DEPARTMENT OF COMPUTER SCIENCE

Semantic Analysis of Financial Headlines Based on Realised Stock Returns

Joshua Felmeden

A dissertation submitted to the University of Bristol in accordance with the requirements of the degree of Master of Engineering in the Faculty of Engineering.

Thursday 7th April, 2022

Abstract

A compulsory section, of at most 300 words

This section should précis the project context, aims and objectives, and main contributions (e.g., deliverables) and achievements; the same section may be called an abstract elsewhere. The goal is to ensure the reader is clear about what the topic is, what you have done within this topic, *and* what your view of the outcome is.

The former aspects should be guided by your specification: essentially this section is a (very) short version of what is typically the first chapter. If your project is experimental in nature, this should include a clear research hypothesis. This will obviously differ significantly for each project, but an example might be as follows:

> My research hypothesis is that a suitable genetic algorithm will yield more accurate results (when applied to the standard ACME data set) than the algorithm proposed by Jones and Smith, while also executing in less time.

The latter aspects should (ideally) be presented as a concise, factual bullet point list. Again the points will differ for each project, but an might be as follows:

- I spent 120 hours collecting material on and learning about the Java garbage-collection sub-system.
- I wrote a total of 5000 lines of source code, comprising a Linux device driver for a robot (in C) and a GUI (in Java) that is used to control it.
- I designed a new algorithm for computing the non-linear mapping from A-space to B-space using a genetic algorithm, see page 17.
- I implemented a version of the algorithm proposed by Jones and Smith in [6], see page 12, corrected a mistake in it, and compared the results with several alternatives.

Dedication and Acknowledgements

I am very fortunate to work with my superviser Rami Chehab, who assisted in my research greatly and inspired the vast quantity of this project. He has been a great mentor.

I would also like to thank my friends and family for always being there for me, especially when the going gets rough. You guys really helped make this university experience what it was.

Declaration

I declare that the work in this dissertation was carried out in accordance with the requirements of the University's Regulations and Code of Practice for Taught Programmes and that it has not been submitted for any other academic award. Except where indicated by specific reference in the text, this work is my own work. Work done in collaboration with, or with the assistance of others, is indicated as such. I have identified all material in this dissertation which is not my own work through appropriate referencing and acknowledgement. Where I have quoted or otherwise incorporated material which is the work of others, I have included the source in the references. Any views expressed in the dissertation, other than referenced material, are those of the author.

Joshua Felmeden, Thursday 7$^{\text{th}}$ April, 2022

Contents

List of Figures

List of Tables

Ethics Statement

A compulsory section

In almost every project, this will be one of the following statements:

- "This project did not require ethical review, as determined by my supervisor, [fill in name]"; or
- "This project fits within the scope of ethics application 0026, as reviewed by my supervisor, [fill in name]"; or
- "An ethics application for this project was reviewed and approved by the faculty research ethics committee as application [fill in number]".

See Section 3.2 of the unit Handbook for more information. If something went wrong and none of those three statements apply, then you should instead explain what happened.

Supporting Technologies

- I used a sample of headlines from Kaggle as training and validation data (https://www.kaggle.com/datasets/miguelaenlle/massive-stock-news-analysis-db-for-nlpbacktests)
- I used the Natural Language Toolkit to assist the preprocessing of data, using their English words and stop words data (https://www.nltk.org/)

Notation and Acronyms

NLP	:	Natural Language Processing
SESTM	:	Semantic Extraction via Screening and Topic Modelling
SESTM Specific Notation		
m	:	Number of words in sample
n	:	Number of articles in sample
$d_{i,j}$	:	Number of times word j appears in text i
$d_{[S],i}$	:	Subset of columns where the only indices are those with sentiment
$D = [d_1, \ldots, d_n]$	:	$m \times n$ Document term matrix
$sgn(y)$	:	Sign of returns of article y
$\hat{x}$	:	Expected value of variable x

Chapter 1

Introduction

Financial news is a widely available resource from which many investors and businesses alike glean information. It is an insight into how many different businesses function and the health of markets. If properly used, can be taken as a measure of ... Analysing the sentiment of an article has been used for a significant amount of time and is utilised for a myriad of purposes, including forecasting financial stocks.

An article is naturally formed of two parts, the headline and the article body. The bulk of the information conveyed by a given article is in the body, however, since headlines are often a summarisation of the body, it has been proven that data mining from the headline itself can be at least as useful as mining the body {citation needed}. The average word counts of a headline are very low, however, the vocabulary is often much higher impact on average, as this is what initially grabs the attention of a reader.

1.1 What to do

This chapter should introduce the project context and motivate each of the proposed aims and objectives. Ideally, it is written at a fairly high-level, and easily understood by a reader who is technically competent but not an expert in the topic itself.

In short, the goal is to answer three questions for the reader. First, what is the project topic, or problem being investigated? Second, why is the topic important, or rather why should the reader care about it? For example, why there is a need for this project (e.g., lack of similar software or deficiency in existing software), who will benefit from the project and in what way (e.g., end-users, or software developers) what work does the project build on and why is the selected approach either important and/or interesting (e.g., fills a gap in literature, applies results from another field to a new problem). Finally, what are the central challenges involved and why are they significant?

The chapter should conclude with a concise bullet point list that summarises the aims and objectives. For example:

> The high-level objective of this project is to reduce the performance gap between hardware and software implementations of modular arithmetic. More specifically, the concrete aims are:
>
> 1. Research and survey literature on public-key cryptography and identify the state of the art in exponentiation algorithms.
> 2. Improve the state of the art algorithm so that it can be used in an effective and flexible way on constrained devices.
> 3. Implement a framework for describing exponentiation algorithms and populate it with suitable examples from the literature on an ARM7 platform.
> 4. Use the framework to perform a study of algorithm performance in terms of time and space, and show the proposed improvements are worthwhile.

Chapter 2

Background

We begin by discussing the technical background that relate to the work in the thesis. This project evaluates and analyses the success of semantic analysis when applied to financial headlines; making use of realised stock returns as a teaching signal.

2.1 Semantic Analysis

Semantic analysis (also known as opinion mining) is the task of identifying opinions or sentiment of authors from an input of text. The ramifications of being able to programmatically extract the intended meaning of an input is extremely powerful in a variety of fields and is particularly prudent in a financial context. The nuances of natural language can sometimes make this difficult to extract, and therefore significant research has been conducted on the topic over the course of the last decades. The main fields of sentiment analysis are lexicon based, and deep learning based.

2.1.1 Lexicon Based Methods

The fundamental concept revolves around investigating a piece of text, and deciding on a binary classification: positive or negative. The simplest method compiles a list words with weights. Each weight corresponds to the positivity of the word (for example 'great' would have a high positivity, while 'terrible' would have very low). The overall sentiment of a piece of text could then be estimated by summing the individual sentiment scores of each word. The dictionary is not limited to single words, and can be expanded to include n-grams (phrases of n words), as the context in which a word is used can dramatically change the sentiment. This may increase the accuracy of the dictionary at the cost of increased dimensionality. The dictionary itself must be compiled before it is possible to utilise this method. One of the most widely used for English text is the Harvard-IV-4 TagNeg (H4N) lexicon which is a general usage model that can be used to estimate the sentiment of a piece of text.

There are many difficulties faced with creating a dictionary of this sort, as language is often a subjective entity, leading to conflicting opinions in assigning tone to a specific word. Furthermore, the context within which a word is used can drastically change the intended sentiment of a word, for example, the word 'great' is naturally a very positive word, however, if used in a sarcastic manner (e.g. 'It's so great that my flight is delayed!') can invert the sentiment entirely. For this reason, simply summing the sentiment of a piece of text on a word by word scale can give an incorrect estimate.

Certain words that may have no meaning at all in one context, may have significant sentiment in another, particularly in the field of finance, where jargon dominates text. Loughran and McDonald [1] conducted an investigation into the use of standard lexicons for use in analysing 10-K filing reports, which are comprehensive reports filed by companies about their financial performance. They discovered that almost 75% of negative words found in the filings based on the H4N file were not typically negative in a financial context. For example, 'liability' is a negatively charged word in a standard context, while it carries no tone at all in the context of the filings. Loughran and McDonald created their own dictionary based on these results called the Loughran McDonald dictionary that is created for the purpose of classifying 10-k filings. For these reasons, lexicon based methods work far better if the it is bespoke for the topic at hand, needing a specific dictionary for each task.

2.1.2 Deep Learning Methods

Deep learning is at the forefront of innovation within the field of computer science and is being applied to increasingly complex problems that were once thought impossible for machines to complete. Naturally, its application in sentiment analysis and NLP has also been explored to some success in recent years.

Discussing the intricacies of modern deep learning methods is out of the scope of this paper, but an overview is sufficient to understand the motives behind the techniques used. Simply, deep learning uses multiple layers of non-linear processing units for feature extraction. Lower layers learn simple features, while those higher up build on this understanding and learn more complex features. In NLP terms, the models need *word embedding* results as input features

2.2 Semantic Extraction via Screening and Topic Modelling

Semantic Extraction via Screening and topic model (SESTM) is a novel text mining algorithm that makes us of a teaching signal developed by Zhang et Al. in 2020. It makes use of stock returns as a teaching signal to develop a model of sentiment words in maximally positive or negative arguments in order to be able to predict the sentiment of out of sample articles.

2.2.1 Generating the sample

- Discuss aligning hte article with returns
  - Discuss private stocks and public stocks
  - Discuss non market days
- Discuss pre processing the data
- Discuss the kaggle sample

2.2.2 Screening for sentiment charged words

The first step in this algorithm is to screen for sentiment words in a collection of articles, since sentiment neutral words are a nuidance and contribute to noise. This strategy isolates the sentiment charged words and uses these alone to calculate sentiment. This is achieved by using a supervised approach with the realised returns of an associated stock, since if a word appears in headlines that result in positive returns, it is reasonable to assume that the word carries positive sentiment. Some notation will be introduced here to facilitate the discussion and explanation of the algorithm.

- The sample is defined as n articles producing a dictionary of m words.
- The word count of article i is recorded in vector d_i
- $D = [d_1, d_2, \ldots, d_n]$ is an $m \times n$ document term matrix
- The count of sentiment charged words in article i is defined as the submatrix $d_{[S],i}$.

For a word j, we define f_j as the fractional representation of the frequency with which a word appears in a positively tagged article versus the frequency with which a word appears in any article:

$$f_j = \frac{\text{count of word } j \text{ in article with } sgn(y) = +1}{\text{count of word } j \text{ in all articles}} \tag{2.1}$$

Here, sgn is simply the sign of the difference in tagged returns of an article. If an article is released on day t (specifically, between 4pm of day $t-1$ and 4pm of day t), the article is tagged with the returns of the associated firm from day $t-1$ to $t+1$ (specifically market close on day $t-2$ to close on day $t+1$).

Next, f_j is compared against positive and negative threshold. We defined $\hat{\pi}$ to be the fraction of articles that have $sgn(y) = +1$, or articles that are tagged with positive returns. In practice, as the sample size increases, this value tends towards 0.5. For sentiment neutral word, it is expected that $f_j \approx \hat{\pi}$ with some variance either side. Therefore, two thresholds are set α_+ and α_-. These thresholds are set such that any word with $f_j \geq \hat{\pi} + \alpha_+$ is determined to be sentiment positive, and the inverse for words such that $f_j \leq \hat{\pi} - \alpha_-$. This filtering technique accepts words that are further away from the expected average sentiment, meaning only those words with extreme sentiment are included. To ensure that words appearing infrequently do not skew sentiment values significantly (for example a word appearing in exactly one article that is maximally positive ending up with $f_j = 1$), we introduce a third parameter κ, restricting words that appear in fewer than κ headlines. The number of articles in which word j appears is defined as κ_j The set of sentiment charged words is therefore defined by:

$$S = \{j : f_j \geq \hat{\pi} + \alpha_+ \cup f_j \leq \hat{\pi} - \alpha_-\} \cap \{j : k_j \geq \kappa\} \tag{2.2}$$

The three parameters introduced here $(\alpha_+, \alpha_-, \kappa)$ are tuned via cross-validation and the best configuration for each is selected.

2.2.3 Learning Sentiment Topics

With the wordlist S calculated, we can now fit a two-topic model to each of the sentiment-charged counts. We introduce matrix $O = [O_+, O_-]$ that determines the expected probability of sentiment charged words in each article using a supervised learning approach. The teaching signal in this case is the realised stock returns of the associated firm for each headline.

In the model, p_i is the headline's sentiment score, described by:

$$p_i = \frac{\text{rank of } y_i \text{ in } \{y_l\}_{l=1}^{n}}{n} \tag{2.3}$$

Where the return of a headline is represented by y, and headlines are ranked in ascending order (meaning an article with the highest returns would have p_i of 1)

2.3 Portfolios and Financial Market Analysis

This project explores investing in portfolios and analyses the returns from the constructed portfolios. In this section, we introduce the concepts that will be utilised and discuss the motivations for using certain factors.

2.3.1 Portfolios

A portfolio is simply a list of stocks that can be invested in by an individual or firm. The returns from a portfolio is defined as the profit accrued from all stocks over a set time period, usually daily, monthly or annually. Due to the nature of stocks, simply listing the returns as a concrete value does not convey the information required. For example, if stock 'A' were to be invested in at value \$50, and it rose to \$60 the following day, the returns could be said to be \$10. However, if stock 'B' were valued at \$1000, and the following day it rose to \$1010, the monetary value would be equivalent at \$10, but the percentage return is vastly different; 20% returns for stock 'A' and 1%. For this reason, returns from portfolios are expressed as a percentage.

Daily returns can be very marginal, as the time period is very small, often being smaller than 1%. In the interest of readability, *basis points* (also known as bps or bips) are used in lieu of a percentage, where 1 bip is equivalent to 0.01%. This makes it much easier to represent very small returns as are common in daily returns.

Creating a portfolio

A portfolio is constructed using a number of stocks and can be either bought (taking the 'long' position) or sold (taking the 'short' position). For stocks that are bought, the returns can be calculated from the difference in price at the time that the stock is sold. More concretely, if a stock has value S_t at time t, and held for n days before being sold, the long returns in percentage form can be calculated using the following formula:

$$\frac{S_t}{S_{t+n}} - 1$$

Similarly, for short returns, as the stock is being sold, profit is acquired if the stock falls in value, therefore the returns can be calculated using the following formula:

$$\frac{S_{t+n}}{S_t} - 1$$

Of course, these simple formulae negect transaction fees that can apply when constructing real portfolios. However, as we are creating portfolios in a theoretical sense, this suits our needs.

Once the portfolio has been constructed, the weighting for each stock must be considered. Each portfolio will have a value, which is the amount of money invested into it, and each stock will in turn get an investment that is a percentage of this overall value. There are two strategies that we will consider: equal weighted and value weighted strategies. Equal weighted is very simple: if a portfolio is comprised of n stocks and has some investment v, each stock has v/n invested into it. This strategy glosses over differences in stock size or price. On the other hand, value weighted portfolios assign much more money to stocks with higher value associated to them. This can be calculated in a number of ways, but the way we calculate this is if stock s_i in portfolio $P = [s_1, ..., s_n]$ has market value $P_{i,t}$ at time t, the weight of stock s_i would be:

$$w_i = \frac{p_{i,t}}{\sum_n^1 p_{n,t}}$$

The amount invested into stock s_i would then be $w_i \times v$.

Equal weighted stocks more closely resemble hedge funds, as well as being the case that smaller companies are able to more quickly encapsulate market share and investor interest. Equal weighted investments ensure a portfolio has a higher representation of smaller stocks, at the higher risk of the stock failing. Conversely, value weighted portfolios tend to be safer, as they prioritise larger companies that are more stable. The downside to utilising this method is that the large percentage increases observed in smaller stocks will have less effect, and therefore some profit can be lost.

2.3.2 Evaluating a portfolio

To successfully determine the success of a portfolio, it is not always as simple as observing the profit alone. While this is a good indicator of the portential returns that could be gleaned from a given portfolio or investment method. The following methods are used to provide more insight into an investment method.

Sharpe Ratio

William Sharpe created the sharpe ratio in 1966 and is one of the most referenced comparison of risk versus return in finance. The formula for this ratio is exceedingly simple — one of the key factors in its wide usage and is as follows:

$$S(x) = \frac{r_x - R_f}{\sigma(r_x)}$$

where x is the investment, r_x is the average rate of return of x, R_f is the risk free rate of return, and σr_x is the standard deviation of r_x. The risk free rate of return is simply the theoretical rate of return on an investment with absolutely no risk. Subtracting these risk free returns from the average rate of returns of x yields the true rate of returns that introduced some risk.

The value of an investment's Sharpe ratio measures the performance with adjustment for risk: the higher the ratio, the better the performance of the investment when adjusted for risk. For comparison methods, a ratio of 1 or higher is good, 2 or better is very good, and 3 or better is excellent.

Fama French 3 and 5 Factor Models

Eugene Fama and Kenneth French co-authored a 1992 paper detailing risk factors in returns on both stocks and bonds. This extends the work Sharpe completed on the Sharpe ratio and goes further in exploring risk factors in returns, along with the capital asset pricing model (CAPM). CAPM is used for describing systematic risk and expected return, especially for that in stocks. The equation for this is:

$$ER_i = R_f + \beta_t(ER_m - R_f)$$

where ER_i is the expected return of the investment, R_f is the risk free rate, β_i is the beta of the investment and $ER_m - R_f$ is the market risk premium. The beta of an investment is the volatility compared to the rest of the market. It encompases the sensitivity of a stock to changes in the market. In essence, this gives the expected returns of an asset based on systematic risk. Building on this, Fama and French observed two additional risk factors: the size premium of an asset, or small minus big (SMB), and the value premium, or high minus low (HML). SMB is used to account for companies with small value stocks that generate high returns, while HML accommodates for stocks with equity that is valued cheaply compared to its book value that generate higher returns in comparison to the rest of the market. These factors are used in conjunction to provide the following formula for the Fama French 3 factor model (FF3 model):

$$ER_i = R_f + \beta_1(ER_m - R_f) + \beta_2(SMB) + \beta_3(HML) + \alpha$$

The values for SMB and HML are available from French's website, and can be collected for daily, monthly, or yearly returns. Computing this model on a series of returns from a portfolio gives useful information on the nature of the returns, since the model explains part of the returns. The values for the βs detail the exposure to exposure to each of the risk factors while the α, or the intercept refers to the amount that a portfolio outperformed the model. This alpha is representative of the amount of private, or new, information that is utilised to construct the portfolio. This means that more simplistic methods of selecting portfolios will not have any private information and therefore the intercept will be much smaller than that of a more complicated model as simple methods have profits that can be explained by these risk factors.

2.4 What to do

This chapter is intended to describe the background on which execution of the project depends. This may be a technical or a contextual background, or both. The goal is to provide a detailed explanation of the specific problem at hand, and existing work that is relevant (e.g., an existing algorithm that you use, alternative solutions proposed, supporting technologies).

Per the same advice in the handbook, note there is a subtly difference from this and a full-blown literature review (or survey). The latter might try to capture and organise (e.g., categorise somehow) *all* related work, potentially offering meta-analysis, whereas here the goal is simple to ensure the dissertation is self-contained. Put another way, after reading this chapter a non-expert reader should have obtained enough background to understand what *you* have done (by reading subsequent sections), then accurately assess your work against existing relevant related work. You might view an additional goal as giving the reader confidence that you are able to absorb, understand and clearly communicate highly technical material and to situate your work within existing literature.

Chapter 3

Project Execution

This chapter is intended to describe what you did: the goal is to explain the main activity or activities, of any type, which constituted your work during the project. The content is highly topic-specific, but for many projects it will make sense to split the chapter into two sections: one will discuss the design of something (e.g., some hardware or software, or an algorithm, or experiment), including any rationale or decisions made, and the other will discuss how this design was realised via some form of implementation.

This is, of course, far from ideal for *many* project topics. Some situations which clearly require a different approach include:

- In a project where asymptotic analysis of some algorithm is the goal, there is no real "design and implementation" in a traditional sense even though the activity of analysis is clearly within the remit of this chapter.
- In a project where analysis of some results is as major, or a more major goal than the implementation that produced them, it might be sensible to merge this chapter with the next one: the main activity is such that discussion of the results cannot be viewed separately.

Note that it is common to include evidence of "best practice" project management (e.g., use of version control, choice of programming language and so on). Rather than simply a rote list, make sure any such content is useful and/or informative in some way: for example, if there was a decision to be made then explain the trade-offs and implications involved.

3.1 Example Section

This is an example section; the following content is auto-generated dummy text. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel foo

Figure 3.1: This is an example figure.

foo	bar	baz
0	0	0
1	1	1
⋮	⋮	⋮
9	9	9

Table 3.1: This is an example table.

magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

3.1.1 Example Sub-section

This is an example sub-section; the following content is auto-generated dummy text. Notice the examples in Figure 3.1, Table 3.1, Algorithm 3.1 and Listing 3.1. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum

for $i = 0$ **upto** n **do**
  $t_i \leftarrow 0$
end

Algorithm 3.1: This is an example algorithm.

```
for( i = 0; i < n; i++ ) {
  t[ i ] = 0;
}
```

Listing 3.1: This is an example listing.

turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Example Sub-sub-section

This is an example sub-sub-section; the following content is auto-generated dummy text. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetuer id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

Nam dui ligula, fringilla a, euismod sodales, sollicitudin vel, wisi. Morbi auctor lorem non justo. Nam lacus libero, pretium at, lobortis vitae, ultricies et, tellus. Donec aliquet, tortor sed accumsan bibendum, erat ligula aliquet magna, vitae ornare odio metus a mi. Morbi ac orci et nisl hendrerit mollis. Suspendisse ut massa. Cras nec ante. Pellentesque a nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Aliquam tincidunt urna. Nulla ullamcorper vestibulum turpis. Pellentesque cursus luctus mauris.

Nulla malesuada porttitor diam. Donec felis erat, congue non, volutpat at, tincidunt tristique, libero. Vivamus viverra fermentum felis. Donec nonummy pellentesque ante. Phasellus adipiscing semper elit. Proin fermentum massa ac quam. Sed diam turpis, molestie vitae, placerat a, molestie nec, leo. Maecenas lacinia. Nam ipsum ligula, eleifend at, accumsan nec, suscipit a, ipsum. Morbi blandit ligula feugiat magna. Nunc eleifend consequat lorem. Sed lacinia nulla vitae enim. Pellentesque tincidunt purus vel magna. Integer non enim. Praesent euismod nunc eu purus. Donec bibendum quam in tellus. Nullam cursus pulvinar lectus. Donec et mi. Nam vulputate metus eu enim. Vestibulum pellentesque felis eu massa.

Quisque ullamcorper placerat ipsum. Cras nibh. Morbi vel justo vitae lacus tincidunt ultrices. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. In hac habitasse platea dictumst. Integer tempus convallis augue. Etiam facilisis. Nunc elementum fermentum wisi. Aenean placerat. Ut imperdiet, enim sed gravida sollicitudin, felis odio placerat quam, ac pulvinar elit purus eget enim. Nunc vitae tortor. Proin tempus nibh sit amet nisl. Vivamus quis tortor vitae risus porta vehicula.

Fusce mauris. Vestibulum luctus nibh at lectus. Sed bibendum, nulla a faucibus semper, leo velit ultricies tellus, ac venenatis arcu wisi vel nisl. Vestibulum diam. Aliquam pellentesque, augue quis sagittis posuere, turpis lacus congue quam, in hendrerit risus eros eget felis. Maecenas eget erat in sapien mattis porttitor. Vestibulum porttitor. Nulla facilisi. Sed a turpis eu lacus commodo facilisis. Morbi fringilla, wisi in dignissim interdum, justo lectus sagittis dui, et vehicula libero dui cursus dui. Mauris tempor ligula sed lacus. Duis cursus enim ut augue. Cras ac magna. Cras nulla. Nulla egestas. Curabitur a leo. Quisque egestas wisi eget nunc. Nam feugiat lacus vel est. Curabitur consectetuer.

Suspendisse vel felis. Ut lorem lorem, interdum eu, tincidunt sit amet, laoreet vitae, arcu. Aenean faucibus pede eu ante. Praesent enim elit, rutrum at, molestie non, nonummy vel, nisl. Ut lectus eros, malesuada sit amet, fermentum eu, sodales cursus, magna. Donec eu purus. Quisque vehicula, urna sed ultricies auctor, pede lorem egestas dui, et convallis elit erat sed nulla. Donec luctus. Curabitur et nunc. Aliquam dolor odio, commodo pretium, ultricies non, pharetra in, velit. Integer arcu est, nonummy in, fermentum faucibus, egestas vel, odio.

Sed commodo posuere pede. Mauris ut est. Ut quis purus. Sed ac odio. Sed vehicula hendrerit sem. Duis non odio. Morbi ut dui. Sed accumsan risus eget odio. In hac habitasse platea dictumst. Pellentesque non elit. Fusce sed justo eu urna porta tincidunt. Mauris felis odio, sollicitudin sed, volutpat a, ornare ac, erat. Morbi quis dolor. Donec pellentesque, erat ac sagittis semper, nunc dui lobortis purus, quis congue purus metus ultricies tellus. Proin et quam. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Praesent sapien turpis, fermentum vel, eleifend faucibus, vehicula eu, lacus.

Example paragraph. This is an example paragraph; note the trailing full-stop in the title, which is intended to ensure it does not run into the text.

Chapter 4

Critical Evaluation

4.1 Unorganised Figures

4.1.1 Daily returns

	Sharpe	Average	Daily	FF3		FF5		FF5 + MOM	
Formation	Ratio	Return	Turnover	α	R^2	α	R^2	α	R^2
EW L-S	1.82	16.48	95.71%	15.00	2.09%	15.43	3.27%		
EW L	0.86	9.14	95.49%	8.77	2.88%	8.56	2.60%		
EW S	0.57	7.34	95.93%	6.23	1.68%	6.17	1.97%		
VW L-S	1.97	30.13	94.99%	27.16	2.21%	27.19	4.86%		
VW L	0.91	12.00	94.67%	10.83	6.29%	10.52	6.21%		
VW S	1.11	18.13	95.30%	16.34	1.16%	16.70	2.21%		

Table 4.1: Performance of Daily News Sentiment Portfolios

Table 4.1 describes the

4.1.2 Most impactful words

Following the construction of matrix O, figure 4.1 demonstrates the list of sentiment charged words on average over all training 19 windows. At each training and validation window, the sentiment lists are generated completely from scratch, and while there is some overlap, each list can vary significantly. The font size corresponds to the average tone (calculated by $\frac{1}{2}(O_+ - O_-)$) of the words across all windows. Of the top 50 positive sentiment words, the following appeared in more than 75% of windows, with words highlighted in **bold** appearing in all windows:

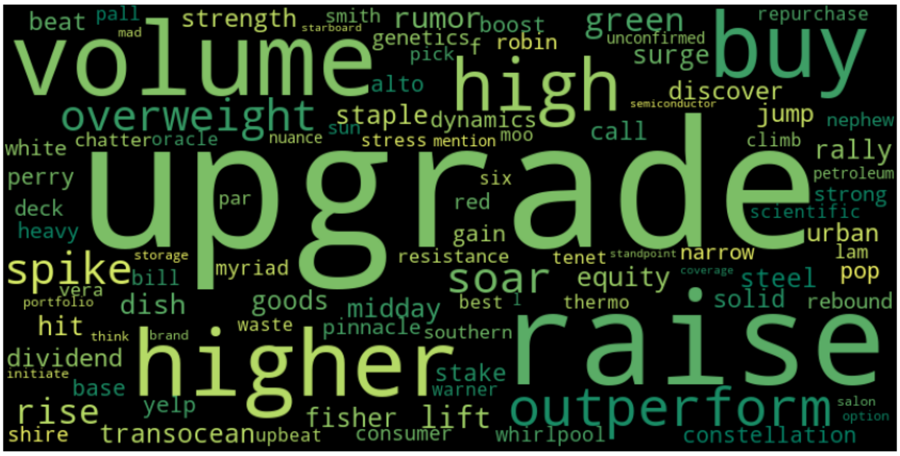
soar, raise, high, volume, **upgrade**

The following words are the equivalent with respect to top 50 negative sentiment words:

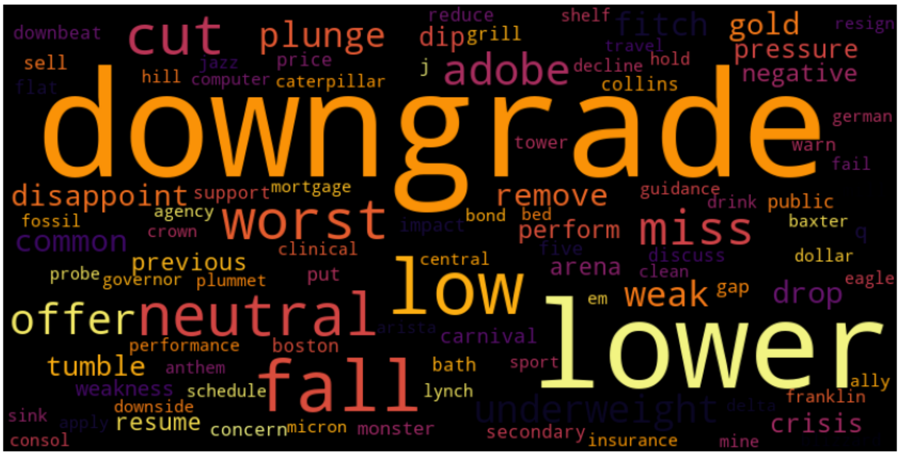
disappoint, fall, cut, plunge, low, miss, weak, **lower, downgrade, underweight**

Simply by inspection, each group appears sane, in the sense that many of the words with high values in either sentiment could be assumed. However, some words are somewhat surprising and this may offer an insight into subconsious bias that exists in writing headlines as opposed to article bodies. For example, the word *volume* is, under normal circumstances, a sentiment neutral word, but according to the model generated by SESTM, is a highly positive word. Examples of headlines including this word include:

- *Agilent spikes to high of $60.40 on Volume*
- *Markets gather some momentum as volume remains light, geopolitical tension improving*
- *Tuesday's Mid-day options for Volume Leaders*



(a) Positive words



(b) Negative words

Figure 4.1: Word clouds demonstrating sentiment charged words. Font size corresponds to average tone across all training samples

Included in the sample are headlines from ‘Benzinga’, which is a company that offers realtime news articles, and has a significant quantity of headlines of the form *Benzinga’s top upgrades* and *Benzinga’s volume*

The headline of a news article is often packed with more impactful words to grab the attention of a reader

4.1.3 Speed of information Assimilation

4.1.4 Comparison to other methods

4.2 What to do

A topic-specific chapter

This chapter is intended to evaluate what you did. The content is highly topic-specific, but for many projects will have flavours of the following:

1. functional testing, including analysis and explanation of failure cases,
2. behavioural testing, often including analysis of any results that draw some form of conclusion wrt. the aims and objectives, and

3. evaluation of options and decisions within the project, and/or a comparison with alternatives.

This chapter often acts to differentiate project quality: even if the work completed is of a high technical quality, critical yet objective evaluation and comparison of the outcomes is crucial. In essence, the reader wants to learn something, so the worst examples amount to simple statements of fact (e.g., "graph X shows the result is Y"); the best examples are analytical and exploratory (e.g., "graph X shows the result is Y, which means Z; this contradicts [1], which may be because I use a different assumption"). As such, both positive *and* negative outcomes are valid *if* presented in a suitable manner.

Chapter 5

Conclusion

The concluding chapter of a dissertation is often underutilised because it is too often left too close to the deadline: it is important to allocate enough attention to it. Ideally, the chapter will consist of three parts:

1. (Re)summarise the main contributions and achievements, in essence summing up the content.
2. Clearly state the current project status (e.g., “X is working, Y is not”) and evaluate what has been achieved with respect to the initial aims and objectives (e.g., “I completed aim X outlined previously, the evidence for this is within Chapter Y”). There is no problem including aims which were not completed, but it is important to evaluate and/or justify why this is the case.
3. Outline any open problems or future plans. Rather than treat this only as an exercise in what you *could* have done given more time, try to focus on any unexplored options or interesting outcomes (e.g., “my experiment for X gave counter-intuitive results, this could be because Y and would form an interesting area for further study” or “users found feature Z of my software difficult to use, which is obvious in hindsight but not during at design stage; to resolve this, I could clearly apply the technique of Smith [7]”).

Bibliography

[1] TIM LOUGHRAN and BILL MCDONALD. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, 66(1):35–65, 2011.

Appendix A

An Example Appendix

Content which is not central to, but may enhance the dissertation can be included in one or more appendices; examples include, but are not limited to

- lengthy mathematical proofs, numerical or graphical results which are summarised in the main body,
- sample or example calculations, and
- results of user studies or questionnaires.

Note that in line with most research conferences, the marking panel is not obliged to read such appendices. The point of including them is to serve as an additional reference if and only if the marker needs it in order to check something in the main text. For example, the marker might check a program listing in an appendix if they think the description in the main dissertation is ambiguous.