Assessment 3: WebCrawler and NLP System

Document 1: Overview of WebCrawler and NLP Tasks

1 Abstract

Social media forums capture the collective crowd discussion on many topics and crowd sentiment or crowd predictions may give valuable insights into topics of interest. This report investigates if a stock market forum can be analysed using NLP Engines to extract those insights and provide a tool for predicting future stock market movements.

This analysis should not be used for real life trading of stocks as there are many assumptions made in this analysis.

2 Introduction / Overview

This report overviews the WebCrawler and NLP Systems developed to investigate if social media posts on a popular stock market forum are predictive of the movement of the stock market based on NLP analysis and sentiment analysis.

The investigation required the development of two WebCrawlers. The first WebCrawler extracted the forum post from https://hotcopper.com.au/ (https://hotcopper.com.au/). Hot Copper is Australia's largest stock trading and investment internet discussion forum with over 250,000 registered members who actively post on the forum, and has 21 million monthly page impressions (HotCopper, 2021). The second WebCrawler extracted historical stock market data from https://au.finance.yahoo.com/ (https://au.finance.yahoo.com/). Yahoo Finance is a source of free historical stock market data available to the public. Yahoo! Finance is a media property that is owned by Verizon Media (Media, 2021).

Two NLP systems were developed to process the extracted forum posts from the WebCrawler and analyse the data to find underlying insights and patterns. This was compared to the historical stock market data to see if the insights were predictive of movements in the stock market.

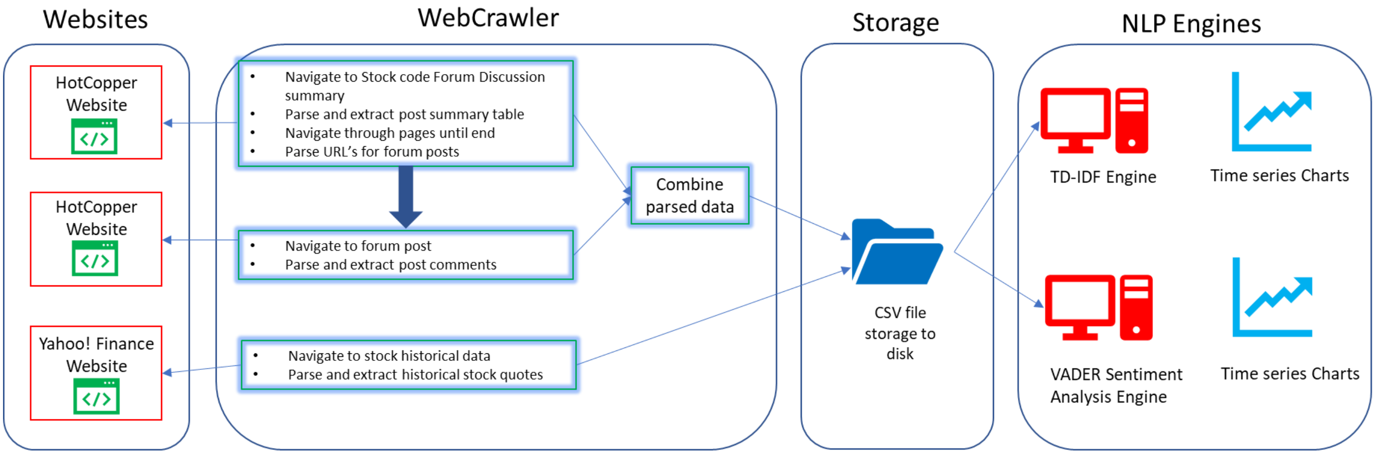
The first NLP system normalised the forum posts using Part of Speech (POS) tags and Lemmatization pre-processing before being passed to the Term Frequency-Inverse Document Frequency(TF-IDF) vectorizer algorithm. The TF-IDF vectorizer produces weights which show the importance or how significant the keyword is in the whole corpus (Goralewicz, 2021). These weights were then plotted on a timeseries chart against the adjusted closing stock price to visually evaluate the predictive ability of the NLP system.

The second NLP system used Vader which is a simple rule-based model for sentiment analysis. VADER (Valence Aware Dictionary for sEntiment Reasoning) uses a combination of qualitative and quantitative methods to construct a gold-standard list of lexical features and their associated sentiment intensity measures, to determine the sentiment score of the forum posts (C.J. Hustto, 2014). These sentiment scores were then plotted on a time series chart against the adjusted closing stock price and again, visually evaluated for its predictive ability.

3 Architecture

To investigate if social media posts on a popular stock market forum are predictive of the movement of the stock market, the following WebCrawler and NLP Engine Architecture in Figure 1 was designed and implemented.

The first WebCrawler extracts the forum post summary, then WebCrawls each forum post parsing the data before combining and saving to disk as a CSV. The second WebCrawler parses the stock price data and save it to a CSV file on disk. The NLP Engines then use these CSV files to process the data and produce time series charts for result analysis.


Websites

WebCrawler

Storage

NLP Engines

- Navigate to Stock code Forum Discussion summary
- Parse and extract post summary table
- Navigate through pages until end
- Parse URL's for forum posts

- Navigate to forum post
- Parse and extract post comments

- Navigate to stock historical data
- Parse and extract historical stock quotes

Figure 1: WebCrawler / NLP Engine Architecture

4 References

C.J. Hustto, E. G. (2014). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf (http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf)

Goralewicz, B. (2021). The TF*IDF Algorithm Explained. Onely. https://www.onely.com/blog/what-is-tf-idf/#:~:text=TF (https://www.onely.com/blog/what-is-tf-idf/#:~:text=TF)*IDF%20is%20an%20information,IDF%20weight%20of%20that%20term.

HotCopper. (2021). About HotCopper. https://hotcopper.com.au/ (https://hotcopper.com.au/)

Media, V. (2021). Verizon Media Terms of Service. https://www.verizonmedia.com/policies/au/en/verizonmedia/terms/otos/index.html (https://www.verizonmedia.com/policies/au/en/verizonmedia/terms/otos/index.html)