Daniel Kadosh
Sankalp Shubham

Sentiment Analysis Report

Abstract

This project focuses on utilizing Natural Language Processing (NLP) techniques for stock prediction through sentiment analysis of news articles. The objective is to develop a comprehensive pipeline that collects news articles from multiple sources, preprocesses the text, analyzes sentiment using the BERT model, and derives actionable insights for trading decisions.

Problem Description

We wanted to make a program that would be able to make a conclusion on a stock using current published information. We created this model so that it would save the user time instead of going and reading articles they could use our model and get a score of how the stock is currently doing

Approach

The pipeline begins by fetching articles from two news resources, Google News and newsNow, we fetch around 100 articles and pre-process by first limiting the number of articles from one news source to 3 to prevent bias, then we split each article into paragraphs and remove all the stop words. The next step is to tokenize the data in the paragraphs so we can give it to our model BERT. After calculating the polarity of all the paragraphs and averaging them, we use a different model, Gemini, for further analysis. BERT and Gemini might experience inconsistency so to battle that we take the average of the polarity scores from both models.

Data Description

The data consists of articles from many news resources that are being fetched using several APIs. The data is mostly formatted into paragraphs when it is fetched but after pre-processing all of them are formatted into paragraphs. The apis cannot fetch some articles' text so we remove them from the dataset.

Experiments and Error Analysis

Daniel Kadosh
Sankalp Shubham

Our first approach to sentiment scoring was splitting the articles into paragraphs and then into sentences. After we have the sentences we get the polarity of each one and then average them to get the polarity score for that specific paragraph, then in the end we average the polarity scores of the paragraphs to get a final polarity score. This approach gave us mixed results with many inaccurate results due to neutral sentences affecting the polarity score of the paragraph so we went with our current approach of just averaging each paragraph's score.

We also tried a different model called VADER but it was also very inconsistent giving us very large polarity scores that when compared with other models did not make sense. Bert gave us better overall results so we went with it.

Conclusion

The conclusion of our program is whether you should hold, invest, or sell the stock that is being provided. We went with a threshold method, after all the polarity scores are calculated we decide using the following threshold: between -1 and 0.2 sell, between 0.2 and 0.5 hold and between 0.5 and 1 buy. The reason for selling being such a big portion of the threshold is due to the inconsistency of the models being used and due to the reason that we are limited to the amount of articles we can fetch

References

BERT - https://huggingface.co/kwang123/bert-sentiment-analysis
VADER - https://www.nltk.org/_modules/nltk/sentiment/vader.html
Scikit Learn - https://scikit-learn.org/stable/