TARP

DIGITAL ASSIGNMENT -3 NAME: KANIKA SHARMA REG. NO. 21MIC0101

**GROUP NO. 6** [21MIC0101, 21MIC0074, 21MIC0153, 21MIC0082, 21MIC0087]

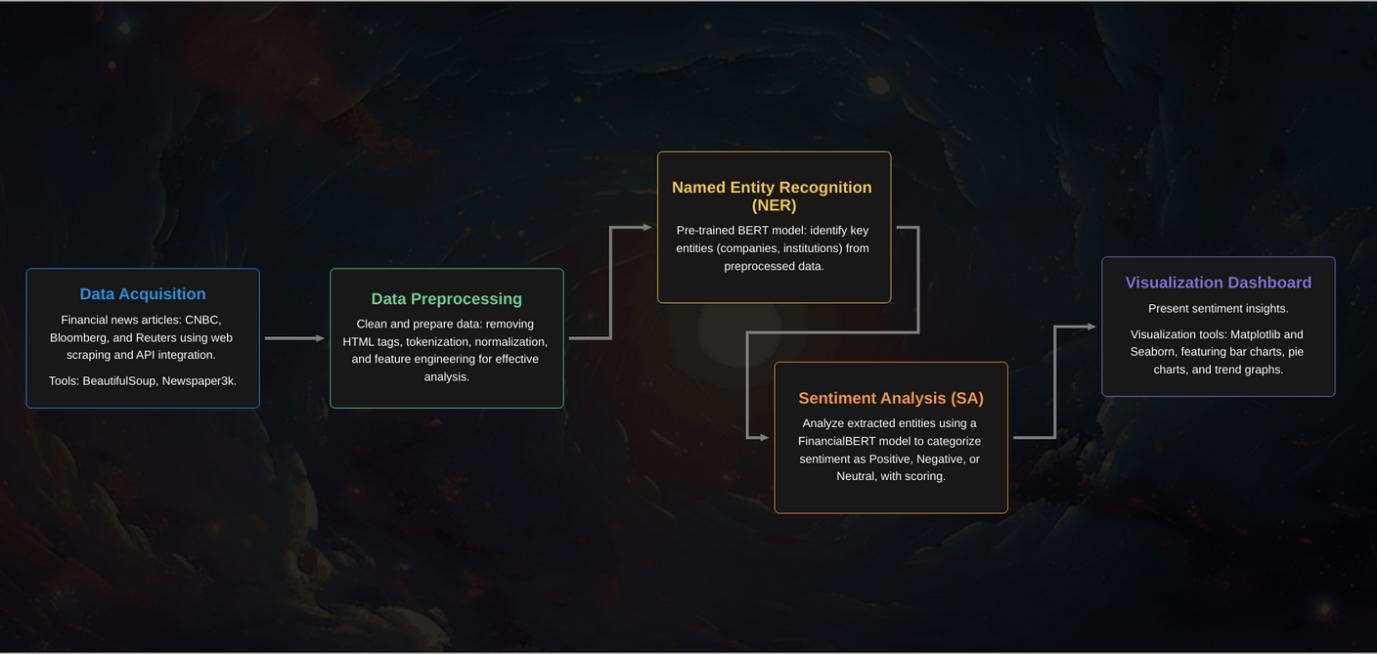
**PROJECT TITLE**: StockSense: An NLP based tool for stock market analysis

**A Comprehensive Overview Of The Proposed Methodology**

1. ​Block Diagram of the Project

The StockSense system is designed to follow a structured approach, ensuring accurate and effective sentiment analysis of financial news articles. It comprises five major components that work in conjunction to extract meaningful insights from financial text data. These components include Data Acquisition, Data Preprocessing, Named Entity Recognition (NER), Sentiment Analysis (SA), and the Visualization Dashboard. Each of these plays a critical role in processing and analyzing the data efficiently.

Block Diagram



1. ​Research Process

The research methodology follows a sequential workflow to maintain accuracy, reliability, and effectiveness in sentiment analysis. The process begins with the collection of financial news articles from trusted sources through web scraping and data retrieval APIs. Once the raw data is acquired, it undergoes preprocessing, where it is cleaned, tokenized, and structured to facilitate further analysis.

Named Entity Recognition (NER) is applied next to identify key financial entities within the articles, such as company names, stock symbols, and financial institutions. This is followed by Sentiment Analysis (SA), where the extracted entities are classified based on the sentiment expressed towards them in the articles. The final step in the research process is performance evaluation and visualization, where sentiment scores are assessed using various performance metrics and the results are presented in an interactive dashboard.

1. ​Description of Each Block

Data Acquisition

Data acquisition is the foundation of the StockSense system, where financial news articles are collected from multiple credible sources such as CNBC, Bloomberg, Reuters, and Financial Times. The system employs web scraping tools such as BeautifulSoup and Newspaper3k to extract textual data from these sources in real time. Additionally, an API integration feature allows direct data retrieval, ensuring a continuous and up-to-date supply of financial news.

The acquired data is then formatted and stored in structured datasets using JSON and CSV formats, making it easier to process in subsequent stages. This structured approach ensures that the extracted data is suitable for analysis and minimizes errors associated with unstructured text data.

Data Preprocessing

The preprocessing stage involves a series of steps to clean and prepare the acquired data for analysis. Initially, unwanted elements such as HTML tags, special characters, and stopwords are removed. The text is then normalized by converting it to lowercase and standardizing abbreviations to maintain uniformity.

Tokenization is performed to break the text into meaningful words and sentences, facilitating efficient entity recognition and sentiment analysis. Feature engineering techniques are applied to extract relevant features that enhance model performance. Lastly, the dataset is split into training and testing subsets to validate the model's performance and ensure its accuracy.

Mathematical functions used in preprocessing include:

* Tokenization formula: X = tokenize(T)
* Text normalization function: 𝑇𝑛𝑜𝑟𝑚 = 𝐿𝑜𝑤𝑒𝑟𝐶𝑎𝑠𝑒(𝑇)

Model Implementation

The StockSense system leverages advanced Natural Language Processing (NLP) techniques to extract insights from financial news data. The two primary models used in the system are Named Entity Recognition (NER) and Sentiment Analysis (SA), both based on pre-trained BERT architectures.

Named Entity Recognition (NER) The NER module employs the dslim/bert-base-NER model, which is fine-tuned to recognize financial entities such as company names and organizations. The model processes financial text data and extracts named entities,

classifying them into predefined categories. This ensures that relevant financial terms are accurately identified and analyzed.

Sentiment Analysis (SA) The Sentiment Analysis module utilizes the FinancialBERT- Sentiment-Analysis model, specifically trained for financial sentiment classification. The model assesses the sentiment expressed towards the extracted financial entities and categorizes them into positive, negative, or neutral sentiments. By assigning sentiment scores to each entity, the system provides valuable insights into market trends and investor sentiment.

The model training process involves fine-tuning pre-trained models using financial datasets from sources like Reuters and Bloomberg. Feature engineering is also employed to refine the classification accuracy.

Equations used in sentiment analysis include:

* Sentiment Score Calculation: 𝑆(𝑐) = ∑(𝑤 ∗ 𝑠)

Performance Measures

The effectiveness of the StockSense system is evaluated using multiple performance metrics to ensure the accuracy and reliability of sentiment predictions. Key metrics include:

* Accuracy: 𝑇𝑃+𝑇𝑁

TP + FP+TN+FN

* Precision: 𝑇𝑃

TP + FP

* Recall: 𝑇𝑃

TP + FN

* F1-Score: The harmonic mean of precision and recall, ensuring a balance between the

two.

* Mean Squared Error (MSE): Used for regression-based sentiment scoring.

These metrics help in fine-tuning the model to improve sentiment classification performance and ensure that it aligns with real-world financial trends.

1. ​Model Training & Prediction

The training process of the StockSense system involves fine-tuning the pre-trained BERT models using domain-specific financial datasets. The models are trained using labeled financial data, allowing them to learn the contextual nuances of sentiment in financial news.

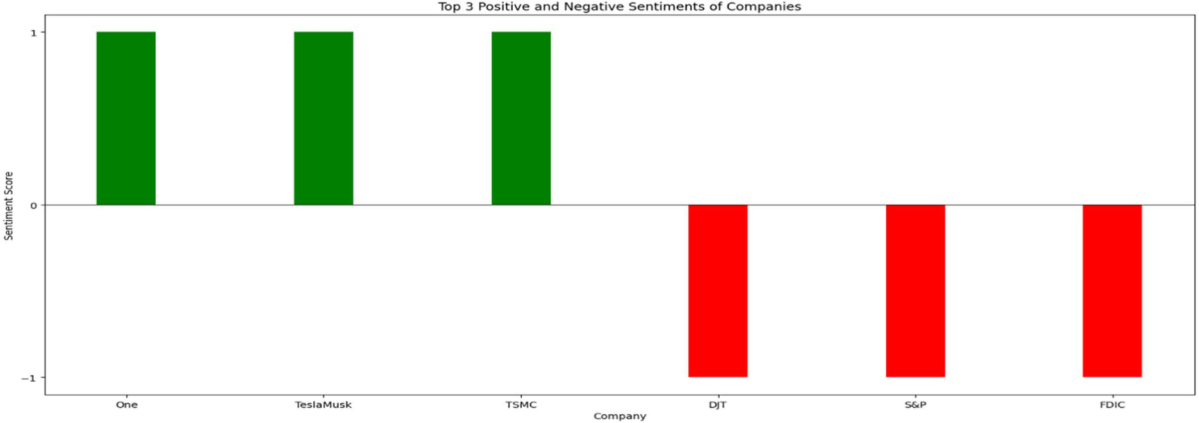
During the prediction phase, the trained models analyze newly acquired financial news articles, extract relevant entities, and determine the sentiment associated with them. The results are presented as sentiment scores, which provide insights into how different financial entities are perceived in the market. These predictions assist investors and analysts in making data-driven decisions.

1. ​Visualization & Insights

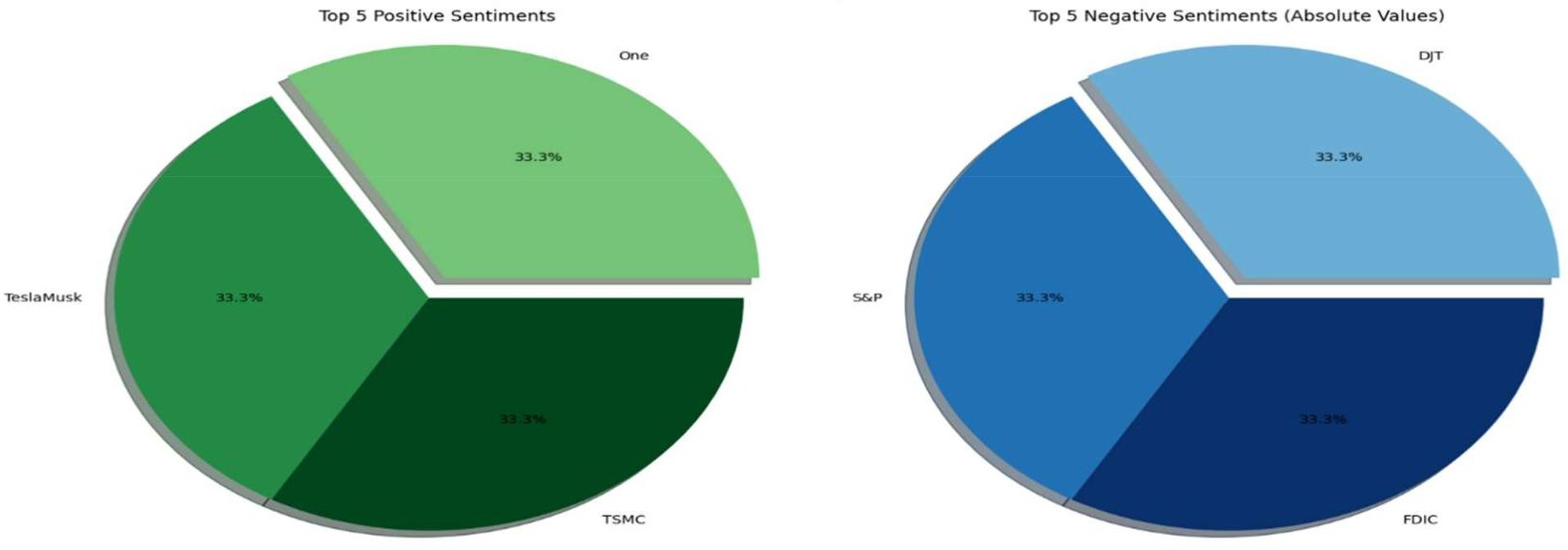
The visualization module plays a crucial role in presenting sentiment analysis results in an interactive and easily interpretable manner. The StockSense system employs visualization tools such as Matplotlib and Seaborn to generate comprehensive graphs and charts.

The primary visualization techniques used include:

* + Bar Charts: Displaying the top companies with the highest positive and negative sentiment scores.



* + Pie Charts: Representing the sentiment distribution across multiple financial entities.



* + Trend Graphs: Showing the variation in sentiment over time for specific companies.



By utilizing these visual elements, the StockSense system provides users with a clear and structured overview of market sentiment, enabling them to identify trends and patterns effectively.

The StockSense project provides a robust and accurate approach to financial sentiment analysis by integrating advanced NLP techniques such as Named Entity Recognition (NER) and Sentiment Analysis (SA). By leveraging high-quality data sources, implementing structured preprocessing steps, and utilizing machine learning models, the system delivers actionable insights for financial decision-making.

Future enhancements to the StockSense system will focus on improving entity recognition accuracy, incorporating real-time interactive visualizations, and expanding support for multi- language analysis. These advancements will further enhance the usability and effectiveness of the system, making it a valuable tool for financial analysts and investors.