Scraping & Summarizing Financial News for Investors

SMITKUMAR B. SARAIYA, University of Calgary, Canada

This project explores the application of natural language processing using NLTK and SpaCy for financial news summarization and sentiment analysis. The goal is to develop a pipeline that scrapes financial news from platforms like Bloomberg, Reuters, or Yahoo Finance, processes the data, and delivers condensed, actionable information to investors. The system will incorporate sentiment analysis using tools like FinBERT or VADER to provide additional market insights, as sentiment has been shown to impact stock prices significantly.

1 Introduction

The financial market is a very volatile space and can change a lot depending on a lot of factors stemming from bold business decisions and strategies by groups of people or powerful individuals with influence in the stock market. For this project I will focus on investors who use financial news from mediums such as newsletters, articles, annual reports and so on to make informed financial decisions.

This project aims to explore how natural language processing using libraries like NLTK and SpaCy can play a role in financial news summarization with a sentiment analysis module on financial news and deliver summarized information to investors. The intention is to scrape data from a chosen financial news platform(s) such as Bloomberg, Reuters or Yahoo Finance in order to gather financial news as data.

2 Related Work

There are a multitude of projects that have leveraged natural language processing (NLP) for different end goals within a financial space or financial topic surrounding their project. Some literature review suggests the use of the AZFin text system to collect financial news data in order to do things such as stock market predictions or even the end goal of this project which is summarization but in parallel with AI and AI techniques which I do not look to implement.

While leveraging NLP libraries such as NLTK and SpaCy, other python libraries such as BeautifulSoup, transformer, and pandas amongst others will be used to fully leverage output from

Author's Contact Information: Smitkumar B. Saraiya, smitkumar.saraiya@ucalgary.ca, University of Calgary, Canada.
Manuscript submitted to ACM

NLP techniques to deliver the final product. This project looks to use pretrained models such as FinBERT which was trained by further training the BERT model for use in financial data processing pipelines, specifically financial sentiment classification. Another option for this particular use is VADER within the nltk.sentiment library to use for sentiment analysis. Sentiment analysis is key in this project as it has been shown that market sentiment can impact stock prices and this valuable information can be of very good use to investors.

3 Proposed Work

The plan to go about this project is to begin scraping news data after reviewing sources and collecting enough in order to be able to start preprocessing and working with the data to finally summarize.

The idea is to have a refined pipeline that in the end can be run to continuously or periodically scrape data from the financial news sources or the web, depending on what is chosen, in order to deliver periodical summarized data that is ready to use for investors.

Summarizations will be done using other libraries such as sumy or spaCy to extract key sentences or other models available on hugging face such as BART or T5 for more human-like text.

As mentioned previously, sentiment analysis will be conducted using libraries/tools such as FinBERT, VADER or both for comparison as to see which one performs better before using one of them into an active pipeline.

In terms of finally storing or making use of the extracted data, the summarized text could be stored in JSON format or within a CSV which can be further used to read and host on a custom web application or within an application on the end users system.

Upon the project's completion, the capabilities of the pipeline would most likely be demonstrated using final extracted data, Although, the hope is still there to do a live demo of the pipeline scraping, processing, and finally giving summarized output for use. For the purpose of the project there will be heavy focus on how NLP libraries and techniques were used within the pipeline to give best results possible.

4 Evaluation

Evaluation of the results is intended to be done by assessing accuracy of summarization using ROGUE scores. Other metrics that could be evaluated are sentiment analysis results compared against historical stock market trends. Finally, the manual route can be taken to evaluate summaries by reviewing them to check for coherence and relevance.

References

[1] Khant, A., & Mehta, M. (2018). Analysis of financial news using natural language processing and artificial intelligence. In *Proceedings of the 2018 International Conference on Business Innovation (ICOBI)*. NSBM, Colombo, Sri Lanka.

[2] Schumaker, R. P., & Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The AZFinText system. *ACM Transactions on Information Systems*, 27(2), Article 12.

[3] ProsusAI. (n.d.). FinBERT. Hugging Face. Retrieved from https://huggingface.co/ProsusAI/finbert