

**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](#_Toc409116450)

[Acknowledgements 3](#_Toc409116451)

[Table of Figures 6](#_Toc409116452)

[1. Introduction 7](#_Toc409116453)

[2. Literature Review 8](#_Toc409116454)

[3. Background Knowledge 9](#_Toc409116455)

[4. Scientific Method 10](#_Toc409116456)

[4.1. Method Overview 10](#_Toc409116457)

[4.2. Sentiment Classification 10](#_Toc409116458)

[4.2.1. Data Acquisition 10](#_Toc409116459)

[4.2.2. Data Labelling 12](#_Toc409116460)

[4.2.3. Data Pre-processing 14](#_Toc409116461)

[4.2.4. Document Representation 14](#_Toc409116462)

[4.2.5. Feature Selection or Reduction 14](#_Toc409116463)

[4.2.6. Classification 15](#_Toc409116464)

[4.3. Price Prediction 15](#_Toc409116465)

[5. Evaluation and Results 15](#_Toc409116466)

[5.1. Method Overview 15](#_Toc409116467)

[5.2. Data Acquisition 15](#_Toc409116468)

[5.2.1. News Article Acquisition 15](#_Toc409116469)

[5.2.2. Stock Data Acquisition 16](#_Toc409116470)

[5.3. Labelling 16](#_Toc409116471)

[5.3.1. Manual Labelling 16](#_Toc409116472)

[5.3.2. Automatic Labelling 18](#_Toc409116473)

[5.3.3. Labelling Discussion 20](#_Toc409116474)

[5.4. Data Pre-processing 20](#_Toc409116475)

[5.5. Feature Selection or Feature Reduction 21](#_Toc409116476)

[5.6. Document Classification 22](#_Toc409116477)

[5.6.1. Manual Classification 22](#_Toc409116478)

[5.6.2. Automatic Classification 27](#_Toc409116479)

[6. Conclusion 28](#_Toc409116480)

[6.1. Keys for Transforming Sentiments to Numbers 28](#_Toc409116481)

[7. Appendix 29](#_Toc409116482)

# Table of Figures

[Figure 4.1 – XML Format for Scrapped News Articles 11](file:///C:\Users\Ester\Documents\Fourth%20Year\G54MIP\Documents\Dissertation.docx#_Toc409116431)

[Figure 5.1 – Number of articles collected for each company 15](#_Toc409116432)

[Figure 5.2 – Table of correlation between sentiment and stock price using manually labelled data 16](#_Toc409116433)

[Figure 5.3 – Actual Stock Price and Projected Price of Disney 18](#_Toc409116434)

[Figure 5.4 – Actual Stock Price and Projected Stock Price of Exxon 19](#_Toc409116435)

[Figure 5.5 – Aligned News Articles with Trends 19](#_Toc409116436)

[Figure 5.6 – Correlation values of the automatically generated data 20](#_Toc409116437)

[Figure 5.7 – Initial features and pre-selected features 21](#_Toc409116438)

[Figure 5.8 - Number of Features depending on whether feature selection or extraction is used 21](#_Toc409116439)

[Figure 5.9 – Words selected for progress classification (manually labelled data) 22](#_Toc409116440)

[Figure 5.10 – Words selected for feeling classification(manually labelled data) 22](#_Toc409116441)

[Figure 5.11 – support for the various classes (manual/progress) 23](#_Toc409116442)

[Figure 5.12 – Confusion matrices (Manual/Progress). Top left – Unigram, Top Right – Bigram, Bottom – Unigram + Bigram 24](#_Toc409116443)

[Figure 5.13 – Table of performance of linear SVM measured by cross validation (manual/progress) 24](#_Toc409116444)

[Figure 5.14 – Support for the classes (Manual/ Feeling) 25](#_Toc409116445)

[Figure 5.15 – Confusion matrices (Manual/ Feeling). Top left – Unigram, Top Right – Bigram, Bottom – Unigram + Bigram 26](#_Toc409116446)

[Figure 5.16 –Table of performance of linear SVM measured by cross validation (Manual/ Feeling) 26](#_Toc409116447)

[Figure 5.17 – Support for the classes (Automatic / Progress) 27](#_Toc409116448)

[Figure 5.18 - Table of performance of linear SVM measured by cross validation (Automatic/ Feeling) 27](#_Toc409116449)

# 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

#### News Articles 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. We do not use all of this data however because of time constraints – there’s no way for a single person to manually label the 12000 articles in the time frame of 4 weeks. Hence, we had to discard a lot of the news articles and aim for classifying a fraction of the news articles (about 2000 articles) – we note also that the number actual number of articles which are used during classification is further reduced to 1690 after preprocessing – this is further discussed in section 4.2.3. The zip file accompanying this document contains all code for scraping the websites.

Prior to the scraping of news articles, we must first determine what it is we hope to find – in this case, we want to scrap enough news articles in order to perform classification on the news articles. Hence the gathering of news articles has to be targeted so that we have enough data for each of the companies we aim to classify. There is therefore a need to determine which companies we aim to classify. The companies selected were chosen from the Dow Jones Industrial Average (DJIA) because the companies listed on the index are major American companies which tend to get a lot of attention from the media.

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 4.1 – XML Format for Scrapped News Articles

#### Stock Price Data Acquisition

The acquisition of the stock data simply involves selecting the companies which are of interest and extracting the data from Yahoo Finance. In this case, the API (ystockquote) was used to extract the data for relevant companies. The data collected for each company includes the following: Date, Adjusted Close, Close, High, Low, Open, and Volume. As interesting as the use of the intraday stock price is, there are other complex aspects of the project and the introduction of intraday stock prices would mean the consideration of other aspects of both finance and implementation details that would only increase the difficulty of the project. An example of such an issue would be determining the period after which we can say a news article’s effect has taken place. On a day-by-day basis, we assume that at the end of the working day, the effects of a news article should already be reflected in the movement of the stock price. In addition, a day-by-day basis means that any projected trend line generated does not have to account for the minute movements during the day. [Check back on this.]

### Data Labelling

At the start of the project, the intention was to crowd source 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. The news articles therefore need to be tokenised (into unigrams and bigrams), stemmed and have stop words removed (section 3) before classification can take place. All these methods have been extensively discussed in section 3 but we summarise them here.

Tokenisationis the splitting of words into smaller groups whilst ignoring all punctuation. We perform tokenisation of news articles into unigrams, bigrams and unigrams and bigrams. In order to perform tokenisation we use nltk’s word\_tokenize. Before tokenisation, we ensure that the words all have the same casing (in this case, lowercase). Stemming is performed using nltk’s implementation of the Porter stemmer. This allows for removal of words that mean the same. Stop words don’t bear enough information to be useful for classification purposes and these are removed by using the scikit-learn in-built capabilities.

After tokenisation, feature extraction from the corpus must take place. Tokenising – the process by which features are extracted – 2242 separate documents can lead to quite a lot of features, most of which are unlikely to be useful for classification even when stop word removal is taken into account. Therefore, a step that’s carried out in pre-processing is selecting the max\_features based on term frequency is a parameter used in order to select only the features that are most likely to be informative during the classification process.

### Document Representation

After the completion of all pre-processing steps, the documents are now ready to be transformed into vectors. The library Scikit-learn provides the TfidfVectorizer that converts the news articles into tf-idf-weighted document-term matrix. Tf-idf has been discussed in (section 3). The process thereby transforms document into number – ready to be used by a classifier – after feature selection or feature reduction. Using all the features can lead to very low accuracy so in this step, we already pre-select the number of features to use. The number of features we select depends on tokens of the documents (bigrams, unigrams or both). It’s important to note that this is not feature selection as is performed later on; this pre-selection is implemented by scikit-learn and allows us to pre-select only the highest tf-idf scores. We discuss the results in section 5.

### Feature Selection or Reduction

In the literature, the chi-squared method for feature selection and the SVD method are popular and we decided to use test these two methods against each other. We very briefly compare the results of both methods in section 5. Section 3 provides an in-depth explanation of both methods. The library Scikit-learn provides the class TruncatedSVD class as well as chi2 function for performing feature selection and feature reduction respectively. SVD is a very expensive method that in practice using our relatively small corpus of approximately 2500 document resulted in training time of about 2 hours. Clearly this is unacceptable for most purposes and while the compare the results of both methods all things being equal, chi2 is much less time consuming giving it an edge over SVD.

### Classification

The finally step is to train the SVM to predict the news article. We assume that the problem is linear (and instead we briefly compared the linear SVM with other types) and the linear SVM seemed to perform best (discussed in section 3). In order to truly evaluate the SVM, cross validation over the data set was performed. 10 fold cross validation ensured we got an average of all performance of the hyper-parameters of the SVM. The results of each fold were evaluated using the following metrics: confusion matrix, recall, precision and f-measure. These results can then be averaged over the number of folds to determine the overall performance of the hyper-parameters. This therefore completes our discussion of the process of sentiment classification of the news articles. Section 5 details the results.

## Price Prediction

# Evaluation and Results

## Method Overview

In this section, we use detail the results of each of the main activities that have been discussed in section 4. The format of this section is the same as that of section 4, this is for easy referencing and comprehension. As there are a lot of possible data, we try as much as possible to reduce the data to only that which illustrates the general idea or result and any redundant but interesting information is put in the appendix.

## Data Acquisition

### News Article Acquisition

As we have already discussed in section 4.2.1.1 the exact process of acquisition, in this section we simply delve into the results. The table below shows the number of articles that were gathered for each company. Please note that the numbers reflected here are numbers after pre-processing had been performed.

|  |  |  |
| --- | --- | --- |
| Company Name | Number of Articles | |
| Chevron | | 88 |
| Cocacola | | 52 |
| Disney | | 108 |
| Exxon | | 120 |
| Goldman | | 731 |
| IBM | | 118 |
| JP Morgan | | 613 |
| Microsoft | | 259 |
| Pfizer | | 115 |
| Visa | | 48 |

Figure 5.1 – Number of articles collected for each company

### Stock Data Acquisition

The price daily values were collected for the period from the 1st of January, 2013 to the 30th of September, 2014. Of course, the stock price data is only released for working days so this accounts for only 440 working days.

## Labelling

### Manual Labelling

We show the results of manual labelling in this section. Principally, this comes in the form of the Pearson’s correlation coefficient carried out between the stock price and the aggregate sentiment. The aggregate sentiment is calculated simply by adding the sentiment for every previous day in the time series. In order to do this, we need to calculate the aggregate sentiment for each working day. This is simply done by adding together all the sentiment values for each article released in the working day.

Supposing therefore that we have 4 articles released on any day labelled as such: (please refer to the appendix – section 7.1 – for how this is derived from the sentiment labels) using progress sentiment. The aggregate sentiment would be given as by adding all the sentiments together. In addition to this, supposing we have any set of 8 days for which the aggregate sentiment for each day is given, we have aggregate sentiments for each day to be the sum of previous sentiments:

We’ve aggregated the sentiments to better show the rise and fall of the sentiment (gotten from the news articles). With this transformed sentiment, we calculate the correlation between the stock price closing values and the sentiment. The table below shows the correlation values between each company’s stock data and the aggregated sentiments.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Company Name | Correlation of Progress Sentiment(Actual) | | Correlation of Feeling Sentiment(Actual) | Correlation of Progress Sentiment(Projected) | Correlation of Feeling Sentiment(Projected) |
| Chevron | | 51.7 | 51.8 | 49.7 | 50.1 |
| Cocacola | | 32.8 | 32.6 | 32.99 | 32.84 |
| Disney | | 97.6 | 97.6 | 97.10 | 97.13 |
| Exxon | | 79.0 | 78.8 | 79.7 | 79.6 |
| Goldman | | 76.74 | 77.42 | 71.98 | 72.5 |
| IBM | | 53.2 | 53.3 | 51.7 | 51.8 |
| JP Morgan | | 84.2 | 84.4 | 84.5 | 84.7 |
| Microsoft | | 95.8 | 95.8 | 95.8 | 95.8 |
| Pfizer | | 51.1 | 51.6 | 50.9 | 51.3 |
| Visa | | 89.14 | 89.20 | 88.80 | 88.86 |

Figure 5.2 – Table of correlation between sentiment and stock price using manually labelled data

Looking at the table above as well as figure 5.1, there are several points raised regarding the values shown. We attempt to discuss some of them here but leave others until we show the results of automatic labelling. In addition, we also generate correlation coefficients between the projected trend lines and the sentiment – for purposes of comparison with the results of automatic labelling. We have chosen to a few of the figures that show the results of piecewise linear approximations in the succeeding section as the results are more relevant there. However, please refer to the appendix for the rest of the figures. [Appendix Reference]

We see that in almost all cases, the values of the correlation projected values are almost always lower than the correlation between the actual stock prices. It is expected that there would be a difference between the correlation between the two pairs of values but assuming that piecewise linear approximation generates trends as accurately as possible, the difference between the correlation values should never be large enough to raise questions. Exxon as we can see has correlation coefficients that are larger in the projected trend line than in the actual trend line. A possible reason for this is that news articles capture better changes in prices over a longer period of time than the daily changes in price. [Check this out later]

In addition, we can see that although feeling sentiment and progress sentiment are intended to be different measures of labelling articles, they lead to in all cases, similar correlation coefficients. However, we note that in almost all cases, progress sentiment has a higher correlation value than projected sentiment.

Finally, we address the fact that something must be done for those days in which news articles are not released (this applies as well when calculating the correlation for automatically labelled data). There are three assumptions we can make when classifying based on progress:

1. There is an underlying feeling of positivity – that is, these cases, the stock price is assumed to go up meaning that the progress sentiment is always positive
2. There is an underlying feeling of neutrality – the stock price will stay the same when there’s no news.
3. An underlying feeling of neutrality – the stock price will go down when there’s no news

This can of course have drastic effects on the correlation values because in most cases, the number of articles released is significantly lower than the number of days in the time period. While it might seem that an underlying feeling of neutrality is the most appropriate, this is not true. The underlying feeling depends on the company being discussed. Prices of most companies change positively when there’s no news hence we assume an underlying feeling of positively. However, for companies such as IBM, there’s an underlying feeling of negatively. This is determined by simply looking at the overall projected trend line. IBM over the course of the time period has shown a gradual decrease in price. This is sentiment is well reflected in articles as IBM over the course of the time period had trouble keeping up with other technology firms who have incorporated cloud computing into the services offered. The complete set of trend lines are in the appendix – [Appendix Reference].

News articles aren’t released solely during weekdays; however stock markets are closed over the weekend. This doesn’t mean that trading does not occur over the weekend – extended hours trading does – however prices changes aren’t released for the weekend. Hence, in order to calculate the correlation, we have to eliminate news articles that are released on days that fall on weekends. This of course is detrimental to our calculations but despite this we see that there still is a high correlation between stock prices and news articles.

### Automatic Labelling

The first few steps needed to be carried out are similar to the steps carried out for manual labelling – summarily, calculating the aggregate sentiment of the news articles over the time period. As previously mentioned in section 4.2.2.2, we only calculate the progress sentiment of the news articles. The sentiment is then correlated both with the projected trend and the actual price trend. Figure 5.3 and 5.4 show generated projected trend lines as well as the actual stock price movements for Disney, Exxon and Pfizer – similar figures for other companies are shown in the appendix – [Appendix Reference].

Using the Disney trend line as a sample case, we show how news articles are classified based on where they fall in the projected trend line – Figure 5.5

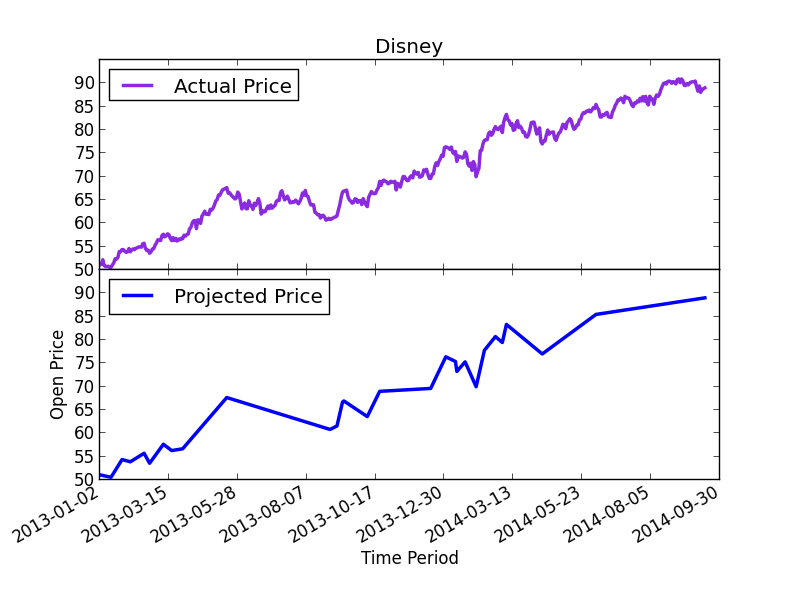


Figure 5.3 – Actual Stock Price and Projected Price of Disney

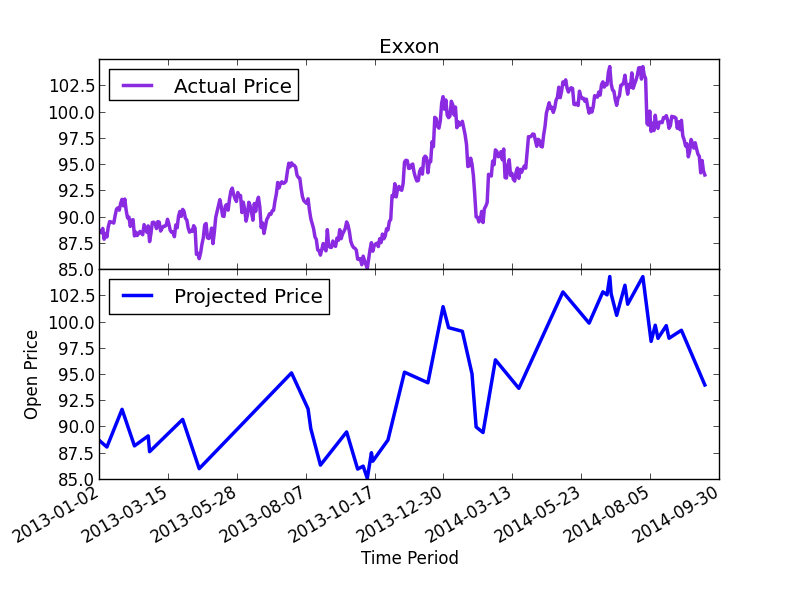


Figure 5.4 – Actual Stock Price and Projected Stock Price of Exxon

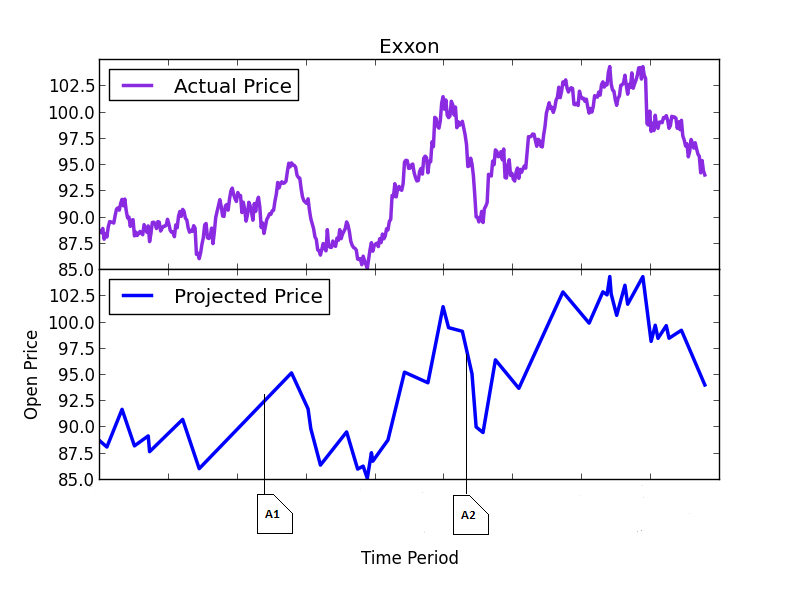
****

Figure 5.5 – Aligned News Articles with Trends

Using Figure 5.5 as a sample case, we would classify article A1 as “up” or positive while A2 would be classified as “down” or negative. After achieving this step, we can then proceed to calculate the correlation between the automatically calculated prices and the labelling. One would assume that it would be a correlation value of 1 as labelling in fact is done with the trend line; however, this is not so as news articles aren’t released every day. The table below shows the correlation results.

|  |  |  |  |
| --- | --- | --- | --- |
| Company Name | Correlation of Progress Sentiment(Actual) | | Correlation of Progress Sentiment(Actual) |
| Chevron | | 50.2 | 51.8 |
| Cocacola | | 32.1 | 32.0 |
| Disney | | 97.0 | 97.5 |
| Exxon | | 80.4 | 79.7 |
| Goldman | | 50.0 | 52.60 |
| IBM | | 52.12 | 53.41 |
| JP Morgan | | 85.37 | 85.14 |
| Microsoft | | 94.99 | 94.99 |
| Pfizer | | 53.08 | 53.33 |
| Visa | | 89.27 | 89.62 |

Figure 5.6 – Correlation values of the automatically generated data

Looking at the data above and comparing them, we see that the correlation values are quite similar to that of the manual data except for the correlation values of Goldman Sachs which is vastly different from the result of manual classification. The reason for this is that Goldman Sachs often is the source of the news (for example, Goldman Sachs often advises on buying and selling other companies) and the news is not about Goldman Sachs. This means that automatic labelling is blind to these issues as it labels both news by Goldman as well as news about Goldman without using any filter. However, when manually labelling, we ensure to label those articles are “neutral” in terms of progress of the entity and the relevant feeling sentiment.

### Labelling Discussion

We see that the results of labelling is in general very positive – providing a reason to go on with classification of the news articles – preferably using the manually labelled data. The high similarities between the progress coefficient of the automatic and manual data also is a form of validation for the manually labelled data and idea that news articles correlate with the stock price. However, the results also make harder to overlook the issues with automatic labelling.

## Data Pre-processing

Using the method described in section 4.2.3, the dataset was pre-processed. However, there’s a decision to be made about which method of tokenisation is best. In our experiments, we performed only 3 types – unigrams, bigrams and combination of both. These types are the most popular in the literature. Unigrams, also known as bag of words are criticised often for not bearing enough information but we see that in all areas, they perform quite well. Bigrams, as we will see also perform comparatively to unigrams. The combination seems to perform the best of all three. In addition, when tokenising, we only select words that have greater than or equals to three characters.

The table below shows the number of features that are extracted and the number pre-selected (based on term frequency) before feature extraction or selection. These values were selected via trial and error, matching the various parameters with results and selecting the parameters that produced the highest result.

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unigram | Bigram | Unigram + Bigram |
| Initial number of features | 26322 | 310660 | 336982 |
| Selected amount of features | 11000 | 30000 | 30000 |

Figure 5.7 – Initial features and pre-selected features

## Feature Selection or Feature Reduction

During the experimentation phase, we compared the results of SVD with chi2 based feature reduction. The results are, very similar with no distinct advantage provided by SVD (rather, it’s disadvantageous as it took up to 2 hours to compute bigrams), we decided to go opt for feature selection based on chi2, hence the classification results in the succeeding section detail the results based on feature selection and not feature reduction simply because the results are virtually the same. However, we provide the settings used for SVD and chi2 (Figure 5.8) as well as confusion matrixes of classification using both methods (Figure 5.10).

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unigram | Bigram | Unigram + Bigram |
| Chi-2 | 5000 | 17000 | 17000 |
| SVD | 4000 | 13000 | 13000 |

Figure 5.8 - Number of Features depending on whether feature selection or extraction is used

Examining figure 5.8, we see that more features are utilised for a comparative result – however, the amount of time taken to perform extraction (45 minutes for unigrams and 2 hours for bigrams and unigram + bigrams) makes it unfeasible when performing classification. This will be even more pronounced as one of the issues with the current project is the lack of data - a realistic system would expect to use many more thousands of news articles.

In addition to the benefits provided by chi2, there are also benefits in terms of ability to examine the features selected more closely. In figure 5.9 and 5.10, we show 10 tokens selected for both the progress and feeling sentiment (please note that these aren’t in any particular order – instead the table is as a result of words selected across all folds during cross-validation).

|  |  |  |
| --- | --- | --- |
| Unigram | Bigram | Unigram + Bigram |
| aaa | accord announced | aaa |
| surplus | zero percent | abnormal |
| illegal | capital declined | analysts predict |
| destroy | company profit | exciting |
| asian | stock buyback | growing market |
| litigation | Creditworthiness decreased | Largest technology |
| embezzlement | stock gained | Lawsuit jpmorgan |
| examination | wall street | laundering |
| fine | volatility index | straight year |
| value | legal claims | breach contract |

Figure 5.9 – Words selected for progress classification (manually labelled data)

|  |  |  |
| --- | --- | --- |
| Unigram | Bigram | Unigram + Bigram |
| acceptable | aaa credit | abandon |
| grossing | aaa rated | billion asset |
| questionable | abc network | cash flow |
| rallied | income drop | government bond |
| suppress | percent called | shrank percent |
| disagree | seek boost | trading stock |
| distort | later acquire | state law |
| expense | gain ground | jpmorgan led |
| investigation | dollar bond | Exclusive |
| policy | compliance action | development |

Figure 5.10 – Words selected for feeling classification(manually labelled data)

We have neglected to include the corresponding tables for the automatically labelled data as automatically labelled data had lower overall accuracy than manually labelled data. In addition, it should be noted that the words shown in either table are not necessarily exclusive to the table. For example, “jpmorgan led” which appears in the 8 row of the Unigram + Bigram column also appears as a bigram feature when classifying based on progress sentiment.

## Document Classification

### Manual Classification

In this section, we detail the results of manual document classification and the settings used to achieve the results. As we use the linear classifier, the parameters that need to be set are the class weights and the cost. Other parameters to be set are default parameters by the classifier. First, discuss the classification of progress sentiment and show the results, followed by the classification of feeling sentiment. To finish up, use only data for a single company (Goldman Sachs) for classification. This is to determine whether or not classification for a single entity is much better than group classification. It’s highly likely that this might be the case as a classifier is more likely to be fine-tuned to the exact negativity and positivity of the news articles about the company.

#### Progress Sentiment Classification

In order to set the weights, we need to look at the support for each class. Figure 5.11 shows the number of articles supporting each class (Figure 5.11). The hyper-parameters used for configuring the SVM are as follows:

LinearSVC (the class used for classification) implements the one-versus-rest classifier for multiclass problems. The penalty term is l2 rather than the typical (as performed better for this problem) used for linear SVMs. We use the default value of “1” for the parameter. [need equation reference]. This setting is however unusual as usually requires quite large numbers. Class weights are set automatically by LinearSVC. The values set the C parameter for each of the classes, dependent on the class frequencies. This needs to be set as Figure 5.11 shows, the classes are not represented equally in the training sets. Using a StratifiedKFold cross validator, we can preserve the percentage of representation for each sample.

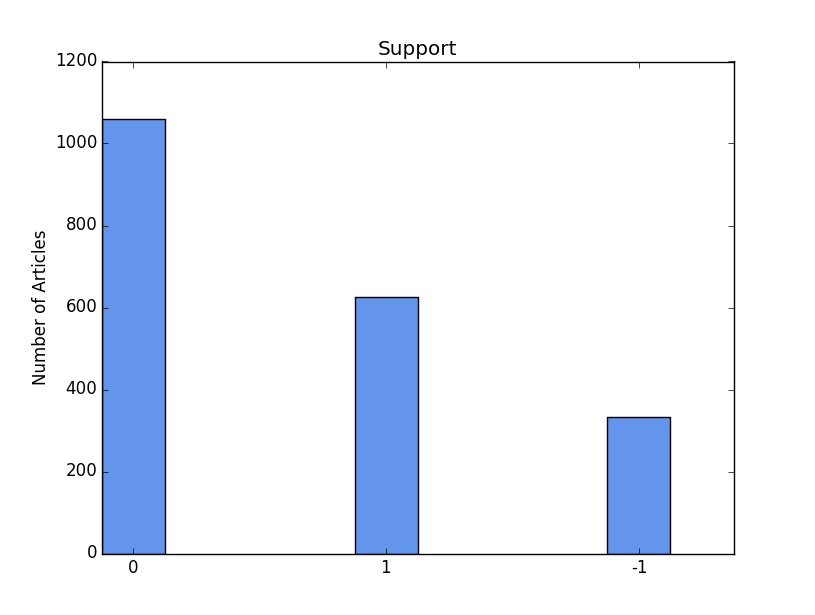
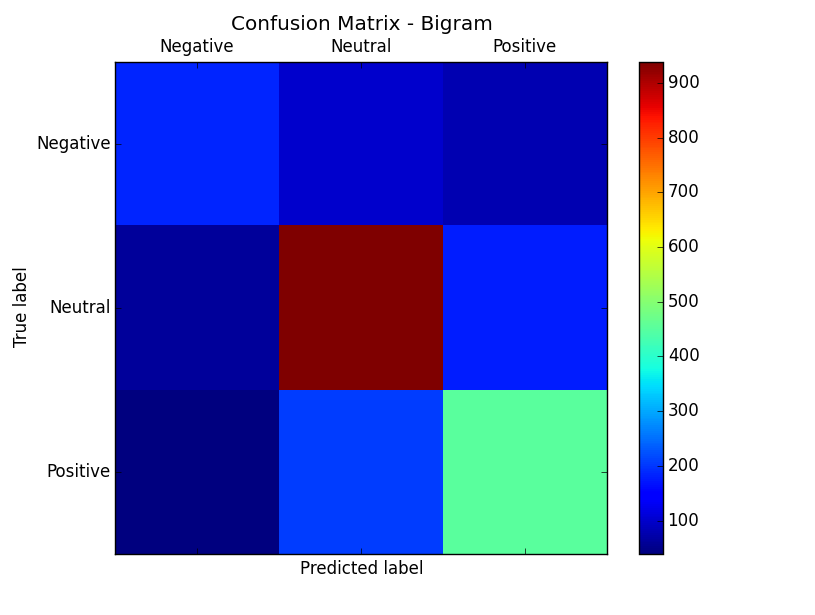
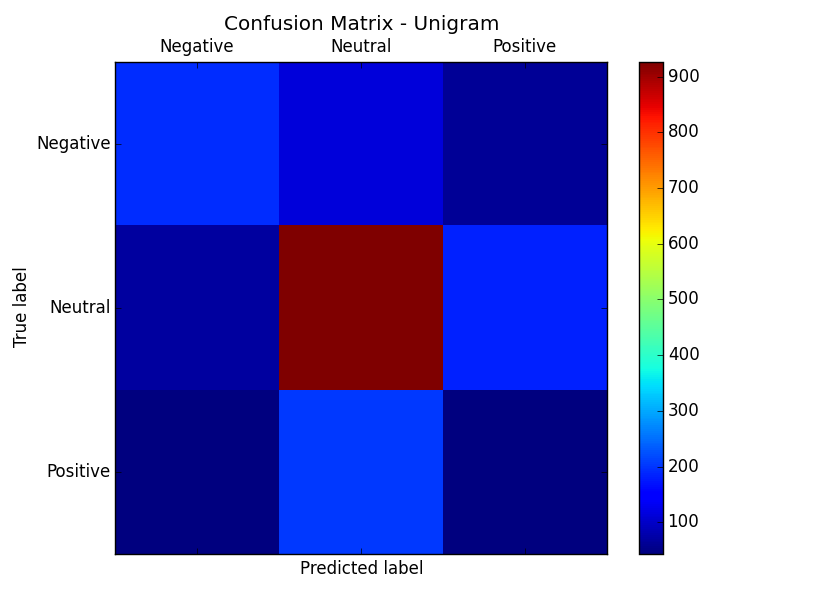


Figure 5.11 – support for the various classes (manual/progress)

Similar settings were used for bigrams, unigrams and combination experiments. In order to determine accuracy, we use cross validation and the following matrices: f-measure, recall, precision and confusion matrix. We compute the average of the scores of all the folds. Please note that we results we show here are for feature selection rather than extraction. In order to view the results for SVD-based feature extraction, please refer to the appendix.The confusion matrixes for unigram, bigram and combination (Figure 5.12) shows an overview of the accuracy for the three classes. For actual numerical values, please refer to the appendix.

The confusion matrixes show that there aren’t very big differences in the performances of the three methods of tokenisation (except when classifying positive articles). It’s very difficult to explain why this is the case except that all three methods carry similar levels of information for this problem.



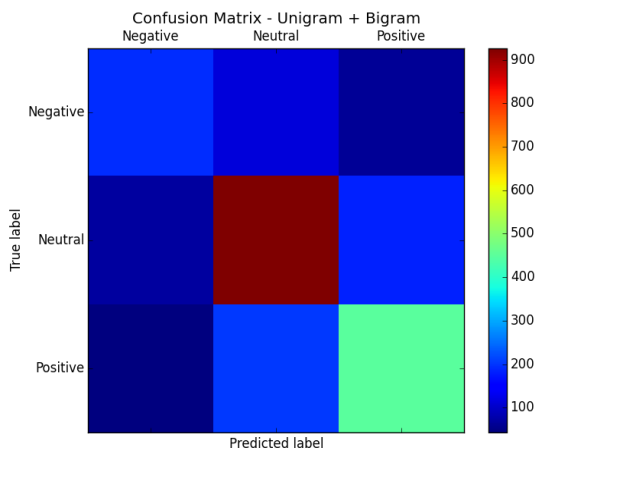


Figure 5.12 – Confusion matrices (Manual/Progress). Top left – Unigram, Top Right – Bigram, Bottom – Unigram + Bigram

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unigram | Bigram | Unigram + Bigram |
| F-measure | 69.48 | 69.82 | 70.51 |
| Recall | 69.58 | 69.97 | 70.58 |
| Precision | 69.77 | 70.17 | 70.62 |

Figure 5.13 – Table of performance of linear SVM measured by cross validation (manual/progress)

Delving into the actual numbers, we see that overall, the bigram does better than the unigram and the combination of both does better than either of them singularly. Combining this information with the confusion matrix, we see that bigrams and the combination perform better due to being able to slightly classify positive news articles better.

#### Feeling Sentiment Classification

Poorer results were achieved for the classification of feeling sentiment in general. This is contradictory to the initial belief that feeling sentiment would be easier than progress sentiment to classify. We performed classification using a linear SVM as before. The settings for feeling sentiment were quite different. In addition, classification performance for the feeling sentiment was quite poor overall. As per the previous section, we start by introducing the frequencies for the classes (Figure 5.14)

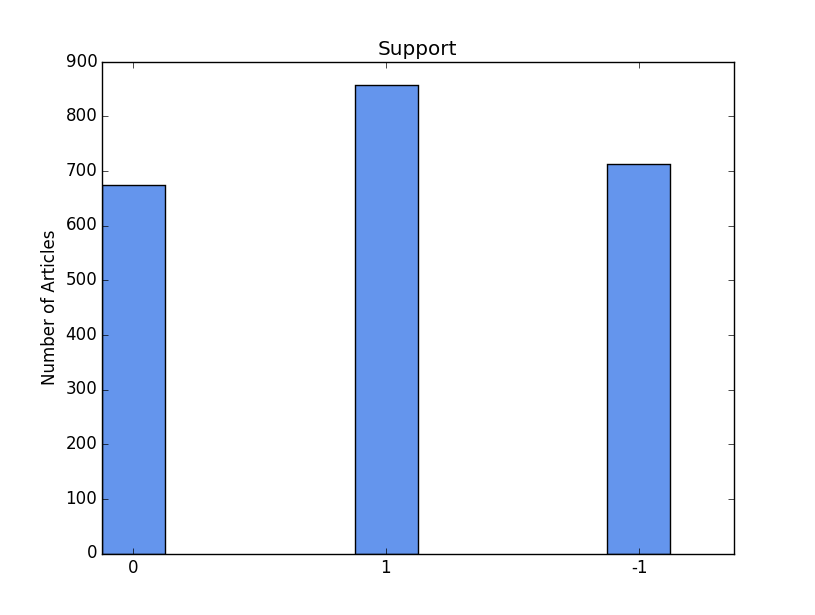
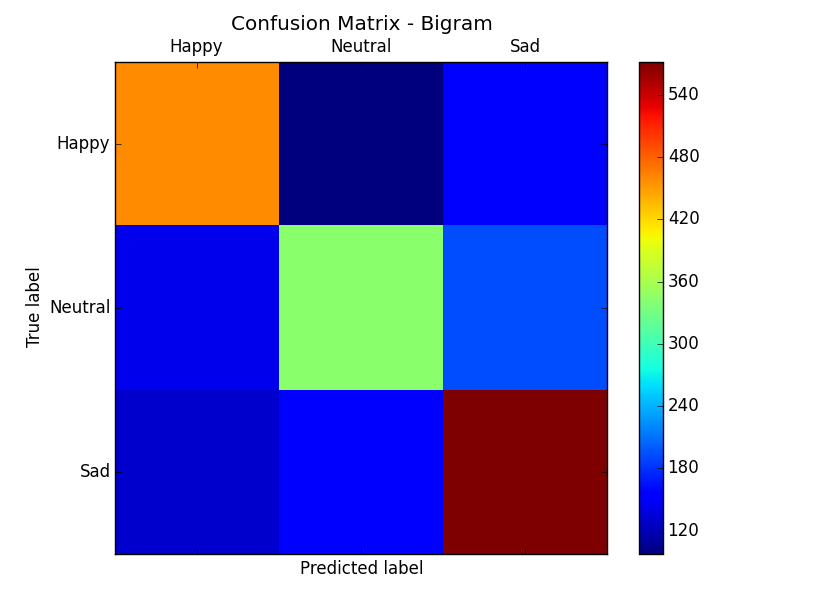
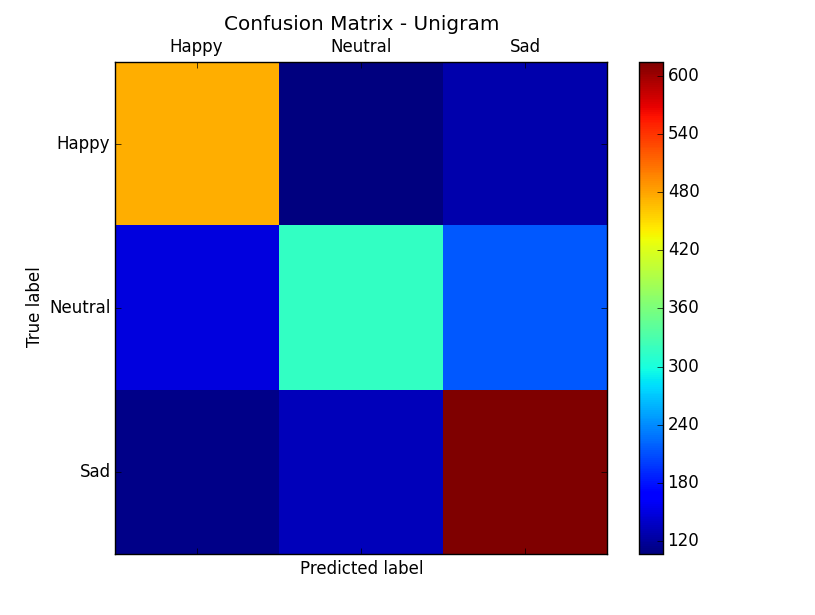


Figure 5.14 – Support for the classes (Manual/ Feeling)

Different hyper-parameter settings were used for classification. The parameter was set to higher levels with a value of . The other parameters, such as the class weight were also set automatically based on the class. We use a loss function of and penalty function of .

Considering the confusion matrices (Figure 5.15), we see that the all three methods of tokenising perform very similarly as before. A possible reason for this is that news articles often bear mixed feelings. On the surface, it may seem that news articles bear feeling sentiment orientations that lean towards one way or the other but this isn’t so. News articles often carry information that lean to both sides. A classic example of such news articles is articles that discuss “happy” sentiment. In a few of these articles, there’s also discussion of past “sad” sentiment that led to perhaps structural changes that result in improvement. Hence, while progress sentiment might be clear, feeling sentiment can often be ambiguous when it comes to classifying neutral articles.



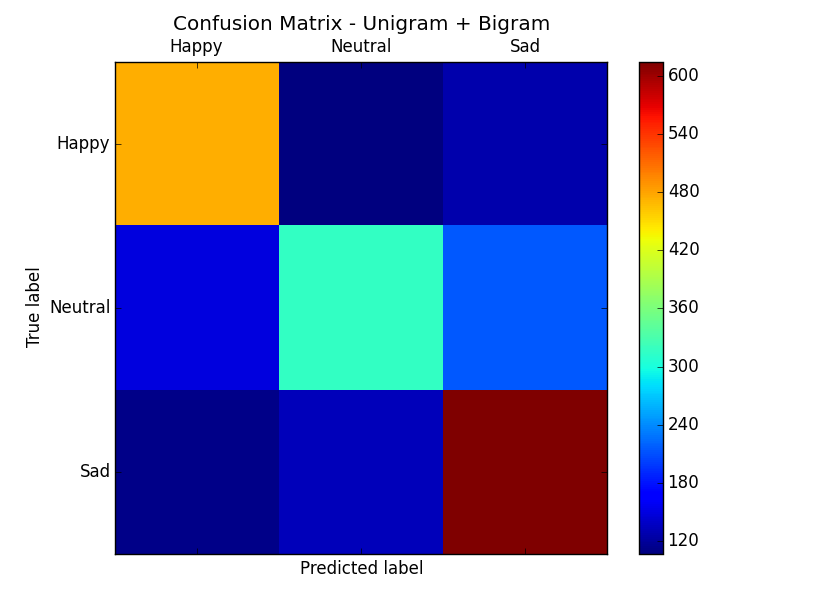


Figure 5.15 – Confusion matrices (Manual/ Feeling). Top left – Unigram, Top Right – Bigram, Bottom – Unigram + Bigram

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unigram | Bigram | Unigram + Bigram |
| F-measure | 62.00 | 60.87 | 63.68 |
| Recall | 62.48 | 61.06 | 63.92 |
| Precision | 62.23 | 61.15 | 63.99 |

Figure 5.16 –Table of performance of linear SVM measured by cross validation (Manual/ Feeling)

Bigrams are typically expected to do better than unigram due to the fact they retain sentence structure but clearly, bigrams doesn’t do very well for this problem looking at the performance measures in Figure 5.16. However, combination of both performs better than either but not by much.

### Automatic Classification

In this section, we follow the same pattern as in section 5.6.1, with the exception that we only discuss progress sentiment. For automatic classification, neutral movements are underrepresented (only about 46 news articles were classified neutral); hence, we only consider positive and negative movements.

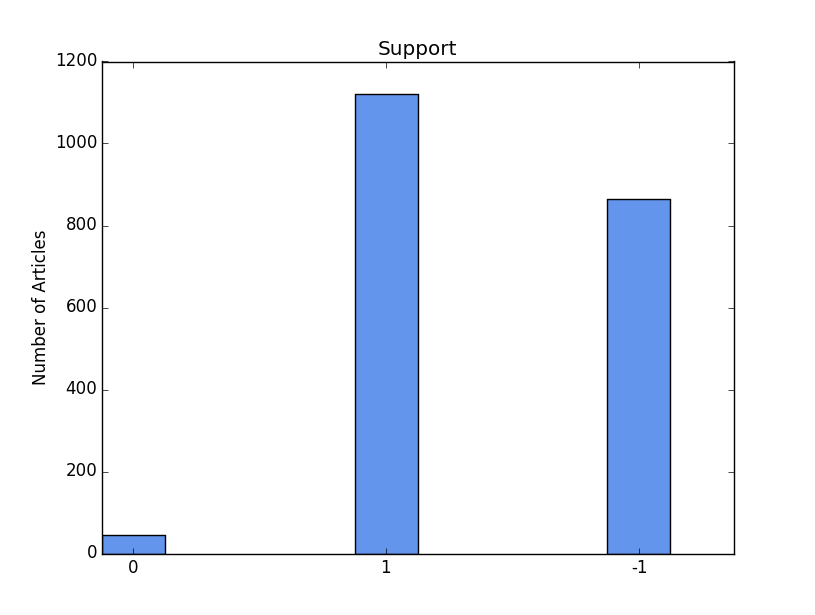


Figure 5.17 – Support for the classes (Automatic / Progress)

Figure 5.18 - Table of performance of linear SVM measured by cross validation (Automatic/ Feeling)

Here, we only present the numerical results to restrict repetitiveness as well as the fact that the manually labelled data will be used (Figure 5.18). It would of course be interesting to consider how well automatically labelled data performs when used subsequently for price prediction, however time restraints prevent this. In addition, the intention was not to use automatic labelling for classification; instead, it provided an adequate benchmark for comparison with the results of manual labelling. We have extensively discussed the pitfalls or issues automatic labelling is susceptible to and we believe that the results here can be explained by these (Reference to discussion 5.3.)

|  |  |  |  |
| --- | --- | --- | --- |
|  | Unigram | Bigram | Unigram + Bigram |
| F-measure | 64.50 | 66.30 | 64.49 |
| Recall | 67.97 | 72.55 | 61.36 |
| Precision | 61.43 | 61.12 | 68.05 |

# Conclusion

## Keys for Transforming Sentiments to Numbers

# Appendix