

**Sentiment Analysis of News Articles for Stock Price Prediction**

By

**Bolanle Esther Onifade**

**(beo01u)**

**School of Computer Science**

**University of Nottingham**

**Supervisor: Michel Valstar**

**Signature: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

**Date: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

# Abstract

Several theories regarding the behaviour of the stock market exist. These theories put forward ideas ranging from those claiming that the stock market cannot be predicted and to those claiming those that with the right assumptions, the stock market can be predicted. Based purely on numbers, it’s hard to see how hard a predictor can predict the stock market as often the numbers are reflective of non-numerical factors. In this dissertation we aim to consider those non-numerical factors that affect the price of specific stock. These non-numerical factors come in form of news articles. In the digital age of rapid response to ongoing situations, news articles about events tend to be released as soon as they occur and the assumption is that if we can track news sources, monitoring them for the release of relevant news articles, we can use the sentiment orientation of the article (whether positive, negative or neutral) to predict the price of the stock market before the market has a chance to react to the article. This of course poses an important question on how the sentiment orientation of articles. There are several approaches that can be taken towards determining the class of articles but for the purposes of this dissertation, we will focus on the use of Support Vector Machines to perform classification. The output of classification will then be fed into a second layer of support vector machines to perform the price prediction.

# Acknowledgements

I would like to firstly thank my supervisor Michel Valstar for the encouragement to explore areas of analysis that I otherwise wouldn’t have thanked. Helping me think outside the box has helped take the dissertation much farther than I thought it would go. It would so be very poor form to neglect to thank Dr Robert Young who has helped me with aspects of finance and economics that I have found difficult to wrap my head around. I also would like to thank the few anonymous students from business school that helped with labelling the news articles.

To my parents, sisters and friends, I would like to say a massive thanks for all these years of supporting me through my education. A lot of time has passed since I started school 18 years ago and this dissertation is a representation of all the knowledge I have gained in that time. I truly appreciate all the support you’ve all expressed for me and this project.

**Table of Contents**

[Abstract 2](#_Toc404951180)

[Acknowledgements 3](#_Toc404951181)

[Table of Figures 5](#_Toc404951182)

[1. Introduction 6](#_Toc404951183)

[2. Literature Review 7](#_Toc404951184)

[3. Background Knowledge 8](#_Toc404951185)

[4. Scientific Method 9](#_Toc404951186)

[4.1. Method Overview 9](#_Toc404951187)

[4.2. Sentiment Classification 9](#_Toc404951188)

[4.2.1. Data Acquisition 9](#_Toc404951189)

[4.2.2. Data Labelling 10](#_Toc404951190)

[4.2.3. Data Pre-processing 10](#_Toc404951191)

[5. Experimentation and Evaluation 11](#_Toc404951192)

[6. Conclusion 12](#_Toc404951193)

# Table of Figures

[Figure 1 – XML Format for Scrapped News Articles 10](file:///\\store\beo01u$\Private\Dissertation.docx#_Toc404951167)

# Introduction

# Literature Review

# Background Knowledge

# Scientific Method

## Method Overview

## Sentiment Classification

As emphasised in previous sections, the first task to be performed is the classification of news articles. As with classification of other types of data, the following steps need to be performed: data acquisition, data labelling, data pre-processing, data analysis and then finally classification. The range of human sentiment is very wide and includes sentiment such as happiness, sadness, calmness, anger, anxious. This range is much too wide for the application at hand. In fact, we have taken a much simpler approach and simply classified the sentiment of the news articles into three categories: happy, sad and ambivalent.

Unlike sentiment classification for domains such as films, cars or music, happiness, sadness and ambivalent news articles may not bear much information about the progress of the entity. While the sentiment of the article is what we aim to extract, a cursory look at any news article that bears financial information will show that news articles aren’t very sentimental. This indicates that classification based simply on human sentiment while might be accurate might not be as successful given that we aim to predict the stock market. Therefore, to supplement classification based on sentiment, we also classify based on the progression of the company. This means that the articles get classified into an additional set of categories (positive, negative and neutral).The aim with the progress classification is that articles get classified based on what the news article evaluator expects that the entity’s stock price will go up or down or simply stay the same.

This new direction of evaluation however raises the question of how to gauge the effect of a news article. For illustrative purposes, if the price of a specific stock has been on the rise for the past three days and then a news article is released and it’s classified as “positive”, how do we factor that in? Does it simply not matter as we have a direction of progress or could we instead watch for the rate of the change of the stock price? We aim to answer some of these questions in the following sections and the over the next chapter.

### Data Acquisition

It’s clear that the very first task to be performed is the acquisition of news articles whether labelled or non-labelled. Although a fair bit of work have been done using this particular approach to stock price prediction, we were unable to find any publicly available datasets that fell in line with the purposes of this dissertation. Hence, a dataset was generated using from online news sources.

Although selecting news sources seems like a trivial task, it requires careful consideration as the news sources has to be able to satisfy the following requirements:

1. Has to be popularly read, especially by traders. This is particularly important because a high level of trust needs to be placed in the news source, enough to determine that significant changes in stock price trend will be reflected in the news articles.
2. The news sources has to have decent coverage of the news sources, again to ensure that we gather as much data as possible.

These requirements are particularly important especially when we consider that we will be aligning news articles with stock prices. Selecting news sources that report events a few days after the fact might skew results as by then the market is sure to have absorbed the new information and any relevant changes in (rate of change of ) price will be missed. Given these requirements, investors on online forums (as well as individuals with knowledge of finance) were asked which news sources were read and the following sources were given: Reuters, Bloomberg, Financial Times, Market Watch, Yahoo Finance.

The next step is to scrap selected websites (Bloomberg, Reuters) for news articles. We do not discuss the exact process of scraping websites as it’s not relevant to this project. However, scraping can involve interesting problems such as logging in to websites via a program (in this case, python) and extracting data. The zip file accompanying this document contains all code for scraping the websites.

Finally, the news data is extracted and put in the following xml format:

<?xml version='1.0' encoding='us-ascii'?>

<news>

<entry author=''…'' datetime=''…'' url=''…''>

<headline>

…

</headline>

<body>

…

</body>

</entry>

</news>

Figure 1 – XML Format for Scrapped News Articles

### Data Labelling

At the start of the project, the intention was to crowdsource the labelling of the articles. This would be done by asking individuals with knowledge of economics and finance to evaluate the news articles. This involved uploading the corpus to a website for easier classification. (sentimentanalysis.bolanleonifade.me). However, the rate at which the articles were getting classified was very slow so the alternative approach taken was to use prior knowledge to label the article. This of course meant that experiments might suffer due to lack of enough financial knowledge. Therefore, in order to provide a baseline or at the very least, a means of evaluating the manually labelled data, a set of automatically labelled data was created as well.

* + - 1. **Manual Labelling**

Manual labelling of data is simply reading each news article and labelling them by hand. The evaluators are asked to estimate the company’s progression based on the news article. Their estimates can fall into the following categories: (up, down, neutral). The evaluators are also asked to provide the sentiment of the article (happy, sad, neutral). From henceforth, for clarity purposes, we shall refer to the former as “progress sentiment” and the latter as “feeling sentiment”

One might think that there is a perfect correlation between the two sets of categories. However, there can be differences between the two. For illustrative purposes, we will examine a few cases in which there are differences:

The headline “Exxon Mobil reports Fire, oil spill at Nigeria’s terminal”, evokes a feeling of sadness but due to the established nature of Exxon, it’s unlikely that this event is going to lead to a massive dent in the stock price, we give the article a progress sentiment of neutral (because the article doesn’t go on to indicate that Exxon will suffer from this incident). Another article that wouldn’t be expected to change the stock price much is “JP Morgan falls to death from building roof in Hong Kong”. It’s however clear that the article is “sad”

If the previous two examples give the impression that from headlines, we can always tell the feeling sentiment of an article, it would be wrong. In fact, a seemling neutral headline such as ”Coca-cola names Waller Finance Chief as Fayard Retires” goes on to discuss the recent struggles of coca-cola, therefore giving it a feeling sentiment of sad and a progress sentiment of neutral. In the same strain, we discovered articles can both be up for progress sentiment and feeling sentiment; this would be the case for articles that discuss an entity’s growing business.

Given these apparent differences in the labelling of the progress and feeling sentiment, one might be concerned that the one of those sentiments is useless. We find out that this isn’t the case when the perform analysis on the dataset (Section 5). In addition, we find out

* + - 1. **Automatic Labelling**

In order to perform automatic labelling of news articles, we need to generate projected trends, this gives us an idea of the overall outlook of the stock price – that is, for example, we can safely say that the overall projected trend of the stock price is an upwards movement if the price over a period of time has changed positively, ignore every minor dip in the price trend. We can perform automatic labelling by using piecewise linear approximation (described in section 3 and the results of which are discussed in section 5) and proposed by Fung et al. (2005). This allows us to align news articles with the projected stock price and simply labelling the articles based on the projected stock price.

There are obvious inaccuracies that can occur from the use of such a method – in fact, as shown in the literature review, classification based purely on price differences, tends not to be very accurate – this is because one cannot say for sure that all articles released during periods of overall upwards price movement are positive and vice versa. However, we are operating under the assumption that news articles strongly reflect the direction of movement of the stock market. We expect the price to move up when news articles discuss increases in sales, innovation, positive restructuring and we expect the price to move downwards when news articles discuss fines, bankruptcy, legal problems, sanctions etc.

Automatic labelling, therefore provides us with a baseline. The more similar the results of automatic labelling is to those of manual labelling, the more “trust”, we can place in the results of manual labelling.

Another issue of note with automatic labelling is the fact that automatic labelling cannot be classified based on feeling sentiment. Feeling sentiment by definition requires an evaluator to label articles based on what feelings are evoked by reading the article. However, since the method by which we automatically label stock data is based on the progression of the stock price, we can automatically label articles based on stock price trends.

* + - 1. **Evaluating Labelled News Articles**

Sentiment labelling generated via automatic categorisation is a reflection of the price movements, not a reflection of the articles themselves. However, since the articles themselves are manually labelled to reflect precisely the sentiment which they carry, we can conclude that if there exists a high similarity the results of manual labelling is to the results of automatic labelling, then we can say confidently that the labelled articles can led to positive results in later classification. It’s important to note however, that any comparison that we do can only be by comparing progress sentiment of automatically labelled articles with progress sentiment of manually labelled articles (as opposed to both progress and feeling sentiment of manually labelled articles).

How therefore do we evaluate the results of labelling? An easy method of doing is is by calculating the pearson’s correlation coefficient (already discussed in section 3). We comprehensively discuss the results of the calculations and the other considerations (specific to calculating the correlation) when discussing the evaluations in section 5). We finish off this section by pointing out that the more similar the correlation values are between the projected trend line (generated via piecewise linear approximation) and the actual stock price trend line, the higher the similarity is between the two trends. Hence, this also provides additional validation for the results of piecewise linear segmentation.

### Data Pre-processing

When we first retrieve data and to display them, the news articles are kept in HTML format. Th

# Experimentation and Evaluation

# Conclusion