Using social media and machine learning to predict financial performance of a company

Sepehr Forouzani
Abstract

Using social media and machine learning to predict financial performance of a company

Sepehr Forouzani

Social media have recently become one of the most popular communicating form of media for numerous number of people. The text and posts shared on social media is widely used by researcher to analyze, study and relate them to various fields. In this master thesis, sentiment analysis has been performed on posts containing information about two companies that are shared on Twitter, and machine learning algorithms has been used to predict the financial performance of these companies.
List of Figures

1  The methodology ..................................... 7
2  Sentiment Analysis methods [18] ....................... 12
3  Machine learning workflow [34] .......................... 15
4  Steps toward financial prediction .......................... 20
5  The format of a feature vector ............................ 22
# List of Tables

1. The datasets used in the experiments. ........................................ 22
2. The companies performance based on the ROA. ...................... 24
3. The two different dictionaries and some example words. ........... 25
4. Confusion matrix ...................................................................... 26
5. The results for experiment 1 using $TW_{BMW}$ dataset. ............. 28
6. The results for experiment 2 using $TW_{BMW}$ dataset. ............. 28
7. The results for experiment 3 using $TW_{BMW}$ dataset. ............. 29
8. The results for experiment 3 using $TW_{VW}$ dataset. ............... 29
9. The results for experiment 4 using $TW_{BMW}$ dataset. ............. 30
10. The results for experiment 4 using $TW_{VW}$ dataset. .............. 30
1 Introduction

Nowadays media and in particular social media is considered as a big data source to researchers due to the large number of people communicating and sharing their ideas, feelings, knowledge, and personal opinions about various topics at any time. During the last ten years, Twitter and Facebook has emerged to be the most popular social networking websites. Facebook has 1.59 billion monthly users and Twitter has 332 million active users [6].

Data from social media provides a unique opportunity to social scientists, economists, and statisticians to understand individuals and human behavioral patterns that has effects on different areas such as finance [4]. As an example, recent research on financial performance prediction using opinion and sentiment analysis of posts that are shared in social media indicates that there is a possibility to predict a company’s stock value [5].

The data available on social media is enormous, unstructured and contains a lot of irrelevant information, therefore it is impossible for individuals to read and analyze all of the data manually. To analyze data from social media, statistical and data mining techniques need to be applied to make the best use of the data [7].

Customer’s opinion about products and services is always a concern for most large-and middle sized companies. Social media is one of the most widely used source of data about customer’s opinion toward a certain company [8]. Most companies use different methods and techniques to find out customer’s opinion about their services and products. However relating the data extracted from social media about customer’s opinion to the co-related sectors of the companies such as productivity, profitability, financial performance and economics is not always possible [21], for example if a firm improves productivity by downsizing, the profitability might be endangered
if the customer satisfaction depends on companies services [23]. Research [1] has shown that there is a relation between opinion and sentiment about a company and the stock price. However, to the best of our knowledge there are no studies that focus on investigating the relation of sentiment analysis of tweets and the financial performance of companies.

In this master thesis we will investigate the correlation between the sentiment of tweets where a certain company is mentioned in a hashtag and the financial performance of that company.

1.1 Objectives

The overall objective of this thesis project is to investigate the relation between sentiment extracted from social media and the financial performance of automotive companies. The goal is to predict the financial performance of a company based on what people write about the company on Twitter. This results in the following more specific objectives:

- Develop techniques for sentiment analysis of data from Twitter with respect to a specific company.

- Use machine learning and train a model to predict the financial performance of a company

- Develop a prototype tool for the proposed method.

1.2 Method

The work in this thesis is done through five steps, as illustrated in Figure 1.
In the first step, the problem and the objectives for the research is defined. In the second step a literature review is done. The literature study focus on reviewing related work as well as gaining knowledge about the techniques that will be used in the project.

In the third step, the experiment setups and configurations will be designed and data will be collected.

In the forth step, a prototype tool is developed in order to collect, prepare and analyze data. The analysis is based on mood and sentiment word lists. For the machine learning components in this project the Weka data mining tool [39] is used. In the fifth step, the results are evaluated by measuring the accuracy of performance prediction.
2 Related Work

In this chapter some work related to sentiment analysis methods and financial predictions using mood and sentiment analysis, will be reviewed.

In [1] the authors are collecting public tweets posted by approximately 2.7 million users. All tweets have an identifier, a publishing time, a submission type and a 140 character text. To make the data suitable for analysis, stopwords (topic independent words that are most common in a language) and punctuation are removed and then the text is filtered by words such as "I feel", "i am feeling", "I'm", "I'm", "I am", and "makes me" because those words state their author's mood state. At the next stage they use the OpinionFinder (OF) tool [13] for sentiment analysis. In order to measure polarity of a sentence in terms of being positive and negative, OF takes a text (e.g. large number of tweets) and uses the OF lexicon to determine the percentage of positive against negative sentiment of the text. To measure mood of a text they use an algorithm called Google-Profile of Mood States (GPOMS). GPOMS measures the mood of a text from six different dimensions, which are: calm, alert, sure, vital, kind, and happy.

To enable normalization of time series and comparison between OF and GPOMS results, the authors of [1] are using z-score statistical measurement which is based on local mean and standard deviation. The authors are also using econometric technique of Granger causality analysis [19] in order to investigate the relation between public mood and stock market closing value changes. The Granger causality indicates that there is a predictive relation of certain mood categories and the closing price of the stock market.

In [3] the authors used machine learning and social media to predict how successful a movie will be. In order to measure success of a movie the authors used return on investment (ROI) which is a profitability metric, and
they applied binary and multi-class classification algorithms such as support vector machines (SVM), multilayer perceptron (MLP), decision trees (J48), random forest and logitBoost algorithm to predict the success. The results shows that random forest was the best classifier, with an accuracy of almost 84%.

In [12] the authors investigate the possibility of predicting electronic devices market sales using social media. In their work they are analyzing sentiment of Twitter comments about a certain product before the product is released. They are using semi-supervised recursive auto encoders for predicting sentiment distribution. Semi-supervised recursive auto encoders is an artificial neural network which its goal is to learn encoding a set of data, typically for the purpose of dimensional reduction. In sentiment analysis semi-supervised recursive auto encoders are used to learn semantic vector representations of a phrases [20]. After running sentiment analysis, the total number of comments, number of positive comments, total number of re-tweeted comments and number of re-tweeted positive comments are extracted and used as features in their model. In the experiments their model showed 35% of accuracy in prediction of iPad3 sale meanwhile linear regression was showing 58% accuracy in iPad3 sale prediction which is a low accuracy and could not be used as a practical model.

In [2] the authors are using Artificial Neural Networks (ANN), Support Vector Machines (SVM) and Relevance Vector Machines (RVM) to predict daily returns for an FX carry basket. A currency basket is a portfolio of selected currencies with different weightings, and FX carry basket is made of a long position in high yielding currencies versus a short position in low yielding ones is a common asset for fund managers and speculative traders. It was found that in general the committee of networks was much more effective
at predicting five day returns than one day returns, and it was on this basis that the optimal configuration was used.

In [9] it is stated that the list of words that is used in general to measure the sentiment of a text is not accurate to be used to measure sentiment of finance related texts. To illustrate this, the authors of [9] did a review of the negative words extracted from 10-k reports (an annual report which contains summary of a company’s financial performance [15]) based on the Harvard dictionary [14] and found out that almost seventy five percent of the words counted as negative are not negative in finance. Therefore they have developed a new word dictionary which reflects the tone of financial texts with a higher accuracy. The authors have used a bag of words (considering a text like a bag for its words, regardless of grammar and order of words) approach to produce vector of words and word counts, and modified one of the most common term weighting scheme to make it adjustable to document length.

In [10] the authors are developing an automated method for sentiment classification. They are using a classifier which is based on a multinomial Naive Bayes classifier to determine the positive, negative and neutral sentiment of a document. They also propose a technique that can be used to determine sentiment of documents in any languages. In their method, the TreeTagger [16] (a language independent part-of-speech tagger) is used for part-of-speech tagging and the differences in distribution of positive, negative and neutral tags are observed. For feature extraction they used N-gram as binary features and the frequency of keywords. Unigrams, bigrams, and trigrams are used for experiments, and the authors are stating that when bigrams are used, the performance is the best.

In [11] four classes of mood: calm, happy, alert and kind are used and
a text is categorized into these four classes using a analysis tool. The tool uses a word list based on the Profile of Mood States (POMS) questionnaire \[17\] where the POMS different states are mapped into their four mood states using static correlation rules. They also filtered down a set of tweets into emotion specific texts using words such as ”feel”, ”makes me”, ”I’m”, ”I am”. In this work the authors are using a new cross validation method called k-fold sequential cross validation to train the model and the model showed 75.56% accuracy in prediction of stock market movements. They have tried four different learning algorithms: linear regression, logistic regression, support vector machines (SVMs), and self organizing fuzzy neural networks (SOFNN) to learn and study correlation of mood and market. The conclusion is that SOFNN performed better compared to the other algorithms.

3 Background theory

3.1 Social media

The tools and platforms that enables users to interact and exchange information in different forms such as text, picture, video and etc. are called social media \[24\]. There are a number of different types of social media for example blogs, discussion boards and networking platforms such as Facebook and Twitter. Twitter is one of the most popular social media services that enable users to publish and share a maximum of 140 characters text called tweets and use hashtags ”#” to relate their tweets to a specific topic, person or a company. Several companies and business strategists consider social media as an important arena and they are constantly trying to find out various ways to increase their profitability using social media\[25\].
3.2 Sentiment analysis

Sentiment analysis is done using natural language processing and information extraction with the goal of obtaining the writer’s feeling as positive, negative or neutral [27]. Sentiment analysis is often used as component in opinion mining when the goal is to analyze sentiment and attitudes [28]. There are a number of various methods that can be used to classify sentiment of a text. A list of methods are shown in Figure 2.

![Figure 2: Sentiment Analysis methods](18)

In this thesis the Dictionary-based approach is used for sentiment analysis.
3.3 Financial performance

Most of the time financial analysts and investors are focusing on return on equity (ROE) as the primary metric for measuring companies performance. Many executives focus heavily on this metric as well, believing that it is the one that seems to get the most attention from the investor community. ROE is calculated by dividing the net income by shareholder’s equity.

\[
\text{Return on Equity} = \frac{\text{Net Income}}{\text{shareholder’s equity}} \quad (1)
\]

Shareholder’s equity is the equity of a company as divided among individual shareholders of company’s stock [48]. Using ROE as performance metric has some shortcomings as well. As an example, companies can artificially maintain a good value of ROE by growing debt leverage and stock buybacks which are funded through accumulated cash. Therefore other metrics such as return on assets (ROA) can be used instead of ROE. ROA directly considers the assets that are used to support business activities and it determines whether a company is able to generate sufficient return on the assets rather than simply showing robust return on sales [29]. ROA is an indicator of a company’s profitability based on its total assets [31], it captures the fundamentals of company’s performance in a general way by looking at both income statement performance and the assets required to run a business [22]. ROA is a good metric to measure performance of a company on generating income by using the assets. ROA is calculated by dividing a company’s earnings by its total assets and displayed as a percentage. Sometimes ROA is referred to as ”return on investment”. ROA is calculated using below formula:

\[
\text{Return on Assets} = \frac{\text{Net Income}}{\text{Total Assets}} \quad (2)
\]
3.4 Data collection

Data collection and dataset creation is the first step when you want to create a statistical model using machine learning. The dataset is commonly divided into three subsets: a training set, a validation set and a test set. The training set is used to train the statistical model, the validation set is used to estimate how well the model is trained and the test set is used to measure the performance of the model.

3.4.1 Feature Vectors

A feature vector is the way an object is presented in machine learning and pattern recognition. Feature vectors are $n$-dimensional vectors where each vector represents an object. A numeric representation of the features (variables) will enhance statistical analysis, therefore many machine learning algorithms requires numerical features.

3.5 Machine learning

Machine learning is a field of computer science which studies and explores ways of making algorithms find patterns or learn how to do certain tasks. In this thesis machine learning is used to predict the performance of a company. Figure 3 shows the workflow for the machine learning process we have used in this thesis.
In the first step (data ingestion) the data is collected and stored in a database. After collecting the data, the data is cleaned and/or transformed. The data is divided into two sets: a training set and a testing set. In the next step a mathematical model is built based on the training set and then the model will be tested against the testing set.

In order to improve the results, the user can make decision about creating or choosing different data and feature vectors (data presentation style), after results are produced from the model.

There are three categories of machine learning that are based on their nature of learning.

- **Supervised Learning**: In supervised learning the computer receives a set of inputs and their related outputs from a teacher. The goal is to find a general mapping model from input to output.

- **Unsupervised Learning**: In unsupervised learning, the computer find structures in the input data without having any input from a teacher.

- **Reinforcement Learning**: In reinforcement learning the computer inter-
acts with an environment to achieve the goal without any help from a teacher.

3.5.1 Classification Algorithms

A classification algorithm task is to pick the right identified categories in data, for the new observations, the classifier estimates categories for new data based on the model parameters that are learned from the training data.

Different classification algorithms use different classifier methods and variables and therefore a number of classification algorithms can be applied on the data in order to find the most suitable and efficient algorithm [30]. In this section a few different classification algorithms that are used in the project will be reviewed.

Random Forest [35] is bagged trees with both bootstrap sampling of the data and a form of attribute bagging. A decision tree is made of a directed series of decisions, based on input variables value, and culminating in a classification of the target variable. Bagging is a method of combining multiple predictors. It will get a bootstrap sample from training set and train a predictor on that sample. Samples with replacement from the known weights called a bootstrap sample. Random forests provide a simple means of analyzing feature importance, and the resulting score is known as the variable importance score. In random forest it is not required to separate a test set from the data to get an unbiased estimate of the error since each tree in random forest is built by using a different bootstrap sample from the original data. Bootstrap is an algorithm, designed to improve the stability and accuracy of machine learning algorithms.

Naive Bayes [33] is a probabilistic classifier that uses Bayes theory with the assumption that the features are independent (occurrence of one feature
does not effect the probability of others). Naive Bayes computes probability $p$ as the probability of feature $x$ represented by a vector $x = (x_1, ..., x_n)$ being in the class $c : p(c|x)$. The conditional probability using Bayes theorem can be shown as:

$$p(c|x) = \frac{p(c)p(x|c)}{p(x)}$$

when training model time is important Naive Bays is useful.

**AdaBoost [32]** stands for adaptive boosting and it assumes that finding many weak models are easier than finding one accurate model. Boosting is an approach to create predictions rules with high accuracy using a combination of weak models and rules that have low accuracy in prediction. Boosting generates a sequence of base models and then decides a final estimate of the target variable based on aggregating the estimates of the base models. AdaBoost generates a numbers of weak classifiers and a final estimate of the target variable is chosen based on aggregating the estimates made by the base models. Similar to the random forest algorithm, AdaBoost also have a variable importance estimation but in a different way. In AdaBoost the more informative variables are used more often, and the less informative features are barely used.

**Cross validation [42]** creates a training set and a test set by partitioning the original data with the goal to train and evaluate the model. In $k$-fold cross validation the original data will be divided into $k$ number of subsamples. One subsample is selected as test dataset and the rest ($k - 1$) number of subsamples are used as training set for the model. The same process will be repeated for $k$ number of times (folds) and each subsample will be used at least once as test set and then the results will be averaged or combined to make the best estimation.
3.5.2 Data balancing

If the number on instances in classification categories in a dataset are having a huge difference, the dataset is called imbalanced. To counter the issues of imbalanced data, methods such as over-sampling (creating new samples of a certain class) and under-sampling (removing instances of a class) have been proposed. Synthetic Minority Oversampling TEchnique (SMOTE) [36] is an over-sampling algorithm which provides more instances of the class with lower number of instances in addition to under-sampling of the class with more number of instances. In SMOTE, based on the required number of over-sampling $K$ number of the nearest neighbor to the data point is selected and then after these steps the synthetic sample will be created:

- Take the difference of a data instance to its nearest neighbor,
- Multiply the number by a random value between 0 and 1,
- Add the new data point to the considered feature vector

3.5.3 Feature selection

The process of selecting a subset of features that should be used to construct the model is called feature selection. In machine learning and statistics, the process is also called variable selection. There are various ways to do feature selection. As an example, information gain $IG$ specify the most important features following the formula:

$$IG(T, a) = H(T) - H(T|a)$$ \hspace{1cm} (4)

where:

- $T$ is set of training example,
- $a$ is the index of a feature
The **H() function** is an entropy (Entropy is a measure of the randomness of a variable and it measures the level of impurity in a group of examples).

# 4 Implementation

In this chapter the design and implementation of the financial performance predictor (FPP) is described.

## 4.1 Financial Performance Predictor design

The financial performance predictor (FPP) is a prototype tool for prediction of companies financial performance using machine learning. The flow of how FPP is used is shown in figure 4.
The first step is to collect relevant data, in this thesis we use data from Twitter. In order to detect the sentiment of a tweet or a group of tweets, we use the bag of word method. The bag of word method focuses on the words or in some cases set of words (a string of words), regardless of the context of sentence. We use a list of words (from a dictionary) and all words that are attached to a sentiment. The words are either positive or negative. In the experiment we have used two different dictionaries one with that is developed for financial purposes and one more general. The second step is to count the number of occurrence of each word present in the dictionaries in the extracted tweets. The result is combined with the ROA for the corresponding
time period and included in the feature vectors. In the forth step machine learning algorithms will be applied on the feature vectors to train a model to predict if the ROA increases or decreases based on the sentiment of the tweets. The classification algorithms that we have used to train the model are Random Forest, Naive Bayes and Adaboost.

4.2 Financial Performance Predictor Implementation

Various programming languages and tools are used in the implementation of the FPP.

4.2.1 Collecting data

In order to download tweets a web scraper is written in python programming language. At the first step a web search query will be made by a python library called selenium [49]. In the second step the HTML contents will be stored to driver’s page source of a web browser.

In the third step a python library called beautifulsoup [41] is used to organize and extract the required data from the HTML source.

At the last step the tweets will be saved as a comma separated version (CSV) file and then stored in a MySQL database to ease the data management.

4.2.2 Feature vectors creation

In this thesis a program for creating feature vectors is written in Java. The program uses the word dictionaries and count the number of occurrence of each dictionary word in the tweets. The result is stored in a vector. The format of a feature vector is shown in Figure 5.
The class variable it the company’s performance. The value of class variable is 1 in case of over-performance and 0 in case of under-performance.

5 Experiments and Results

In this section the experimental setup along with the results are described. The results are further analyzed in Section 6.

5.1 Dataset

Two datasets are used for the experiments. The first dataset denoted as $TW_{BMW}$ contains tweets where BMW is either mentioned or used in a hashtag (#BMW). The second dataset is called $TW_{VW}$ contains tweets where Volkswagen is either mentioned or used in a hashtag (#Volkswagen). The two datasets are described in Table 1

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Description</th>
<th>Size</th>
<th>Time period</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{BMW}$</td>
<td>Tweets related to BMW</td>
<td>677596</td>
<td>2007-2015</td>
</tr>
<tr>
<td>$TW_{VW}$</td>
<td>Tweets related to Volkswagen</td>
<td>151648</td>
<td>2012-2015</td>
</tr>
</tbody>
</table>

An example of a negative tweet from $TW_{BMW}$ is:

"BMW is ruining the M-division brand by releasing crap like the "X6 M"
- http://tinyurl.com/cb2nq7"
An example of a positive tweet from the same dataset is:

"Track drive reveals excellent balance of the 2015 BMW 228i - Torque News http://bit.ly/1xk4xj7 - #BMW"

An example of a neutral tweet (neither positive or negative) from the same dataset:

"mclaren should come back later in the race when ferrari and bmw have to use the hard tyres hopefully, anyway"

The sentiment of each tweet is determined by counting the occurrence of positive and negative words. If a tweet contain more positive words than negative words, the sentiment is considered positive, if there are more negative words than positive words, the sentiment is considered negative. If a tweet contain the same amount of positive and negative words the sentiment is considered to be neutral.

5.2 Quarterly reports

To obtain the value on return on asset (ROA) for each quarter, BMW quarterly reports (10-Q reports) are downloaded from [44] and Volkswagen quarterly reports are downloaded from [45]. The value of ROA is not explicitly mentioned in the quarterly reports and therefore it is calculated manually using the value of the total income and and the total assets value. In Table 2 performance of BMW and Volkswagen in different quarter of the year is shown.

5.3 Dictionaries

We have used two different dictionaries to determine the sentiment of tweets. The first dictionary (called the regular dictionary) is inspired by the positive and negative emotions from the tool Linguistic Inquiry and Word Count
Table 2: The companies performance based on the ROA.

<table>
<thead>
<tr>
<th>Year</th>
<th>Quarter</th>
<th>BMW</th>
<th>Volkswagen</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015</td>
<td>Quarter 1</td>
<td>Over-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>Over-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Under-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td>2014</td>
<td>Quarter 1</td>
<td>Over-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>Over-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Under-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td>2013</td>
<td>Quarter 1</td>
<td>Under-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>Over-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Under-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td>2012</td>
<td>Quarter 1</td>
<td>Over-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>Under-perform</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Under-perform</td>
<td>Over-perform</td>
</tr>
<tr>
<td>2011</td>
<td>Quarter 1</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td>2010</td>
<td>Quarter 1</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td>2009</td>
<td>Quarter 1</td>
<td>Under-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Under-perform</td>
<td>—</td>
</tr>
<tr>
<td>2008</td>
<td>Quarter 1</td>
<td>Under-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Under-perform</td>
<td>—</td>
</tr>
<tr>
<td>2007</td>
<td>Quarter 1</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 2</td>
<td>Over-perform</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Quarter 3</td>
<td>Over-perform</td>
<td>—</td>
</tr>
</tbody>
</table>

(LIWC) [37]. The second dictionary (called the financial dictionary) is called Loughran-McDonald master dictionary[38]. The Loughran-McDonald mas-
ter dictionary is an extension of the 2of12inf wordlist that includes an addition of the words that are appearing in companies annual reports. The 2of12inf is a wordlist from SCOWL (Spell Checker Oriented Word Lists) and Friends consisting of English words that are useful for creating high-quality list of words for spell checkers [43].

Table 3: The two different dictionaries and some example words.

<table>
<thead>
<tr>
<th>Regular dictionary</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive Emotions</td>
<td>happy, pretty, good</td>
</tr>
<tr>
<td>Negative Emotions</td>
<td>hate, worthless, enemy, hurt</td>
</tr>
<tr>
<td>Financial dictionary</td>
<td>Example</td>
</tr>
<tr>
<td>Positive Emotions</td>
<td>best, achieve, able</td>
</tr>
<tr>
<td>Negative Emotions</td>
<td>abandoned, misprice, untrusted</td>
</tr>
</tbody>
</table>

Table 3 shows some sample words from the two different dictionaries we have used.

5.4 Weka

All experiments are done using Weka [39]. Weka has a collection of data mining algorithms, predictive modeling and tools for visualization and a graphical user interface for ease of access to its functions.

Three different classification algorithms are used in our experiments: Random forest, Naive Bayes and AdaBoost. Information Gain feature selection method is been used for Naive Bayes classifier. For data balancing, the SMOTE algorithm [36] and Weka Randomize filter are used. The default settings for each algorithm in Weka are:

- Random Forest: Number Of Trees: 100, Seed = 1.
• AdaBoost: Number of Iteration = 10, Seed = 1, Weight Threshold = 100.

• SMOTE: Nearest Neighbor = 5, Percentage (percentage of SMOTE instances to create) = 100, Random seed = 1.

5.5 Experiments

We have done four different experiments to get an understanding on the possibilities to predict a company’s performance based on public opinion extracted from social media. The experiments are different in terms of the number of feature vectors used, the features and the choice of classifier. All experiments have the same classifier setup. For each relevant time period, a number of feature vectors are created from the datasets. For each time period a variable describing if the company was under-performing or over-performing (relative to previous quarter) is added. The differences between the experiments are the number of feature vectors that are created for the dataset and what dictionary that is used.

The results for the different classifiers are described as confusion matrices in which we present the number of true positives, false negatives, true negatives, and false positives as illustrated in Table 4.

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Predicted class</th>
</tr>
</thead>
<tbody>
<tr>
<td>True Neg. (TN)</td>
<td>False Pos. (FP)</td>
</tr>
<tr>
<td>False Neg. (FN)</td>
<td>True Pos. (TP)</td>
</tr>
</tbody>
</table>

Table 4: Confusion matrix

To evaluate the results we use the measures accuracy, precision, recall and F-score that can be derived from the confusion matrix.
Accuracy is defined as:
\[
\frac{TP + TN}{TP + FP + TN + FN}
\]

precision is defined as:
\[
\frac{TP}{TP + FP}
\]

recall as:
\[
\frac{TP}{TP + FN}
\]

and F-score (to measure test’s accuracy) as:
\[
\frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}
\]

5.5.1 Experiments with the regular dictionary

Experiment 1: Combined tweets

In the first experiment all tweets that were published during each year’s quarter are combined and one feature vector representing a quarter of a year is created. The words in the regular dictionary are used as features together with a variable representing the total sentiment of the tweets and a variable that indicates whether the company was over performing or under performing during specific quarter of the year.

In the experiment, a model was trained and evaluated on 27 instances using 10-fold cross validation.

Table 5 shows the results for experiment 1 using three different classifiers and the TW_{BMW} dataset.

Experiment 2: Combined tweets and changes in sentiment

In the second experiment all tweets published during each year’s quarter are combined and the total sentiment is specified. Feature vectors are created
Table 5: The results for experiment 1 using $TW_{BMW}$ dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Over-perform</th>
<th>Under-perform</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{BMW}$</td>
<td>Random Forest</td>
<td>10</td>
<td>3</td>
<td>74.07%</td>
<td>0.714</td>
<td>0.769</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$</td>
<td>Naive Bays</td>
<td>10</td>
<td>3</td>
<td>74.07%</td>
<td>0.714</td>
<td>0.769</td>
<td>0.74</td>
</tr>
<tr>
<td></td>
<td></td>
<td>4</td>
<td>10</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$</td>
<td>AdaBoost</td>
<td>9</td>
<td>4</td>
<td>62.96%</td>
<td>0.6</td>
<td>0.692</td>
<td>0.64</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8</td>
<td>6</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

using the changes of sentiment from one quarter to another. The words in regular dictionary are used as features together with a variable representing the total sentiment of the tweets and a variable that indicates whether the company was over performing or under performing during specific quarter of the year. In the experiment, a model was trained and evaluated on 27 instances using 10-fold cross validation.

Table 6: The results for experiment 2 using $TW_{BMW}$ dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Over-perform</th>
<th>Under-perform</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{BMW}$</td>
<td>Random Forest</td>
<td>5</td>
<td>9</td>
<td>25.92%</td>
<td>0.313</td>
<td>0.357</td>
<td>0.33</td>
</tr>
<tr>
<td></td>
<td></td>
<td>11</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$</td>
<td>Naive Bays</td>
<td>13</td>
<td>1</td>
<td>77.77%</td>
<td>0.722</td>
<td>0.929</td>
<td>0.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$</td>
<td>AdaBoost</td>
<td>10</td>
<td>4</td>
<td>66.66%</td>
<td>0.667</td>
<td>0.714</td>
<td>0.688</td>
</tr>
<tr>
<td></td>
<td></td>
<td>5</td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Experiment 3: One feature vector per 100 tweets

In experiment 3 one feature vector is created per 100 tweets and Y variables

28
of feature vectors are assigned based on their published time. The Y variable
(value to be predicted) is zero if the company is under-performing and one
if the company is over-performing.

In this experiment the data is balanced using SMOTE algorithm and the
randomize algorithm \[47\]. The randomize algorithm randomly shuffles the
order of instances passed through and is used to prevent over-fitting.

Table 7: The results for experiment 3 using $TW_{BMW}$ dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Over-perform</th>
<th>Under-perform</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{BMW}$</td>
<td>Random Forest</td>
<td>2564</td>
<td>308</td>
<td>84.61%</td>
<td>0.786</td>
<td>0.893</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td></td>
<td>698</td>
<td>2968</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$</td>
<td>Naive Bays</td>
<td>1947</td>
<td>925</td>
<td>68.09%</td>
<td>0.626</td>
<td>0.678</td>
<td>0.65</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1161</td>
<td>2505</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$</td>
<td>AdaBoost</td>
<td>1136</td>
<td>1736</td>
<td>62.05%</td>
<td>0.604</td>
<td>0.396</td>
<td>0.478</td>
</tr>
<tr>
<td></td>
<td></td>
<td>745</td>
<td>2921</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: The results for experiment 3 using $TW_{VW}$ dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Over-perform</th>
<th>Under-perform</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{VW}$</td>
<td>Random Forest</td>
<td>604</td>
<td>194</td>
<td>86.17%</td>
<td>0.953</td>
<td>0.757</td>
<td>0.842</td>
</tr>
<tr>
<td></td>
<td></td>
<td>30</td>
<td>792</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{VW}$</td>
<td>Naive Bays</td>
<td>567</td>
<td>231</td>
<td>77.22%</td>
<td>0.804</td>
<td>0.711</td>
<td>0.752</td>
</tr>
<tr>
<td></td>
<td></td>
<td>138</td>
<td>684</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{VW}$</td>
<td>AdaBoost</td>
<td>448</td>
<td>350</td>
<td>60.86%</td>
<td>0.612</td>
<td>0.561</td>
<td>0.584</td>
</tr>
<tr>
<td></td>
<td></td>
<td>284</td>
<td>538</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
5.5.2 Experiments using the financial dictionary

Experiment 4: One feature vector per 100 tweets In the forth experiment one feature vector is created per 100 tweets and Y variables of feature vectors are assigned based on their published time.

In this experiment in order to balance the data instances, SMOTE and randomize algorithms are used.

Table 9: The results for experiment 4 using $TW_{BMW}$ dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Over-perform</th>
<th>Under-perform</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{BMW}$ Random Forest</td>
<td>1261</td>
<td>175</td>
<td>81.12%</td>
<td>0.888</td>
<td>0.758</td>
<td>0.816</td>
<td></td>
</tr>
<tr>
<td></td>
<td>442</td>
<td>1390</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$ Naive Bays</td>
<td>1110</td>
<td>326</td>
<td>71.54%</td>
<td>0.79</td>
<td>0.67</td>
<td>0.724</td>
<td></td>
</tr>
<tr>
<td></td>
<td>604</td>
<td>1228</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{BMW}$ AdaBoost</td>
<td>268</td>
<td>1168</td>
<td>60.67%</td>
<td>0.594</td>
<td>0.93</td>
<td>0.724</td>
<td></td>
</tr>
<tr>
<td></td>
<td>117</td>
<td>1715</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10: The results for experiment 4 using $TW_{VW}$ dataset.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Classifier</th>
<th>Over-perform</th>
<th>Under-perform</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>$TW_{VW}$ Random Forest</td>
<td>1304</td>
<td>492</td>
<td>83.03%</td>
<td>0.914</td>
<td>0.726</td>
<td>0.808</td>
<td></td>
</tr>
<tr>
<td></td>
<td>122</td>
<td>1702</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{VW}$ Naive Bays</td>
<td>1110</td>
<td>686</td>
<td>65.85%</td>
<td>0.669</td>
<td>0.618</td>
<td>0.64</td>
<td></td>
</tr>
<tr>
<td></td>
<td>550</td>
<td>1274</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$TW_{VW}$ AdaBoost</td>
<td>1629</td>
<td>167</td>
<td>57.59%</td>
<td>0.544</td>
<td>0.907</td>
<td>0.678</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1368</td>
<td>456</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6 Discussion

In the first experiment one feature vector was created for each quarter of the year, which means 27 data instances in total. Low number of data instances can be one of the reasons that the accuracy is lower in compare to other experiments. In the second experiment, instead of counting number of words and use them as features, the differences of word counts from previous quarter is used and the prediction accuracy has dropped for random forest algorithm while it showed a little improvement in other classifiers. The reason for getting low accuracy with random forest classifier could be that the sentiment in feature vectors should not be created in relation to other feature vectors. In the third and forth experiment, one feature vector is created per 100 tweets and the datasets are balanced, then the prediction accuracy improves. This could be due to balanced number of instances.

Among all of the experiments that is done, except experiment 2, the most accurate classifier was Random forest classification algorithm, from the third experiment which provided 86.17% accuracy in an experiment where 100 tweets from TWVW dataset were combined into one feature vector and the regular dictionary was used as features.

The best results was obtained when using random forest. Random forest ranks the variables in the feature vector, and also relation between each variables while splitting nodes, in order to produce higher accuracy. The data used to train the random forest classifier was balanced and therefore a more accurate classification model could be produced.
7 Conclusion

Customer’s opinion about products and services is always a concern for most large-and middle sized companies because it has effects on the company’s financial performance. Social media is one of the most widely used source of data about customer’s opinion toward a certain company. We have presented a machine learning approach toward predicting two companies financial performance using tweets that are related to them from twitter. We use two different set of features based on two different sentiment analysis dictionaries. Three different classification algorithms (Random forest, Naive Bays and AdaBoost) are used to find the best model to predict changes of Return on Assets (ROA) from one quarter to another quarter. Our experiments shows that with an accuracy of 86.17% tweets can predict whether a company will over-perform or under perform in the upcoming quarter of the year. However more research on various companies need to be done in order to find the most optimal prediction accuracy percentage.

8 Future work

In this thesis, sentiment of twitter and changes of ROA from one quarter of a year to another quarter have been used to predict financial performance of a company. Changes of ROA is not the only way to predict the financial performance of a company. There are many different variable and metrics such as Internal rate of return (IRR), Cash-flow return on investment (CFROI), Discounted cash flow (DCF) and Return on Equity (ROE) that could also be used and it would be interesting to investigate possibilities to predict these metrics as well.

We focused on Twitter in this work but there are many other online
forums and social media that may have more effect on companies performance or reflect the opinion of certain companies user better than Twitter. A direction for future work would be to investigate other forms of social media and how well they can predict the performance of a company.

In this work finding we used a bag of words method to detect the sentiment of a text. There are many other sentiment analysis methods which can be used to find sentiment of a text.

In this work the features that we considered consist of word counts only. There might be many other factors that are important in predicting the performance of a company. An obvious direction for future work is to extend the set of features and to do more experiments on different data and on different companies.

References


[16] TreeTagger’ http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/


[43] *Release 4 of the 12dicts word lists* [http://wordlist.aspell.net/12dicts-readme-r4/](http://wordlist.aspell.net/12dicts-readme-r4/)


[46] SCOWL (And Friends) wordlist http://wordlist.aspell.net/

