Minería de Textos para Clasificación y Análisis de Sentimientos de Relatos Personales

Soledad Ruiz Diaz[1], Miguel Mendez-Garabetti[2]

[1] Departamento de Posgrados, Universidad CAECE, Mar del Plata, Argentina.
[2] Free and Open Source Software/Hardware Research Laboratory (FOSSHLab), Argentina.
soledad.ruizdiaz@atlantida.edu.ar, mendezgarabetti@fosshlab.org

Resumen. El presente trabajo busca implementar herramientas y técnicas de aprendizaje automático para automatizar el proceso de análisis de los relatos recopilados en tres ediciones del libro "Matilda y las Mujeres en Ingeniería en América Latina", con el fin de identificar factores que influyen en la elección y ejercicio de la carrera de ingeniería por parte de las mujeres. La metodología seguirá los lineamientos propuestos para un proceso de Descubrimiento de Conocimiento en Textos (KDT). El trabajo se dividirá en varias etapas: comprensión del dominio de aplicación, extracción de datos, limpieza, procesamiento y transformación de datos, y desarrollo del modelo. En la actualidad, el proyecto se encuentra en la fase de construcción del corpus y supresión de patrones de información no significativos. Luego se realizará una tokenización del texto para entender las características del mismo y se evaluará la técnica más adecuada para cuantificar el set de palabras presentes en el corpus. Se construirá un modelo de aprendizaje automático supervisado para predecir la temática principal del relato y se analizará el sentimiento del mismo en función de su temática. El análisis de sentimientos se realizará considerando el sentimiento como la suma de los sentimientos de cada una de las palabras que lo conforman.

Palabras Clave: Minería de Textos, Aprendizaje Automático, Clasificación, Análisis de Sentimientos

1 Introducción

La minería de textos se define como una técnica de análisis de documentos escritos en lenguaje natural para extraer información de ellos, clasificarlos o identificar patrones [8]. Dentro de la minería de textos, la minería de opiniones se enfoca en evaluar las opiniones, sentimientos y emociones expresadas en forma textual [16]. La identificación del sentimiento de un autor en un texto puede ser interpretada como una categorización o clasificación del texto según sus características [13].

En este trabajo se busca analizar automáticamente relatos personales para identificar características que permitan categorizarlos y estimar sus sentimientos. Para esto, se utilizarán los relatos recopilados en las cuatro ediciones del libro 'Matilda y las Mujeres en Ingeniería en América Latina', publicados por el Consejo Federal de Decanos de

Ingeniería de la República Argentina (CONFEDI) [6] y el Latin American and Caribbean Consortium of Engineering Institutions (LACCEI) [15]. Estos libros reúnen las experiencias y vivencias de ingenieras de siete países con el fin de visibilizar el rol de la mujer en la ingeniería.

Para automatizar el proceso de análisis de los relatos, se utilizarán herramientas y técnicas de aprendizaje automático. Esto facilitará el análisis de las anécdotas o experiencias expresadas en los libros en busca de factores que influyen en la elección de la carrera y/o el ejercicio de la profesión, ya sea de forma positiva o negativa, para poder cuantificar el impacto de cada factor.

2 Metodología y Resultados

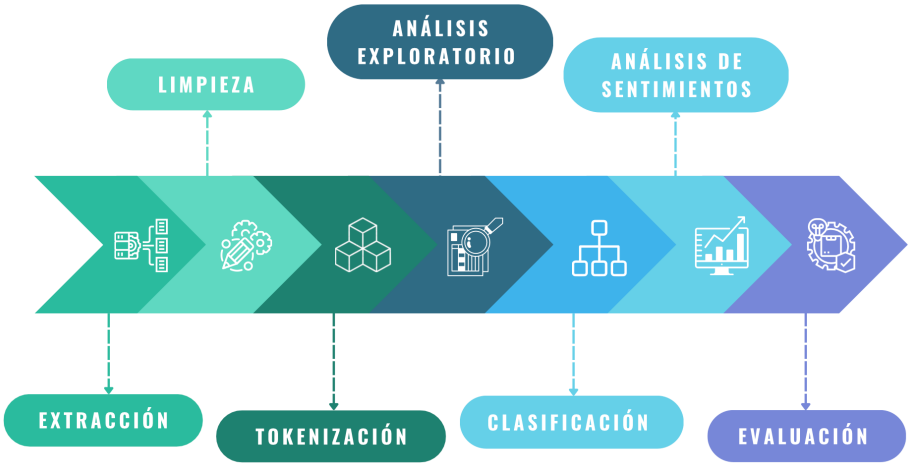
En este trabajo, se adopta la metodología de Descubrimiento de Conocimiento en Textos (KDT) propuesta por Gómez, Puertas y Luaces [10]. Esta metodología se basa en el proceso de Descubrimiento de Conocimiento en Datos (KDD), pero se enfoca específicamente en el análisis de textos para extraer conocimiento relevante.

Según Gómez, Puertas y Luaces, el proceso de KDT sigue fases similares al proceso de KDD, pero se adapta para abordar las particularidades de los datos de texto. Esto implica incluir etapas específicas para el preprocesamiento y la transformación de los datos textuales.

La etapa de preprocesamiento es esencial en el análisis de textos, ya que los datos de texto suelen ser no estructurados y requieren de técnicas de procesamiento del lenguaje natural (NLP). Autores como Bird, Klein y Loper [2] respaldan la importancia del preprocesamiento en el análisis de texto, ya que implica tareas como la tokenización, lematización y etiquetado gramatical, que permiten una mejor representación y comprensión de los datos.

Además del preprocesamiento, la metodología KDT también destaca la importancia de la selección de textos relevantes y la extracción de características específicas. Estas etapas se basan en técnicas de minería de textos y análisis de contenido. Al respecto, Liu [16] menciona la relevancia de la extracción de características en la minería de opiniones, lo cual implica identificar patrones y atributos relevantes en los textos para comprender las opiniones y sentimientos expresados. Con base en lo anterior, a continuación se resumen de manera gráfica las tareas que se realizarán en cada fase de este trabajo teniendo en cuenta el marco establecido por la metodología KDT, tal como se muestra en la Figura 1.

Fig. 1. Diagrama de Flujo del proceso KDT aplicado a la clasificación y análisis de relatos personales.



2.1 Comprensión del dominio de Aplicación

En esta etapa, se busca comprender las dificultades presentes en el procesamiento del lenguaje natural y análisis de sentimientos, para lo cual se puede recurrir a investigaciones previas que establezcan el estado del arte sobre la temática [1][4][5][7][12][14].

2.2 Extracción, Limpieza y Procesamiento de Datos

Para procesar los relatos se deberá crear el corpus del trabajo; el cual es la estructura que contiene el texto que se va a utilizar. La carga del corpus requerirá, de la extracción manual de cada relato y su posterior almacenamiento; posteriormente, por medio de la librería Natural Language Toolkit (NLTK) se procederá a la carga del mismo en el entorno de desarrollo [2] para iniciar con su preprocesamiento. La limpieza de texto es un paso crítico en el análisis de sentimientos [18]. Este proceso incluirá la eliminación de información no significativa para que el texto resultante sea sencillo de analizar y se reduzca el ruido en los datos. Luego se procederá a la tokenización del texto. Ésta tarea es esencial, ya que permite dividir el texto en unidades más pequeñas y significativas para su análisis. Existen diferentes métodos de tokenización; para el trabajo en particular, se dividirá el texto por palabras [9]. Finalmente, para el desarrollo del modelo se necesita entender las características del texto por lo que se realizara un análisis exploratorio del mismo [17].

2.3 Desarrollo del Modelo y Evaluación

La clasificación de textos, requerirá crear una representación numérica del mismo. Las técnicas más comunes para la cuantificación de palabras son la bolsa de palabras y la matriz término-documento [3]. A partir de la representación del texto, se construirá un modelo de aprendizaje automático supervisado para predecir la temática principal del relato. Luego, se deberá realizar el análisis de sentimientos del relato en función de la misma. Una de las formas de analizar el sentimiento de un texto es considerarlo como la suma de los sentimientos de cada una de las palabras que lo forman; por lo que se deberá crear un diccionario léxico en el que se le asocie a cada palabra una polaridad. Esta técnica de análisis de sentimientos se conoce como enfoque léxico; el cual se basa en la idea que las palabras tienen una carga emocional positiva o negativa y que el sentimiento de un texto puede ser determinado por la suma de sus polaridades [11]. Finalmente, se evaluarán las métricas resultantes con el objetivo de optimizar el modelo.

3 Conclusiones y Trabajo Futuro

El proyecto tiene como objetivo identificar los factores que influyen en la elección y ejercicio de la carrera de ingeniería por parte de las mujeres en América Latina. El mismo se encuentra en la fase de construcción del corpus y supresión de patrones de información no significativos.

Este proyecto tiene el potencial de proporcionar información valiosa sobre los factores que influyen en la elección de la carrera de ingeniería por parte de las mujeres en América Latina y puede ayudar a diseñar políticas y/o programas para fomentar la participación de las mujeres en esta área.

Referencias

1. Amor, M. N., Monge, A., Talamé, M. L., & Cardoso, A. C. (2020). Clasificación de sentimientos en opiniones de una red social basada en dimensiones emocionales. ReDDI: Revista Digital del Departamento de Ingeniería, 5(1), 1–13.
2. Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media, Inc.
3. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching word vectors with subword information.
4. Castillo Landínez, S. P. ., & Caicedo Rodríguez, P. E. . (2019). Análisis de sentimientos, una herramienta para valorar la actitud del estudiante frente a un curso. Encuentro Internacional De Educación En Ingeniería. Recuperado a partir de https://acofipapers.org/index.php/eiei/article/view/141
5. Colón-Ruiz, C., Segura-Bedmar, I., Martínez, P. (2019). Análisis de sentimiento en el dominio salud: analizando comentarios sobre fármacos. Procesamiento del Lenguaje Natural, [S.l.], v. 63, 15-22. ISSN 1989-7553.
6. CONFEDI. (s. f.). CONFEDI. Recuperado 4 de junio de 2022, de https://confedi.org.ar/

7. Dasgupta, S., Singh, K., Saini, S., & Kumar, P. (2022). Sentiment Analysis Techniques and Challenges: A Review. IOP Conference Series: Materials Science and Engineering, 1369(1), 012026.
8. Feldman, R., & Sanger, J. (2007). The text mining handbook: Advanced approaches in analyzing unstructured data. Cambridge University Press.
9. García-Silva, A., & Díaz-Galiano, M. C. (2021). Tokenization of texts in natural language processing: A systematic review. Mathematics, 9(20), 2344.
10. Gómez, F., Puertas, E., & Luaces, Ó. (2018). Metodología de descubrimiento de conocimiento en textos (KDT): una revisión sistemática. Procesamiento del Lenguaje Natural, 61, 77-84.
11. Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177).
12. Huang, J., Zhang, Q., Chen, J., & Fu, H. (2021). Study on Corpus Building and Analysis Methods of College English Teaching Materials Based on Python. Journal of Physics: Conference Series, 1928(1), 012035.
13. Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Eighth International Conference on Weblogs and Social Media. AAAI Press.
14. Justicia de la Torre, M.C. (2017). Nuevas técnicas de minería de textos: Aplicaciones. Universidad de Granada, Granada. http://hdl.handle.net/10481/46975
15. LACCEI. (s. f.). LACCEI. Recuperado 4 de junio de 2022, de https://laccei.org/
16. Liu, B. (2012). Sentiment analysis and opinion mining. Morgan & Claypool Publishers.
17. Pandey, M., & Shukla, A. (2021). Exploratory Text Analysis: A Comprehensive Review of Techniques and Tools. Journal of Information Technology Management, 32(1), 1-18.
18. Singh, N., & Singh, R. (2021). Pre-processing and cleaning of text data for sentiment analysis: A comprehensive review. Journal of Ambient Intelligence and Humanized Computing, 12(9), 9799-9816.