|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **FACULTAD DE ADMINISTRACION** |  |  |  | http://smpmanizales.blogspot.es/img/logo-universidad-nacional-manizales.gif |

**Análisis de Sentimientos en Ambientes con Comentarios aplicando Procesamiento de Lenguaje Natural (NLP)***Autor: Daniel Cortés Pérez, Profesor director: PhD. Ingeniería y Néstor Darío Duque Méndez*

**Universidad Nacional de Colombia**

*30/05/2024*

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Resumen**

El análisis de sentimientos, una subdisciplina del procesamiento de lenguaje natural (NLP), es fundamental para interpretar los millones de comentarios generados diariamente en plataformas digitales. Esta subdisciplina permite clasificar automáticamente las emociones expresadas en los textos como positivas, negativas o neutras, aportando información valiosa para la toma de decisiones en áreas como marketing, ventas y desarrollo de productos. En este trabajo se desarrolló un algoritmo de análisis de sentimientos aplicado a comentarios en línea, empleando técnicas avanzadas de NLP y aprendizaje automático. Se utilizaron bibliotecas de Python, como pandas, numpy y sklearn, para la manipulación y transformación de los datos, mientras que los modelos de clasificación más destacados fue la regresión logística, el cual fue entrenado y evaluado utilizando técnicas de validación cruzada. El modelo de regresión logística obtuvo un gran desempeño, alcanzando un F1 Score de 0.89 y un AUC de 0.91, lo que demostró su eficacia para la clasificación de sentimientos. Estos resultados validan la utilidad del análisis de sentimientos en la interpretación de grandes volúmenes de datos textuales, permitiendo a las empresas y organizaciones optimizar sus estrategias basadas en la retroalimentación de los usuarios. Además, el análisis de sentimientos ofrece un enfoque práctico y escalable para comprender las emociones del público, contribuyendo al desarrollo de mejores productos y servicios, así como a una toma de decisiones más informada y precisa.

**Palabras Clave:** Análisis de sentimientos, Procesamiento de lenguaje natural, NLP, Aprendizaje automático

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1. **Objetivos General**

Desarrollar un algoritmo basado en NLP para el análisis automático de sentimientos en comentarios en línea en diversos contextos y dominios.

1. **Objetivos Específicos**

* Revisar la literatura sobre análisis de sentimientos en ambientes con comentarios.
* Identificar y seleccionar las técnicas, algoritmos y herramientas adecuadas para el desarrollo de la solución.
* Diseñar y desarrollar el algoritmo de análisis de sentimientos.
* Evaluar el desempeño del algoritmo en un conjunto de datos de comentarios en línea.
* Aplicar el algoritmo a un conjunto de datos de comentarios en línea en un contexto específico.

1. **Introducción**

En la era digital actual, la cantidad de datos textuales generados por los usuarios es inmensa. Comentarios, reseñas y opiniones se publican constantemente en plataformas de redes sociales, sitios de reseñas y tiendas en línea, lo que crea un vasto volumen de información potencialmente valiosa para las empresas. Para interpretar estos datos de manera efectiva, el análisis de sentimientos se ha convertido en una herramienta esencial. Esta técnica permite determinar si un texto expresa una emoción positiva, negativa o neutra, y ayuda a las organizaciones a entender mejor las percepciones de sus clientes y adaptar sus estrategias en consecuencia.

Este proyecto tiene como objetivo desarrollar un modelo de análisis de sentimientos que permita clasificar los comentarios de manera eficiente utilizando algoritmos de aprendizaje automático. El enfoque propuesto combina técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático, lo que permite transformar datos textuales en información valiosa para la toma de decisiones. Utilizando bibliotecas de Python como pandas y sklearn, se implementaron modelos de clasificación que fueron optimizados y evaluados utilizando métricas de rendimiento para garantizar su eficacia. Los resultados subrayan la importancia del análisis de sentimientos en la era de la información.

**4. Metodología**

Este trabajo de grado adoptó una metodología sistemática para desarrollar y evaluar un algoritmo de análisis de sentimientos utilizando técnicas de procesamiento de lenguaje natural (NLP) y aprendizaje automático. El primer paso fue la recolección y preparación de datos, en la cual se utilizaron bibliotecas de Python como pandas para cargar, explorar y preparar los datos de comentarios en línea obtenidos de plataformas como sitios de reseñas y redes sociales. Esta fase incluyó la recolección de datos relevantes y su limpieza para eliminar elementos no deseados como caracteres especiales, enlaces, menciones y otros componentes irrelevantes.

El preprocesamiento de los datos se llevó a cabo eliminando ruido y normalizando el texto, convirtiendo todo a minúsculas y eliminando stopwords. Posteriormente, se aplicaron técnicas de lematización y tokenización utilizando la biblioteca nltk, lo que permitió que el texto fuera dividido en tokens y sus palabras fueran reducidas a su forma base, facilitando así la manipulación del contenido textual. Una vez preprocesados, los textos fueron transformados en vectores numéricos utilizando TF-IDF (Term Frequency-Inverse Document Frequency), que asigna a cada palabra un valor basado en su frecuencia dentro del documento y en relación con el conjunto completo de datos.

En la etapa de desarrollo del modelo, se implementaron algoritmos de clasificación como Regresión Logística y Naive Bayes utilizando la biblioteca sklearn. Los modelos fueron entrenados utilizando técnicas de validación cruzada para evaluar su rendimiento en diferentes subconjuntos del conjunto de datos. La evaluación del modelo se realizó mediante métricas clave como ROC AUC y F1 Score, que permitieron medir la precisión y efectividad de cada modelo en la clasificación de sentimientos. Finalmente, los resultados fueron visualizados a través de las bibliotecas matplotlib y seaborn, lo que facilitó la interpretación de los hallazgos y proporcionó una visión clara del comportamiento del algoritmo en la clasificación de sentimientos.

**5. Planteamiento**

En un mundo digitalizado, las empresas enfrentan el desafío constante de interpretar grandes volúmenes de datos textuales generados diariamente por los usuarios en plataformas en línea. Los comentarios, opiniones y reseñas de productos y servicios no solo contienen información sobre la satisfacción del cliente, sino también sobre sus emociones, expectativas y áreas de mejora. Sin embargo, el análisis manual de estos datos es prácticamente imposible debido a la magnitud y diversidad de las plataformas donde se generan. Esto plantea la necesidad de una solución automatizada que permita interpretar estos datos de manera eficiente y precisa.

El análisis de sentimientos proporciona una respuesta eficaz a este reto. Al utilizar técnicas de procesamiento de lenguaje natural (NLP), es posible clasificar automáticamente las emociones contenidas en los textos, lo que ayuda a las empresas a comprender mejor las percepciones de sus clientes. Este proyecto se propone desarrollar un algoritmo que aborde esta problemática, aplicando modelos de aprendizaje automático que permitan no solo clasificar los sentimientos de manera eficiente, sino también optimizar el proceso de toma de decisiones estratégicas basadas en datos. A través de la implementación de técnicas avanzadas de NLP, se espera que este trabajo contribuya al desarrollo de herramientas más precisas para el análisis de sentimientos en diversos contextos y dominios.

**6. Revisión del Trabajo**

Este trabajo revisa las investigaciones más relevantes en el campo del análisis de sentimientos. Una de las principales referencias fue el trabajo de **Pang y Lee (2008)**, que sentó las bases teóricas para la clasificación de sentimientos en textos, enfatizando la necesidad de distinguir entre hechos y opiniones. Este trabajo subraya la importancia de identificar las emociones contenidas en los textos y sugiere el uso de técnicas como la regresión logística para lograr una clasificación efectiva.

Por otro lado, **Liu (2012)** exploró el uso de técnicas avanzadas de procesamiento de lenguaje natural y aprendizaje automático para mejorar la precisión en la clasificación de sentimientos. Liu presentó una revisión exhaustiva de las diferentes técnicas disponibles, desde enfoques basados en léxico hasta el uso de modelos de aprendizaje supervisado, destacando la efectividad de los algoritmos de clasificación como Naive Bayes y la regresión logística.

En estudios más recientes, se ha demostrado el potencial de las redes neuronales y los modelos pre-entrenados como **BERT** para mejorar aún más la precisión en la clasificación de sentimientos. **Medhat et al. (2014)** llevaron a cabo una revisión exhaustiva de los algoritmos de análisis de sentimientos, explorando tanto métodos supervisados como no supervisados. Estos estudios subrayan la importancia de seleccionar cuidadosamente los algoritmos y técnicas de preprocesamiento para garantizar una clasificación precisa y robusta.

El uso de bibliotecas como **nltk**, **pandas** y **sklearn** ha sido esencial para implementar de manera efectiva los algoritmos en este trabajo. Estas herramientas proporcionaron funcionalidades robustas para la manipulación y transformación de datos, lo que permitió aplicar las técnicas de procesamiento de lenguaje natural de manera eficiente y efectiva en los comentarios en línea.

**7. Cuerpo del trabajo**

El cuerpo del trabajo se centró en la creación de un algoritmo de análisis de sentimientos aplicable a comentarios generados en línea. El desarrollo del algoritmo siguió un proceso sistemático que incluyó varias fases críticas. La primera fase fue la **recolección de datos**, en la que se recopilaron comentarios en línea provenientes de reseñas de películas y plataformas de redes sociales como Twitter. Estos datos fueron almacenados y organizados utilizando la biblioteca **pandas**, que permitió manipular grandes volúmenes de datos de manera eficiente.

Posteriormente, los datos pasaron por una fase de **preprocesamiento**, donde se eliminaron caracteres especiales, enlaces, menciones y stopwords. Además, el texto fue normalizado, y mediante técnicas de **lematización** y **tokenización**, se redujeron las palabras a su forma base, facilitando la manipulación y análisis de los textos. Una vez preprocesados, los datos fueron transformados en vectores numéricos mediante la técnica **TF-IDF**, lo que permitió representar los comentarios en una estructura numérica adecuada para los modelos de clasificación.

El siguiente paso fue el **desarrollo del modelo**, donde se probaron varios algoritmos de clasificación, entre ellos la **Regresión Logística** y **Naive Bayes**. Estos modelos fueron entrenados utilizando técnicas de validación cruzada para asegurar la robustez del algoritmo y evitar el sobreajuste. Los resultados obtenidos fueron evaluados mediante métricas como **ROC AUC** y **F1 Score**, que permitieron medir la precisión del algoritmo en la clasificación de sentimientos.

Finalmente, los resultados del análisis se visualizaron mediante herramientas como **seaborn** y **matplotlib**, lo que facilitó la interpretación de los hallazgos. Estas visualizaciones permitieron identificar patrones y tendencias en los comentarios, destacando los términos más utilizados en comentarios positivos y negativos, lo que contribuyó a una comprensión más profunda de los datos y la efectividad del algoritmo.

**8. Conclusiones**

**Importancia del Preprocesamiento de Datos:** El preprocesamiento adecuado de los datos, incluyendo la limpieza, tokenización y eliminación de ruido, es crucial para mejorar la precisión de los modelos de análisis de sentimientos. Sin un preprocesamiento riguroso, los resultados pueden ser significativamente menos precisos.

**Comparación de Algoritmos de Clasificación:** Al comparar diferentes algoritmos de clasificación, como la Regresión Logística y Naive Bayes, se encontró que la Regresión Logística ofreció una mejor precisión general en la identificación de sentimientos positivos, negativos y neutros.

**Relevancia de la Visualización de Datos:** La visualización de datos con herramientas como Seaborn y Matplotlib resultó esencial para identificar patrones y tendencias en los datos. Estas visualizaciones no solo facilitaron la comprensión de los resultados, sino que también mejoraron la comunicación de los hallazgos.

**Uso de Métricas de Evaluación:** Las métricas de evaluación como el F1 Score y el ROC AUC fueron fundamentales para medir la eficacia y precisión de los modelos de clasificación. Estas métricas proporcionaron una visión clara de las fortalezas y debilidades de cada modelo, permitiendo ajustes y optimizaciones adicionales.

**Aplicabilidad en Distintos Contextos:** El análisis de sentimientos demostró ser una herramienta poderosa y versátil aplicable en diversos contextos, desde la retroalimentación de reseñas hasta el monitoreo de redes sociales. La capacidad de entender y analizar opiniones en tiempo real ofrece una ventaja competitiva significativa.

**Desafíos en la Recolección de Datos:** La recopilación de datos de fuentes como reseñas de películas y tweets presentó desafíos en términos de acceso y limpieza. La implementación de técnicas de scraping y el uso de APIs específicas fueron necesarios para obtener datos relevantes y de calidad.

**Impacto del Tamaño del Conjunto de Datos:** El tamaño y la diversidad del conjunto de datos impactan directamente en la precisión del modelo. Conjuntos de datos más grandes y variados tienden a proporcionar modelos más robustos y generalizables.

**8. Referencias Bibliográficas**

# Bibliografía

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. 1798-1828.

Bird, S. K. (2009). Natural Language Processing with Python. *O'Reilly Media.*

Bo Pang, L. L. (202). Thumbs up? Sentiment Classification using Machine Learning Techniques. *Conference on Empirical Methods in Natural Language Processing*, 59-67.

González Guerra, C. (2020). *Paridad de herramientas para el análisis de sentimientos sobre un CORPUS en español generado desde twitter.* Obtenido de Repositorio Institucional Academico: http://repositorio.unab.cl/xmlui/handle/ria/15339

H. Cunningham, D. M. (2002). GATE: A framework and graphical development environment for robust NLP tools and applications. *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics*.

Hunter, J. (2010). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering.*

Hutto, C. &. (2014). *VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.* USA: Vol. 8 No. 1 (2014): Eighth International AAAI Conference on Weblogs and Social Media.

Lee, B. P. (08 de 07 de 2008). *Opinion Mining and Sentiment Analysis.* USA: Foundations and Trends® in Information Retrieval: Vol. 2.

Lei Zhang, S. W. (2018). Deep learning for sentiment analysis: A survey.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* USA: Springer Cham.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies, 5.* USA.

McKinney, W. (2010). Data Structures for Statistical Computing in Python. *Proceedings of the 9th Python in Science Conference.*

Md Shad Akhtar, A. K. (2016). A Hybrid Deep Learning Architecture for Sentiment Analysis. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 482-493.

Medhat, W. H. (2014). Sentiment analysis algorithms and applications: A survey. . *Ain Shams Engineering Journal, 5(4),* , 1093-1113.

Mikolov, T. S. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems, 26.*

Oliphant, T. (2006). A guide to NumPy. . *Trelgol Publishing.*

Pedregosa, F. e. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research.*

Saif M. Mohammad, S. K. (Atlanta, USA). NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. *Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)*.

Ungar., M. A.-M. (2017, Vancouver, Canada). EmoNet: Fine-Grained Emotion Detection with Gated Recurrent Neural Networks. *In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728.

Vidal Castro, C. L. (2017). *Comparación de rendimiento de técnicas de aprendizaje automático para análisis de afecto sobre textos en español.* Obtenido de SIBUBB: http://repobib.ubiobio.cl/jspui/handle/123456789/1772

Waskom., M. (2021). Seaborn: Statistical Data Visualization. *Journal of Open-Source Software.*