar de que KMeans proporciona una forma efectiva de agrupar datos, su rendimiento en este caso específico no alcanzó el nivel de precisión y efectividad observado con los otros métodos mencionados. Esto sugiere que, aunque KMeans, CNN y NN pueden ser útiles para ciertas aplicaciones, en este contexto particular, otros algoritmos ofrecen mejores resultados en términos de clasificación y agrupamiento.

4 Referencias

[1] Research on text sentiment analysis model and its application for mental health based on social media data. Journal of Educational Research and Policies, 2023. [2] Gunjan Ansari, Muskan Garg, and Chandni Saxena. Data augmentation for mental health classification on social media. In ICON, 2021. [3] Nafiz Al Asad, Md. Appel Mahmud Pranto, Sadia Afreen, and Md. Maynul Islam. Depression detection by analyzing social media posts of user. 2019 IEEE International Conference on Signal Processing, Information, Commu- nication "&" Systems (SPICSCON), pages 13–17, 2019. [4] Smita Ghosh. Depression detection using machine and deep learning models to assess mental health of social media users. Machine Learning Techniques and Data Science Trends, 2022. [5] Jina Kim, Daeun Lee, and Eunil Park. Machine learning for mental health in social media: Bibliometric study. Journal of Medical Internet Research, 23, 2020. [6] Priya Mathur, Amit Kumar Gupta, and Abhishek Dadhich. Mental health classification on social-media: Systematic review. Proceedings of the 4th International Conference on Information Management "&" Machine Intelligence, 2022. [7] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4):1093–1113, 2014. [8] Anja Thieme, Danielle Belgrave, and Gavin Doherty. Machine learning in mental health. ACM Transactions on Computer-Human Interaction (TOCHI), 27:1 – 53, 2020. [9] Mahesworo Langgeng Wicaksono, Rusdah Rusdah, and Diwi Apriana. Sen- timent analysis of mental health using k-nearest neighbors on social media twitter. Bit (Fakultas Teknologi Informasi Universitas Budi Luhur), 2022.