

Project Proposal

Course: DATS 6312.11- NLP

Members: Abirham Gettie, Julio Montoya

Financial markets are heavily influenced by the tone of news headlines. Investors often react to positive or negative financial news, which can lead to noticeable stock market movements. However, with the large volume of news generated daily, manually interpreting these sentiments is impractical. Therefore, the objective of this project is to leverage Natural Language Processing (NLP) techniques to automatically analyze financial news headlines and classify them as positive, negative, or neutral. The resulting sentiment labels can be correlated with stock market behavior to help forecast potential stock market impacts.

The project's methodology involves preprocessing financial news headlines by cleaning text, tokenizing, lemmatizing, removing stopwords, and normalizing it for consistency. Features will be extracted using Bag of Words (BoW), TF-IDF, and word embeddings such as Word2Vec to represent semantic meaning. Machine learning models like Naive Bayes and Logistic Regression, followed by deep learning models such as LSTM and transformer-based models like BERT for advanced sentiment classification. Pre-trained embeddings will be fine-tuned on financial-domain data to enhance model accuracy and contextual understanding. The implementation will utilize libraries such as NLTK, SpaCy, Scikit-learn, Gensim, PyTorch, Pandas, NumPy, Matplotlib for preprocessing, model training, evaluation, and visualization. In conclusion, this project aims to apply Natural Language Processing (NLP) techniques to automatically evaluate financial news headlines and determine their sentiment as positive, negative, or neutral. These sentiment classifications can then be linked to stock market patterns, providing valuable insights for predicting potential market movements and supporting data-driven investment decisions.

Dataset:

Link: https://www.kaggle.com/datasets/miguelaelle/massive-stock-news-analysis-db-for-nlpbacktests?select=raw_partner_headlines.csv

This dataset compiles headline information from financial news articles aggregated from multiple reputable sources, with the primary source being Benzinga.com. It provides a comprehensive collection of financial news headlines covering a wide range of companies, industries, and market events. The dataset includes several key attributes such as the index, headline text, URL, publisher, publication date, and the corresponding stock symbol associated with each news item. Spanning the years 2009 to 2020, it captures over a decade of financial reporting, making it an invaluable resource for analyzing how news sentiment and tone influence stock market movements.