Cardiff University
School of Computer Science and Informatics

Initial Plan
Final Year Project
CM3202 – 40 Credits

Sentiment Analysis of Financial News Headlines with Market Comparison

Author: Mr Harvey Allen

Supervisor: Professor Irena Spasic

7th February 2022

Table of Contents:

1 Project Description

Growth of the internet and the digital economy, along with technical advances in computer and data science have supported a wave of alternative data sources that can be used to measure and predict the financial markets. Alternative data refers to non-traditional data sets that an investor can use to guide their investment strategy and portfolio. In addition to traditional metrics alternative data can provide insight into a securities earnings, economy, expectation, and emotions. This project will focus on market sentiment, the collective attitude towards a security or market, and investigate the hypothesis that the sentiment of financial news articles reflects and directs the performance of the US stock market. In turn this could help to prove or disprove the efficient market hypothesis and random walk theory popularized by Van Horne, 1967.

Consequently, the study will investigate securities within the S&P500 (A stock market index tracking the performance of 500 large companies listed on stock exchanges in the United States) as well as the broader index fund. Inherently the financial news articles will be targeted at the corporations encompassed by this fund, and it will provide the greatest reflection of the US economy at a given time.

Objectively this project can be divided into four core areas: data collection, data processing, data analysis and data visualization. These four areas should all be met assuming they have been completed under the constraints of time.

Initially, the data will need to be collected. This project will be based on the following Kaggle dataset: https://www.kaggle.com/notlucasp/financial-news-headlines

Within this set, data has been scraped from CNBC, the Guardian, and Reuters official websites, the headline in these datasets reflects the overview of the US

economy and stock market every day for the past year to 2 years. Each dataset contains the headline, the last updated date and CNBC and Reuters contain preview text of the articles.

Stock data will be collected from 'Yahoo Finance' and exported into a csv file. The data will include close prices of the stocks during the analysed time period.

Information extraction will take place, in particular, named entity recognition will be used to remove the noise from the data set. In turn this will reduce overfitting within the training data towards specific companies.

Sentiment Analysis will be used to determine the polarity of a headline. Sentiment analysis is the use of natural language processing, text analysis, computational linguistics, and machine learning to analyse people's opinions, appraisals, attitudes, and emotions toward entities, individuals, issues, events, topics, and their attributes (Liu 2011).

Broadly, machine learning techniques will be used to determine the polarity of a headline. A training set corpus will be produced with rows representing independent feature vectors containing information about a specific document (Headline), particular words and its sentiment, relying on public judgement for annotation. Amplifiers, De-Amplifiers, Negators and Adversative Conjunction will also need to be taken in to account for improved accuracy. This training data will then be used to supervise the algorithm, to form the predictive model, based on a tested classification technique to form the classification probabilities. Once this model is formed and can successfully calculate the sentiment of the headlines the mean sentiment for specific company's and the market will be calculated.

Input: Document (d), Fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$
Training Set: Set of n labelled documents $(d_1, c_1), (d_2, c_2), \ldots, (d_n, c_n)$
Output: $c \in C$
(Paul 2021)

Additionally, Statistical Analysis will be performed to determine the correlation coefficient between the sentiment and the price of the given security. If deemed statistically significant the hypothesis can be rejected or accepted if there is causality present.

Finally, the results must be visualized through plotting the valuation of a security alongside its sentiment on a time series graph covered over the range of the dataset. Additionally, a linear regression will be plotted to demonstrate any correlation between the variables. This will provide insight and understanding into the drivers behind the market's gains/ losses.

1.1 Development Methodology

A waterfall development methodology will be employed to structure a sequential chronological development process through all phases of the project. This method has been chosen as the requirements for the project can be gathered and

understood upfront. This choice of methodology also allows a clear measured timeline to be produced.

1.2 Ethical Considerations

Ethical approval will be required for this project as data will be gathered from various online sources including news articles and market data. An additional ethical consideration is the survey that will be conducted to determine the polarity of specific terms when building the training data set. This should negate any bias from my own judgement and add additional insight into sector specific terms.

2 Project Aims and Objectives

As aforementioned in the problem description, this project is bound by time and thus its scope and scalability are not limitless. Thus, it is necessary to categorise the aims and objectives of the project.

A1: Complete research into the various aspects of the project.
A2: Construct an accurate and complete training data set suitable for algorithmic supervision.
A3: Produce a functioning sentiment analysis classification model through machine learning.
A4: Conduct statistical analysis on a minimum of 5 security's including the overarching market.
A5: Complete Documentation and reflection.

2.1 Functional Requirements

The 'MoSCoW' task prioritization method has also been implemented to manage the individual functional requirements of each aim.

A1: Complete research into the various aspects of the project.

- **SHOULD** Research techniques that can be used to implement the tool.
- **SHOULD** Research technologies that can be used to implement the tool.

A2: Construct an accurate and complete training data set suitable for algorithmic supervision.

- **MUST** gain ethical approval from university.
- **SHOULD** conduct a survey to determine the sentiment value of specific terms for labelling.
- **MUST** construct the training set corpus.

A3: Produce a functioning sentiment analysis classification model through machine learning.

- **MUST** produce a sentiment classification tool in python.
- **MUST** implement machine learning.

- **SHOULD** account for Amplifiers, De-Amplifiers, Negators and Adversative Conjunction.

A4: Conduct statistical analysis on a minimum of 5 security's including the overarching market.

- **MUST** plot the price and sentiment of individual securities on a time series.
- **MUST** calculate the correlation coefficient for the price against sentiment.
- **SHOULD** plot the linear regression graph for the price against sentiment.

A5: Complete Documentation and reflection.

- **MUST** write a detailed final report describes all aspects of the project in detail along with supporting information.
- **SHOULD** produce legible, understandable, and maintainable code documentation.
- **SHOULD** provide a justification and evaluation of the produced results.

3 Work Plan

3.1 Research

A fixed period at the beginning of the work plan will be set aside to conduct in-depth research into the fundamental concepts such as machine learning and natural language processing as well as the technologies that will be used to implement the project. Knowledge will be acquired through conducting online research, building background with academic journals, online tutorials, and studies.

3.2 Meetings

Professor Spasic and I have arranged to meet every Monday at 2:30pm. These meetings should ultimately avoid unexpected setbacks and ensure the project remains on track for its duration. Furthermore, the three meetings, in weeks 3, 7 & 10 (when a milestone should be near completion), will be longer in time to review and assess the progress towards each milestone.

3.3 Milestones

The milestones used to assess the projects progress, tied against the aims:

Milestone Map		
Milestone:	Description:	Related Aims:
M1	Research completed into the fundamental concepts/ technologies and all decisions have been made.	**A1**
M2	Production of the sentiment classification tool is complete, and it is fully functional.	**A2, A3**
M3	The project has been concluded and the results visualized and reflected upon.	**A4, A5**

3.4 Deliverables

The following deliverables will be submitted as part of the project:

- **D1:** The final report, documenting all aspects of the project along with any supporting information.
- **D2:** All source code and accompanying files.
- **D3:** Any accompanying documents such as a technical guide.

3.5 Risk Assessment

A risk map has been produced to visualize the specific risks this project faces. It will help to identify and prioritize the risks associated with the project as well as help to clarify the thinking on the nature and impact of the risks.

Risk Map			
Risk:	Probability of Occurrence:	Impact on Project:	Mitigation Response:
Time Constraint	Medium	High	• Adhere to the weekly schedule. • Discuss any issues with the supervisor. • Refine requirements if necessary.
Data Loss	Low	High	• Store backups of data. • Regular documentation.
No Ethical Approval	Low	Medium	• Reliant on the

3.6 Timeline

A workplan has been produced to demonstrate what is expected to have been achieved by what date and the individual tasks the build towards completing a specific milestone. A timeline structure was sufficient over the implantation of a ‘Gnatt Chart’ as tasks could be confined to individual weeks within the project and didn’t extend for longer periods of time.

Weekly Schedule		
Week:	Outline:	Milestones and Deliverables:
01 (31/01/22)	1. Attend weekly meetings with supervisor. 2. Research into NLP and ML techniques. 3. Research technologies that can be used to implement the tool. 4. Produce the initial report. 5. Outline the structure of the final report.	**M1**
02 (07/02/22)	1. Attend weekly meetings with supervisor. 2. Continue research into NLP and ML techniques.	**M1**

	3. Make a justified decision as to the technologies that will be used to implement the tool. 4. Request ethical approval.	
03 (14/02/22)	1. Attend **milestone review** meeting with supervisor. 2. Gather sentiment classification data for labelling. 3. Structure and format the headline data for processing.	**M2**
04 (21/02/22)	1. Attend weekly meetings with supervisor. 2. Determine the labels of training set used for classification. 3. Test different classification techniques. 4. Produce the 'bag-of-words' model. 5. Prepare the training set corpus.	**M2**
05 (28/02/22)	1. Attend weekly meetings with supervisor. 2. Implement a basic neural network in framework of choice.	**M2**
06 (07/03/22)	1. Attend weekly meetings with supervisor. 2. Train the classification model to predict the performance. 3. Tune the neural network through optimising the model.	**M2**
07 (14/03/22)	1. Attend **milestone review** meeting with supervisor. 2. Test the sentiment analysis tool. 3. Complete the sentiment analysis tool. 4. Select the securities that will be processed and visualized. 5. Collect, structure, and format the securities data for visualization.	**M2, M3**
08 (21/03/22)	1. Attend weekly meetings with supervisor. 2. Plot the price and sentiment of individual securities on a time series. 3. Calculate the correlation coefficient for the price against sentiment.	**M3**
09 (28/03/22)	1. Attend weekly meetings with supervisor. 2. Plot the linear regression graph for the price against sentiment.	**M3**
10 (04/04/22)	1. Attend **milestone review** meeting with supervisor. 2. Produce legible, understandable, and maintainable code documentation. 3. Finalise all code and supporting files.	**D2**
11 (11/04/22)	1. Attend weekly meetings with supervisor. 2. Write a justification and evaluation of the produced results. 3. Finalise the report and any accompanying documents.	**D1, D3**

12 (18/04/22)	1. Attend weekly meetings with supervisor. 2. Planned addition time for any tasks not completed. 3. Submit all work.	Submission of deliverables.

Bibliography

Liu, Bing. 2011. Web Data Mining: Opinion Mining and Sentiment Analysis. (Chapter 11). doi: 10.1007/978-3-642-19460-3_11

Paul, S. 2021. Python Sentiment Analysis Tutorial. Available at: https://www.datacamp.com/community/tutorials/simplifying-sentiment-analysis-python

Van Horne, James C.; Parker, George G.C. 1967. The Random-Walk Theory: An Empirical Test. Financial Analysts Journal. doi: 10.2307/4470248