Polarized analysis of notes using Artificial Intelligence

Análisis polarizado de notas mediante el uso de Inteligencia Artificial

Hermenegildo-Domínguez, Cesar*[a], Reyes-Nava, Adriana [b] and López-González, Erika [c]

[a] ROR TecNM: Tecnológico de Estudios Superiores de Jocotitlán• LNQ-6377-2024 • 0009-0000-2675-4570• 2073873
[b] ROR TecNM: Tecnológico de Estudios Superiores de Jocotitlán • LNP-9046-2024 • 0000-0002-4440-909X • 786639
[c] ROR TecNM: Tecnológico de Estudios Superiores de Jocotitlán • LNQ-5524-2024 • 0000-0001-7279-5111

CONAHCYT classification:

DOI: https://doi.org/10.35429/H.2024.12.41.55

Area: Engineering
Field: Technological sciences
Disciplina: Computer technology
Subdisciplina: Artificial intelligence

Key Handbooks

The main contributions of this research to science and technology focus on the implementation of advanced natural language processing tools and the development of an automated system for the polarized analysis of news stories. These technologies combine to create a platform capable of effectively classifying media content according to its polarity (positive, negative or neutral), which contributes to a better understanding of the media's influence on public opinion. In addition, the creation of an efficient web application allows real-time visualization of analyzed data, facilitating access to relevant information for journalists, researchers and analysts. To apply this knowledge to the generation of universal knowledge, it is essential to understand the principles of natural language processing and sentiment analysis, as well as key entity extraction techniques. These aspects are essential to design systems that can analyze large volumes of text and provide clear and accessible results through interactive interfaces. The authors of this project come from various institutions, including technological and public universities, highlighting the academic collaboration in the search for innovative solutions. The keywords most commonly used in the development of this research include “polarized analysis”, “natural language processing”, “sentiment analysis” and “media bias”.

Citation: Hermenegildo-Domínguez, Cesar, Reyes-Nava, Adriana and López-González, Erika. 2024. Polarized Analysis of Notes Using Artificial Intelligence. 41-55. ECORFAN.

* 🖂 [2020150480477@tesjo.edu.mx]

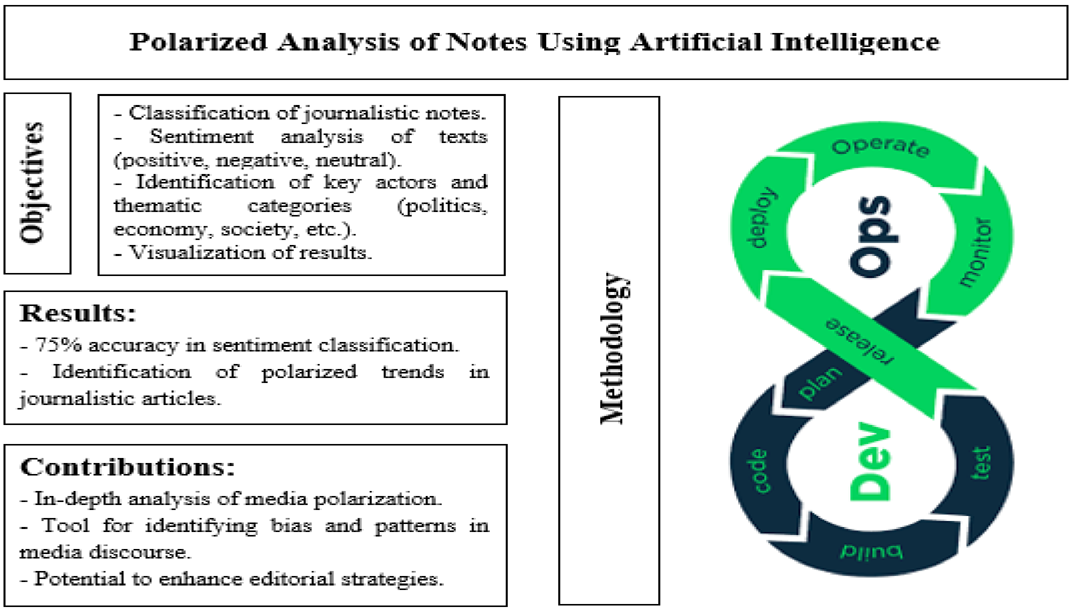
Handbook shelf URL: https://www.ecorfan.org/handbooks.php

Peer Review under the responsibility of the Scientific Committee MARVID®- in contribution to the scientific, technological and innovation Peer Review Process by training Human Resources for the continuity in the Critical Analysis of International Research.

RENIECYT
Registro Nacional de Instituciones y Empresas Científicas y Tecnológicas
1702902 CONAHCYT

Abstract

In a polarized media landscape, critical news analysis is crucial. This project develops an automated system for analyzing journalistic content based on polarity positive, negative, or neutral using advanced natural language processing (NLP) and machine learning techniques. By focusing on emotions and critical stances in texts, the system highlights key actors and topics in news articles. It employs sentiment analysis models like BERT for accurate polarity detection and extracts important figures from the content. The findings are presented through graphical visualizations to illustrate trends in the analyzed texts. Users can upload Excel files with news articles, utilizing a Flask-based web application for real-time analysis. This tool aims to assist researchers, journalists, and analysts in studying media bias and news coverage trends.



Polarized Analysis of Notes Using Artificial Intelligence

Objectives

- Classification of journalistic notes.
- Sentiment analysis of texts (positive, negative, neutral).
- Identification of key actors and thematic categories (politics, economy, society, etc.).
- Visualization of results.

Results:

- 75% accuracy in sentiment classification.
- Identification of polarized trends in journalistic articles.

Contributions:

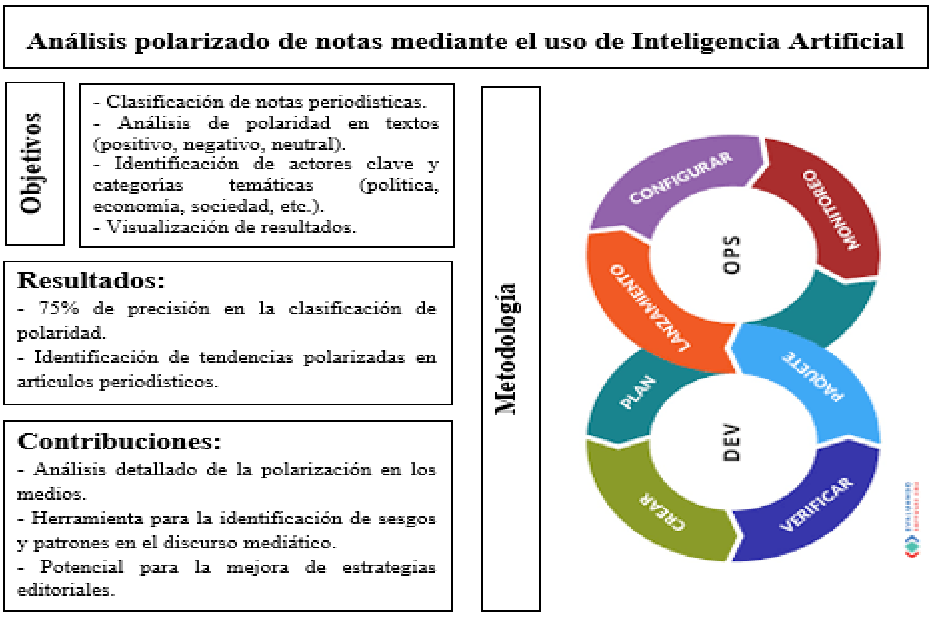
- In-depth analysis of media polarization.
- Tool for identifying bias and patterns in media discourse.
- Potential to enhance editorial strategies.

Methodology

Polarized Analysis, Natural Language Processing, Machine Learning

Resumen

En un panorama mediático polarizado, el análisis crítico de noticias es crucial. Este proyecto desarrolla un sistema automatizado para analizar el contenido periodístico basado en la polaridad positiva, negativa o neutral utilizando técnicas avanzadas de procesamiento de lenguaje natural (NLP) y aprendizaje automático. Al centrarse en las emociones y posturas críticas en los textos, el sistema destaca actores clave y temas en los artículos de noticias. Emplea modelos de análisis de sentimientos como BERT para una detección precisa de la polaridad y extrae Figures importantes del contenido. Los resultados se presentan a través de visualizaciones gráficas que ilustran tendencias en los textos analizados. Los usuarios pueden cargar archivos Excel con artículos de noticias, utilizando una aplicación web basada en Flask para análisis en tiempo real. Esta herramienta tiene como objetivo ayudar a investigadores, periodistas y analistas a estudiar el sesgo mediático y las tendencias en la cobertura informativa.



Análisis polarizado de notas mediante el uso de Inteligencia Artificial

Objetivos

- Clasificación de notas periodisticas.
- Análisis de polaridad en textos (positivo, negativo, neutral).
- Identificación de actores clave y categorias temáticas (politica, economia, sociedad, etc.).
- Visualización de resultados.

Resultados:

- 75% de precisión en la clasificación de polaridad.
- Identificación de tendencias polarizadas en articulos periodisticos.

Contribuciones:

- Análisis detallado de la polarización en los medios.
- Herramienta para la identificación de sesgos y patrones en el discurso mediático.
- Potencial para la mejora de estrategias editoriales.

Metodología

Análisis Polarizado, Procesamiento de Lenguaje Natural, Aprendizaje Automático

Introduction

In the current media landscape, journalism plays a central role in shaping opinions and disseminating information about relevant events at local, national, and international levels. However, with the expansion of social media and the fragmentation of the media ecosystem, a growing polarization in the presentation of facts has become apparent. Media outlets do not just inform; they often influence public perception, shaping opinions and attitudes toward certain topics or individuals. This media polarization, driven by editorial agendas, political ideologies, or commercial interests, affects the objectivity and impartiality of journalism, generating distrust among the public and complicating the understanding of facts.

In this context, there is an evident need to develop technological tools that allow for an objective analysis of journalistic content to identify potential biases and polarized trends. The polarized analysis of journalistic notes is essential for studying how the media presents certain themes, events, or actors and how they emotionally influence their audience. This is crucial not only for journalists and researchers but also for organizations seeking to better understand the dynamics of the media and its impact on public opinion.

The main objective of this project is to create an automated system that conducts an in-depth analysis of journalistic notes, classifying them according to their polarity: positive, negative, or neutral. This analysis will enable the detection of the emotional orientation of the content, determining whether the media tends to describe a topic or actor favorably, unfavorably, or impartially. Furthermore, the project will focus on extracting key actors, identifying the most mentioned figures or entities in the text, providing additional context to understand who the protagonists in the news are and how they are represented. The ultimate goal is to offer journalists, analysts, and researchers a tool that allows them to objectively evaluate media coverage and its influence on audience perception.

This system is characterized by integrating various advanced technologies for natural language processing (NLP) and machine learning. First, pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers), widely used in sentiment analysis, will be implemented to determine the polarity of journalistic texts. These models are designed to identify nuances in language and classify texts according to the emotional tone they convey.

Additionally, the project will include a key actor extraction module, utilizing named entity recognition (NER) techniques to identify people, organizations, and other relevant actors mentioned in the notes. This will help visualize not only the polarity of the note but also whom or what the coverage is directed towards.

An additional feature of the system will be the graphical visualization of results, allowing users to observe patterns and trends in media content. The visualizations will include graphs showing the overall polarity of the notes and its evolution over time, as well as graphs representing the frequency with which certain key factors are mentioned in the media. The system will also support the uploading and processing of Excel files, facilitating the integration of large volumes of journalistic data for analysis. The platform will be available through a web application developed in Flask, providing an intuitive interface where users can upload files, run the analysis, and visualize results in real time.

The central problem this project aims to address is the lack of accessible tools that allow journalists, researchers, and organizations to conduct a systematic and objective analysis of polarization in the news. In a context where media content can significantly influence public opinion, it is essential to have means to identify potential biases in the coverage of relevant events and actors. Currently, there are not enough automated solutions capable of processing large amounts of journalistic data and providing detailed analysis of the emotional orientation of notes and the individuals or entities involved. This gap limits users' ability to understand how the media affects public perception and hinders the identification of polarized patterns in published information.

This project seeks to fill that need by providing a platform that not only analyzes the polarity of journalistic texts but also extracts key actors and offers clear visualizations of the results, enabling users to detect media trends and their possible impact on public opinion. With this tool, it is expected to offer an efficient and scalable solution for the polarized analysis of news.

The hypothesis guiding this project is that, through the combination of advanced natural language processing techniques and sentiment analysis models, it is possible to accurately identify the polarity in journalistic notes and extract key actors from the content, thus revealing patterns in media coverage. It is anticipated that the analysis will provide evidence of how certain media tend to favor or harm specific actors or topics, demonstrating informational bias that influences public perception. Furthermore, it is expected that the detected polarization in the notes will vary according to the involved actors or discussed topics, providing a clearer view of the dynamics of journalistic coverage.

This tool could significantly contribute to research on media bias, helping users identify patterns of favoritism or criticism towards specific actors, and enhancing transparency in how events and figures are covered in the media. Additionally, its potential for real-time and large-scale use makes it a valuable resource for ongoing and detailed analysis of media content.

This work is organized into five sections:

- Section 1. Introduction: Presents the context of the polarized analysis of journalistic notes, the problem, the objectives, and the hypothesis of the project.
- Section 2: Theoretical Framework: Explores the conceptual and technical foundations, including natural language processing, sentiment analysis, and key entity extraction, with a review of previous works.
- Section 3. Development: Describes the implementation process of the system, the tools used, and the workflow from data loading to result visualization.
- Section 4. Results: Presents the analyses conducted on journalistic notes, including sentiment analysis to determine their polarity (positive, negative, or neutral), categorization of the notes into themes such as politics, economy, sports, among others, detection of key actors mentioned in the texts, and identification of the journalistic genre to which the notes belong (news, editorial, chronicle, etc.).
- Section 5. Conclusions: Discusses the findings, evaluates the effectiveness of the system, and suggests possible improvements for future research.

2. Theoretical Framework

2.1. Natural Language Processing (NLP)

Human language is full of ambiguities that make it incredibly difficult to write software that accurately determines the intended meaning of text or voice data. Homonyms, homophones, sarcasm, idioms, metaphors, grammatical exceptions, and variations in sentence structure are just some of the irregularities of human language that people take years to learn.

Natural Language Processing (NLP) is the driving force behind many of the technologies we use in our daily lives, from virtual assistants like Siri and Alexa to language translation tools and the increasing accuracy of predictive text. It focuses on the interaction between computers and human language, assuming the ability of a computer system to analyze, interpret, and generate meaningful and useful human language. This can come in the form of text or audio, and once audio is transcribed, both types of data undergo a common analysis (ISO, n.d.).
According to IAAR (n.d.), the common elements of any standard architecture for a Natural Language Processing system are:

- Speech Recognition: Converting spoken words into a set of words.
- Language Understanding: Generating meaning for the spoken words, which will be used by the next element (dialogue management).
- Dialogue Management: Coordinating and keeping all parts of the system and users connected and linking with other systems.
- Communication with External Systems: Such as expert systems, database systems, or other computer applications.
- Response Generation: Establishing the message that the system should deliver.
- Voice Output: Using different techniques to produce the message from the system.

Natural Language Processing algorithms often rely on machine learning algorithms. Instead of manually coding large sets of rules, NLP can depend on machine learning to automatically learn these rules by analyzing a set of examples and making statistical inferences. Generally, the more data analyzed, the more accurate the model will be.

2.2 Sentiment Analysis Models

Sentiment analysis is a key technique within Natural Language Processing (NLP) that allows for the evaluation of emotions and opinions expressed in a text. It is used in various fields such as marketing, customer service, reputation management, and competitive analysis. There are different approaches and models for conducting this analysis:

- Lexicon-Based Models: These models use dictionaries of words with assigned sentiment values. Tools like AFINN or SentiWordNet count the words related to emotions in a text and determine whether the sentiment is positive, negative, or neutral, based on the frequency of those words.
- Machine Learning-Based Models: These models train algorithms with large volumes of labeled data. Notable among them are Naive Bayes and Support Vector Machines (SVM) for binary or multiclass classifications (positive, negative, neutral). More complex neural networks, such as Recurrent Neural Networks (RNN) or Convolutional Neural Networks (CNN), are also used to capture more complex patterns and dependencies in texts.
- Aspect-Based Sentiment Analysis (ABSA): This type of analysis focuses on evaluating sentiments regarding specific aspects of a product or service. For example, a company might analyze how users perceive the interface of its application without considering other factors.
- Emotion Detection: This technique goes beyond polarity (positive/negative) and attempts to identify more specific emotions such as frustration, enthusiasm, or surprise, which is useful in more detailed analyses, such as those related to customer service.

Sentiment analysis models are widely used in social media monitoring, to improve customer service, product development, and competitive analysis, providing valuable insights that can enhance business decision-making.

2.3 Text Classification

Text classification is based on the ability to automatically assign labels or categories to documents using Natural Language Processing (NLP) algorithms. This task is crucial in analyzing large volumes of text to extract valuable information and conduct efficient analyses.

There are two main approaches to text classification:

- Supervised Classification: This approach requires a labeled dataset, where machine learning models learn from previous examples to assign labels to new texts. It uses algorithms such as Naive Bayes, Support Vector Machines (SVM), and neural networks. For instance, Naive Bayes is effective for text classification problems because it assumes that the features are independent of each other, simplifying the probabilistic calculation (Jurafsky & Martin, 2021).
- Unsupervised Classification: This approach does not require labeled data and employs clustering techniques to organize texts into groups based on their characteristics. It is useful when labeled data is not available and patterns in the texts need to be discovered (Manning et al., 2008).

2.4 Actor Detection

Actor detection in Natural Language Processing (NLP) is a technique used to identify and extract key entities, such as people, organizations, or places, within a text. This task is commonly used in various fields, including media analysis, social studies, and commercial applications.

One of the most common approaches to detecting actors is the use of Named Entity Recognition (NER) models, which identify proper names and other types of entities within a text. Tools like SpaCy and Transformers are widely used for this task, where texts are processed, and entities are automatically tagged, which can be crucial for analyzing large volumes of data. These systems can use pre-trained models or be customized to detect specific entities within a particular domain (Osaba et al., 2021; Macêdo et al., 2021).

Another relevant approach is the use of topic modeling, such as the Latent Dirichlet Allocation (LDA) technique, which has been applied in research like the analysis of tango lyrics, where recurring themes related to emotions or images of the city were detected (Rosati, 2022). Although LDA is not exclusively focused on actor detection, it allows for the extraction of thematic patterns that may include mentions of key figures in the texts.

These models enable scalable, replicable, and efficient text analysis, which is especially useful for studies in social sciences and in the analysis of news or journalistic notes, where identifying relevant actors plays an important role (Márquez, 2021).

2.5 Categorization of Journalistic Genres

The categorization of journalistic genres has been a widely debated topic in the field of communication. Traditionally, genres are divided into informative, opinion, and interpretive categories. Informative genres, such as news and reports, focus on the objective transmission of facts, while opinion genres, such as editorials or columns, include the subjective perspective of the journalist. Interpretive genres, such as chronicles, combine narratives of events with personal analysis.

Currently, new classification proposals are being discussed due to the evolution of journalism and the influence of digital media, which have promoted the emergence of mixed genres. These new formats require a more flexible typology that better reflects the complexity of the current media landscape (Fernández Parratt, 2012).

2.6. DevOps Methodology

DevOps emerges as a philosophy that combines development (Development) and operations (Operations) practices. This approach seeks to integrate development and operations teams to improve efficiency and speed of software delivery, eliminating silos and promoting continuous collaboration. DevOps teams work throughout the entire software lifecycle, from planning and development to deployment and operation, using automation and monitoring tools to achieve continuous integration (CI) and continuous delivery (CD).

The fundamental principles of the DevOps methodology are as follows:

- Continuous Integration (CI): Involves the frequent integration of code into a shared repository, supported by automated tests to ensure software quality.
- Continuous Deployment (CD): Enables the rapid and reliable release of new software versions to production, ensuring that changes are implemented with lower risk.
- Infrastructure as Code (IaC): ConFiguretion and management of infrastructure are done through scripts and automated definitions, facilitating deployment in cloud environments.
- Continuous Feedback: Constant feedback between teams and users improves software iterations, which is key to the agile development cycle. (Atlassian, 2023)

3. Development

In the following section, the detailed steps taken to conduct the polarized analysis of journalistic notes are presented. This process includes data collection, text preparation and cleaning, as well as the implementation of natural language processing and sentiment analysis models. Through these stages, the aim is to classify and understand the polarity of the analyzed content, thus providing a solid foundation for the study's conclusions.

3.1 Data Collection

The data collection was carried out through a meticulous process aimed at ensuring the quality and relevance of the journalistic notes selected for the polarized analysis. Initially, extensive research was conducted to identify the most appropriate sources of information. This included reviewing digital platforms, media archives, and databases containing journalistic articles on various topics and formats.

The selection prioritized reliable and recognized sources, such as reputable newspapers and magazines, as well as informative websites specializing in political and social analysis. To ensure a broad perspective, both local and national media outlets were included, covering a diverse range of opinions and approaches to the topics addressed. The time frame for data collection was set to the last two years, allowing the analysis to focus on recent events and situations that have impacted the public sphere. This time frame not only provides an updated view of trends in media coverage but also facilitates the analysis of how media reflects and contributes to the construction of public opinion during critical moments.

In terms of content, a diversified approach was adopted, selecting notes from different journalistic genres, including reports, opinion columns, and news articles on politics, economics, and culture. This variety not only enriches the analysis by incorporating different writing styles and narrative approaches but also allows for capturing a broader range of sentiments and perspectives on the topics addressed.

Once the documents were gathered, a preliminary analysis was conducted to evaluate the relevance of the selected texts. This process included reviewing the topics covered and identifying relevant actors mentioned in the notes. Only those notes that showed a clear relationship with the study's objectives were included, thus ensuring that the dataset was representative and suitable for subsequent analysis. This careful approach to data collection laid the groundwork for a deeper and more meaningful analysis of the polarity in journalistic content, allowing for a better understanding of how media influences public perception of the events and actors involved.

Box 1

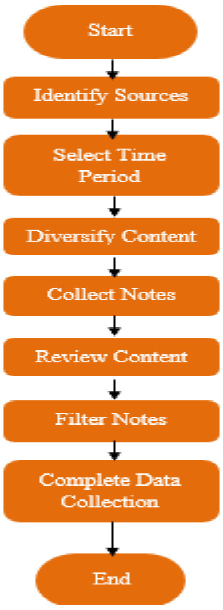


Figure 1

Flowchart of the Data Collection Process for Polarized Analysis

Source: Own Work

3.2 Data Preparation

3.2.1 Text Cleaning

Text cleaning is a critical step in the data processing of text analysis projects, as it helps ensure that the data is consistent, clear, and relevant for subsequent analysis. This process involves various techniques that transform raw data into a form suitable for analysis. Below, each step performed during the text cleaning process is described in detail:

3.2.1.1 Data Loading

The first step in text cleaning is loading the data from the original sources. In this case, the data was collected from CSV files containing journalistic notes. Using libraries like pandas, the files are imported into a DataFrame, allowing for easy and efficient manipulation of the data. Upon reviewing the structure of the data, relevant columns are identified, and initial issues such as missing data or duplicate records are detected.

3.2.1.2 Noise Removal

Once the data is loaded, the next step is to remove noise from the text. This noise includes special characters, numbers, and other elements that do not contribute to the understanding of the content. Using regular expressions, non-alphanumeric characters are eliminated. This process is essential to ensure that the analyzed text focuses solely on meaningful words.

3.2.1.3 Filtering Irrelevant Content

In this step, a deeper filtering of the content is carried out. Only the most relevant sections of the notes are selected, such as the title and the body of the text, while parts that do not add analytical value, such as reader comments or advertisements, are removed. This filtering process ensures that the analysis focuses on the most pertinent and valuable information.

3.2.1.4 Duplicate Detection

The detection and elimination of duplicates is essential to ensure data integrity. Duplicate records can distort the analysis results, so a procedure is implemented to identify and remove entries that are identical in content. This guarantees that each journalistic note is analyzed uniquely, avoiding biases in the results.

3.2.1.5 Text Normalization

Text normalization involves several processes that transform the content into a standard form. One of the first steps in this stage is converting all text to lowercase, which helps avoid discrepancies between similar words (for example, "Política" and "política"). Additionally, the removal of Stop Words is performed, which are common words that do not add meaning in the analysis (such as "and," "the," "of"). This removal is crucial to highlight the words that truly contain relevant information.

3.2.1.6 Lemmatization

Once the text has been normalized, lemmatization is performed. This process reduces words to their base form or lemma, grouping different variants of a word under a single representation. For example, the words "run," "running," and "runs" are reduced to "run." Lemmatization helps simplify the analysis and improves the accuracy of natural language processing models, allowing related terms to be analyzed more effectively.

3.2.1.7 Final Review

Before concluding the cleaning process, a thorough review of the texts is conducted. This review includes checking for unwanted characters, ensuring that stop words have been correctly removed, and confirming that lemmatization has been effectively applied. This validation is crucial to ensure that the texts are ready for the subsequent stages of analysis.

3.2.1.8 Storage of Cleaned Texts

Finally, after completing the cleaning and review of the texts, the results are stored in a new file or DataFrame. This file contains the cleaned and normalized texts, ready to be used in the polarized analysis of notes. The organized structure of the data facilitates further analysis, allowing machine learning models and other data analysis techniques to be applied smoothly.

3.3 Text Analysis

Text analysis is a fundamental process in research, especially in the context of polarized analysis of journalistic notes. This section describes the methodologies employed to extract valuable information and understand the trends present in the analyzed texts. Below, the steps taken during text analysis are detailed.

3.3.1 Actor Extraction

Actor extraction is a crucial component of text analysis, particularly in the context of polarized analysis of journalistic notes. This process involves identifying and classifying relevant actors mentioned in the text, such as people, organizations, and entities. Below, the development of the actor extraction model, the approach used, and the justification behind the choice of this specific method are outlined.

3.3.1.1 Actor Extraction Models

Before implementing the actor extraction model, a comprehensive investigation was conducted on various techniques and methodologies available in the field of natural language processing (NLP). During this phase, different approaches were explored, including the use of rule-based techniques, machine learning algorithms, and pre-trained named entity recognition (NER) models. Popular NLP libraries such as SpaCy, NLTK, and Stanford NLP were reviewed, evaluating their capabilities and results in entity identification. Initial tests were also conducted using different conFiguretions and parameters of the models, aiming to determine which offered the best accuracy and robustness in extracting actors from journalistic notes.

Box 2

Table 1

Comparative Table of Actor Identification Models

Feature	spaCy	NLTK	Stanford NLP
Supported Languages	Multilingual (over 60 languages)	Primarily English, but supports other languages	Multilingual (various languages)
Ease of Use	Intuitive interface, easy to use	Requires more setup and prior knowledge	More technical, requires prior knowledge
Speed	Very fast and efficient in processing	Slower, suitable for research projects	Moderately fast, depends on the model
Model Size	Lightweight models optimized for production	Large and resource-intensive models	Large models, less suitable for production
Main Features	Tokenization, part-of-speech tagging, entity recognition, syntactic analysis, etc.	Tokenization, part-of-speech tagging, syntactic analysis, translation, etc.	Entity recognition, syntactic analysis, dependency analysis, etc.
Integration	Compatible with other libraries like TensorFlow and PyTorch	Can integrate with other libraries but requires more effort	Can be integrated into Java applications and other languages
Community and Support	Active community, good documentation and resources	Large community, good documentation, but may be outdated in some areas	Smaller community, good technical documentation
Typical Use	Industrial applications, real-time processing, production	Academic research, prototyping, education	Research, advanced natural language processing

Source: obtained from (Honnibal et al., 2020), NLTK (Bird et al., 2009) and Stanford NLP (Manning et al., 2014)

After a comparative analysis, the decision was made to use the SpaCy model, which offers a robust and efficient approach for named entity recognition (NER). This pre-trained model is based on neural networks and has proven effective in identifying different types of entities, including people (PER), organizations (ORG), and locations (LOC). The choice of SpaCy was based on its ability to handle texts in Spanish and its superior performance in NER tasks.

3.3.1.2 Model Implementation

Once the tool and approach for actor extraction were selected, the model was implemented using the SpaCy library. This section details the implementation process, the code design, and how the model's effectiveness in extracting relevant actors from journalistic notes is ensured. The first step in the implementation was to set up the development environment. The SpaCy library was installed, and the Spanish model was downloaded, which is essential for processing text in this language. Subsequently, the extraer_actores function was designed, which is responsible for analyzing a text and extracting the mentioned actors and locations.

3.3.2 Text Classification

Text classification is one of the fundamental stages in the polarized analysis of journalistic notes, as it allows for the organization of information into thematic and contextual categories, facilitating interpretation and analysis. The main objective of this stage is to assign a category to each processed text so that its predominant theme can be identified and its political, social, or ideological orientation evaluated. Before proceeding with the implementation of the classification model, a thorough investigation was conducted into different text classification approaches. Various methods were explored, ranging from traditional supervised learning techniques, such as Naive Bayes and Support Vector Machines (SVM), to more advanced approaches based on neural networks and pre-trained language models like BERT (Bidirectional Encoder Representations from Transformers).

The selection of models to consider was based on their ability to handle texts in Spanish, as well as their performance in text classification tasks. During the research phase, various NLP libraries and frameworks were evaluated, including Scikit-learn, TensorFlow, and Hugging Face's Transformers, which provide efficient implementations of models like BERT.

The code presented aims to train a model based on BERT for classifying texts into various categories, such as politics, economics, sports, among others. Here I explain step by step what each section does: The first step is to load a CSV file that contains the texts and their respective category labels.

These labels are converted into numerical values so that the model can process them. The dataset is split into two subsets: one for training (80%) and the other for evaluation (20%). This ensures that the model can be trained and then evaluated to measure its performance.

For the BERT model to understand the texts, they must be tokenized. This involves converting each text into a sequence of tokens (numbers) that the model can process. The pre-trained BERT tokenizer is used for this task.

3.4 Sentiment Analysis

The sentiment analysis section of the project focuses on understanding the emotions and perceptions conveyed through journalistic texts. To carry out this analysis, a series of functions were developed that, step by step, allow for the cleaning, translation, and ultimately the analysis of sentiment in news articles using the VADER sentiment analyzer approach. First, simple methods like TextBlob, Flair, VADER, and Roberta were explored for sentiment analysis. Below is a comparative table of the models:

Box 3

Table 2

Comparative Table of Sentiment Analysis Models

Model	Advantages	Disadvantages
VADER (Valence Aware Dictionary and Entiment Reasoner)	- Lightweight and fast. - Does not require large amounts of data for training. - Performs well with short and social texts. - Provides a clear interpretive output in terms of polarity.	- Primarily optimized for English texts, requiring translation of Spanish texts before analysis. - May be less accurate with long texts and complex structures.
TextBlob	- Easy to implement and use. - Provides both sentiment analysis and text subjectivity.	- Lower accuracy with more complex structured texts. - More suitable for general texts and not as focused on detailed analysis as VADER.
Flair	- Excellent accuracy in sentiment classification tasks. - Works well with long texts and more formal language.	- Slower compared to lightweight models like VADER. - Requires more computational resources and is more complex to implement.
Roberta	- High accuracy in sentiment analysis. - Adaptable to different text classification tasks.	- Requires substantial resources and time for training.

Source: VADER (Hutto & Gilbert, 2014), TextBlob (Loria, 2018), Flair (Akbik et al., 2019) and RoBERTa (Liu et al., 2019)

Finally, it was decided to opt for VADER, a lightweight rule-based model specifically designed for sentiment analysis in informal texts such as social media posts and news articles. This model demonstrated high efficiency and speed, making it ideal for analyzing the journalistic notes of this project. Moreover, VADER can interpret the overall tone of a text and provides a composite score that indicates the overall polarity of the document. Although VADER is optimized for English texts, the implementation of an automatic translation process allowed its adaptation to journalistic notes in Spanish. This approach achieved a balance between speed and accuracy, making VADER the most suitable model for sentiment analysis in this context.

The sentiment analysis process begins with a thorough cleaning of the text to remove irrelevant elements and enhance the accuracy of the analysis. Next, the text is translated into English to be processed by the VADER analyzer, which returns a set of polarity scores. These scores are then interpreted to determine whether the text reflects a negative, neutral, or positive trend, providing valuable insights into the tone of the journalistic notes.

To determine the sentiment trend in the journalistic notes, value ranges were assigned based on the composite score provided by the sentiment analysis model.

Box 4

Table 3

Sentiment Trend Range

Sentiment Trend Range	
1 : 0.66	Positive
0.65 : 0.26	Neutral - Positive
0.25 : -0.25	Neutral
-0.26 : -0.65	Neutral - Negative
-0.66 : -1	Negative

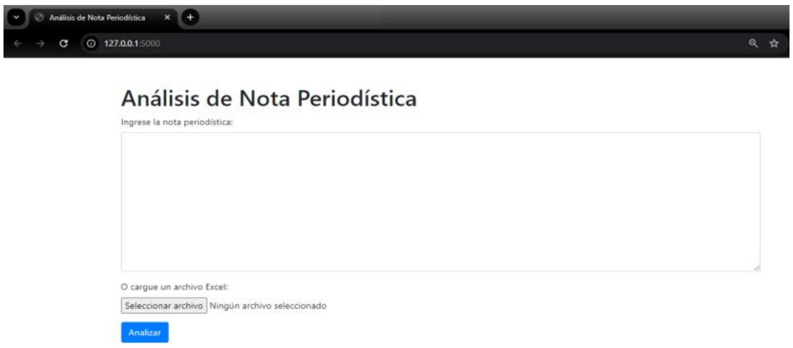
4. Results

The following are the results obtained so far in the project for Polarized Analysis of journalistic notes. These results reflect the process of text classification, extraction of key actors, and sentiment analysis, highlighting the trends identified in the various documents analyzed.

4.1 Main Page

A user-friendly interface was developed to facilitate the use of the Polarized Analysis system for journalistic notes. This interface includes two main sections that allow users to interact with the tool in a straightforward and efficient manner. The first section is designed to analyze a single text note, providing immediate results on the classification, extraction of actors, and sentiment analysis of the entered content. The second section allows for the upload of files in .CSV format, enabling the simultaneous analysis of multiple notes, optimizing the bulk processing of information, and facilitating comparative studies of large volumes of data.

Box 5



Análisis de Nota Periodística

Ingrese la nota periodística:

O cargue un archivo Excel:

Ningún archivo seleccionado

Figure 2

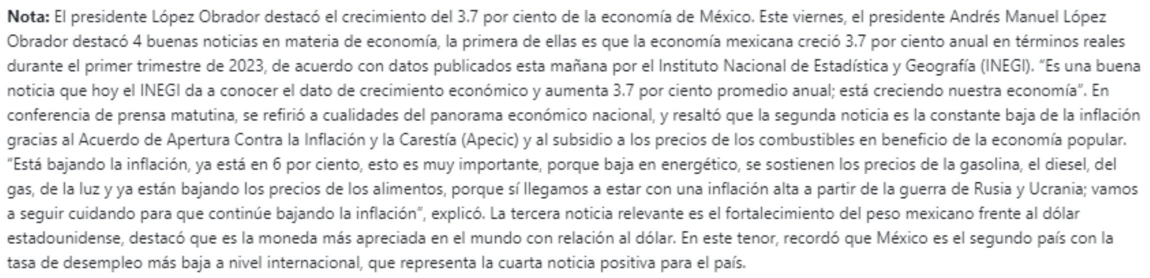
Main Page

Source: Own Work

4.1 Text Cleaning

The text cleaning process was implemented efficiently to ensure that the input content is processed correctly before analysis. When a text is entered in the interface, it undergoes a cleaning process that removes special characters, accents, and irrelevant words such as stop words, ensuring that the data is in an optimal format for analysis. Once the cleaning is completed, the system displays the processed text at the bottom of the main page, allowing the user to visualize the result and confirm that the text has been cleaned properly before proceeding with the analysis. This ensures that the content is suitable for the subsequent processing stages, such as classification and sentiment analysis.

Box 6



Nota: El presidente López Obrador destacó el crecimiento del 3.7 por ciento de la economía de México. Este viernes, el presidente Andrés Manuel López Obrador destacó 4 buenas noticias en materia de economía, la primera de ellas es que la economía mexicana creció 3.7 por ciento anual en términos reales durante el primer trimestre de 2023, de acuerdo con datos publicados esta mañana por el Instituto Nacional de Estadística y Geografía (INEGI). "Es una buena noticia que hoy el INEGI da a conocer el dato de crecimiento económico y aumenta 3.7 por ciento promedio anual; está creciendo nuestra economía". En conferencia de prensa matutina, se refirió a cualidades del panorama económico nacional, y resaltó que la segunda noticia es la constante baja de la inflación gracias al Acuerdo de Apertura Contra la Inflación y la Carestía (Apecic) y al subsidio a los precios de los combustibles en beneficio de la economía popular. "Está bajando la inflación, ya está en 6 por ciento, esto es muy importante, porque baja en energético, se sostienen los precios de la gasolina, el diesel, del gas, de la luz y ya están bajando los precios de los alimentos, porque sí llegamos a estar con una inflación alta a partir de la guerra de Rusia y Ucrania; vamos a seguir cuidando para que continúe bajando la inflación", explicó. La tercera noticia relevante es el fortalecimiento del peso mexicano frente al dólar estadounidense, destacó que es la moneda más apreciada en el mundo con relación al dólar. En este tenor, recordó que México es el segundo país con la tasa de desempleo más baja a nivel internacional, que representa la cuarta noticia positiva para el país.

Figure 3

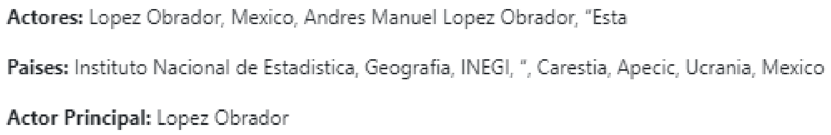
Cleaned text without special characters

Source: Own Work

4.2 Actor Extraction

Similarly, the results of the actor extraction demonstrated that the process was successfully implemented, accurately identifying the main actors mentioned in the journalistic notes. Using a natural language processing model, entities of type person and location were detected within the analyzed texts. The developed code enabled the extraction of names of individuals and countries, removing accents to ensure greater consistency in the results.

Box 6



Actores: Lopez Obrador, Mexico, Andres Manuel Lopez Obrador, "Esta

Paises: Instituto Nacional de Estadistica, Geografia, INEGI, ", Carestia, Apecic, Ucrania, Mexico

Actor Principal: Lopez Obrador

Figure 4

Actor Identification

Source: Own Work

The model's accuracy was validated through various tests, demonstrating that key actors were correctly recognized across a wide range of notes, regardless of the length or complexity of the texts.

4.3 Identification of Category and Genre of the Journalistic Note

Currently, the process of identifying the category and genre of journalistic notes is in the training stage. Using the BERT model, work is being done on classifying the notes into various categories such as politics, economy, sports, culture, technology, science, society, international affairs, and climate.

Additionally, different approaches are being tested for identifying the genre of the notes, determining their journalistic genre as information, opinion, or political columns. As the model progresses in its training, it is expected to improve its accuracy and ability to classify more precisely, contributing to more efficient segmentation and higher-quality polarized analysis.

Although promising advances have been observed, the system is still being adjusted and optimized to ensure it can robustly handle a variety of texts with different structures and tones. The final results will be evaluated once the model has completed this training phase.

5. Conclusions

In conclusion, this chapter addressed the development and application of a polarized analysis system for journalistic notes, utilizing advanced techniques in natural language processing (NLP) and machine learning. Throughout the project, various key components were successfully implemented, including text cleaning, actor extraction, sentiment analysis, and classification of the notes into specific categories. Each of these stages was fundamental to ensuring that the analysis of the notes was conducted accurately and efficiently.

The use of models such as BERT for category classification and tools like VADER for sentiment analysis enabled the automation of processes that traditionally required a manual approach. While some of the models are still being fine-tuned and trained, preliminary results are promising, indicating that the system will be capable of processing large volumes of journalistic data with high accuracy.

Furthermore, a user-friendly interface was designed to facilitate interaction with the system, allowing for the analysis of both individual notes and sets of notes through .CSV files. This makes the project accessible to users with different levels of technical expertise, ensuring that the polarized analysis can be applied in various contexts, such as academic, journalistic, or public opinion studies.

Overall, this chapter demonstrates the value of integrating NLP and machine learning techniques in the analysis of journalistic content, providing a versatile and efficient tool for understanding polarization, trends, and key actors within the media discourse. It is anticipated that, with future adjustments and the completion of the model training, the system will reach its full potential, offering high-quality and useful results for users.

Declarations

Conflict of interest

The authors declare that they have no conflict of interest. They have no financial interests or personal relationships that could have influenced this book.

Authors' contribution

Hermenegildo-Domínguez, Cesar: Design of the project idea. Choice of tools, methods, architecture and system techniques. Software development.

Reyes-Nava, Adriana and López-González, Erika: Support in the idea of the project, Distribution of parts of the project, Review of the article.

Availability of data and materials

All data used for this research were derived from our own data analysis, no information from third parties was used.

Funding

This research did not receive any type of funding.

Acknowledgments

We thank the Tecnológico de Estudios Superiores de Jocotitlán and the entire academic community for their support during the development of this project. Their contribution in technical resources and academic guidance was essential to carry out this research.

Abbreviations

ABSA Aspect-Based Sentiment Analysis
BERT Bidirectional Encoder Representations from Transformers
CD Continuous Deployment

CI Continuous Integration
CNN Convolutional Neural Networks
CSV Comma-Separated Values
IaC Infrastructure as Code
LDA Latent Dirichlet Allocation
NER Named Entity Recognition
NLP Natural Language Processing
NLTK Natural Language Toolkit
RNN Recurrent Neural Networks
SVM Support Vector Machine

References

Backgroud

ISO. (n.d.). Artificial Intelligence – Natural Language Processing.

IAAR. (n.d.). Natural Language Processing.

Jurafsky, D., & Martin, J. H. (2021). Speech and Language Processing (3rd ed.). Prentice Hall.

Basic

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval. Cambridge University Press.

Díaz, J. E. M. (2021). Natural language models in scientific research: A technical description. Innovation and Science.

Osaba, A., Gorman, L., & Liu, Y. (2021). Ethics and limitations in natural language models.

Rosati, G. F. (2022). Natural Language Processing applied to social sciences: Topic detection in tango lyrics. Latin American Journal of Social Research Methodology, 12(23), 38-60.

Support

Parratt, S. F. (2001). El debate en torno a los géneros periodísticos en la prensa: nuevas propuestas de clasificación. Zer: Revista de estudios de comunicación= Komunikazio ikasketen aldizkaria, 6(11)..

Atlassian. (2023). What is DevOps? Atlassian. Retrieved from https://www.atlassian.com/devops/what-is-devops

Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Explosion AI.

Bird, S., Klein, E., & Loper, E. (2009). Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit. O'Reilly Media.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., & McClosky, D. (2014). The Stanford CoreNLP natural language processing toolkit. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, 55-60.

Differences

Hutto, C., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. Proceedings of the International AAAI Conference on Web and Social Media, 8(1), 216-225.

Loria, S. (2018). TextBlob: Simplified text processing. TextBlob Documentation.

Discussions

Akbik, A., Bergmann, T., & Vollgraf, R. (2019). Pooled contextualized embeddings for named entity recognition. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 724-728.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692.