Correlation between emotional tweets and stock prices

Kätriin Kukk

supervisor: Michael Dunn
Abstract

Social media platforms such as Facebook and Twitter have enormous amounts of data that can be extracted and analyzed for various purposes. Stock market prediction is one of them. Previous research has shown that there is a correlation between Twitter sentiment – the proportion of positive, negative and neutral tweets – and the changes in companies’ stock prices. The present study investigates if categorizing tweets into a bigger number of categories – anger, disgust, joy, surprise, none - results in stronger correlations being found. In total, 5985 tweets in English about American Airlines, American Express, AstraZeneca and ExxonMobil were extracted and analyzed with the help of sentiment and emotion classifiers trained. Tweet sentiment showed stronger correlations with stock returns than emotion did, although the type of correlation found differed between the companies considered. It is suggested that dividing tweets into fewer categories results in semantically more distinct labels that are easier to distinguish between and that therefore show stronger correlations. Furthermore, the results indicate that the pairs of values showing the strongest correlations depend on the characteristics of each individual company.
Table of Contents
Abstract ................................................................................................................................. i
1 Introduction ......................................................................................................................... 1
  1.1 Background ..................................................................................................................... 1
  1.2 Purpose .......................................................................................................................... 2
2 Literature review ............................................................................................................... 3
  2.1 Language and emotion on Twitter .................................................................................. 3
  2.2 Twitter sentiment analysis ............................................................................................. 4
  2.3 Twitter sentiment and the stock market ......................................................................... 5
  2.4 Twitter emotion analysis ............................................................................................... 5
  2.5 Emotional tweets and the stock market ........................................................................ 6
3 Method ................................................................................................................................. 8
  3.1 Data ................................................................................................................................ 8
  3.2 Tweet classification ........................................................................................................ 9
    3.2.1 The Sentiment Model ............................................................................................... 10
    3.2.2 The Emotion Model ................................................................................................. 11
  3.3 Shortcomings ................................................................................................................ 12
4 Results ................................................................................................................................... 13
  4.1 American Airlines .......................................................................................................... 13
  4.2 American Express ......................................................................................................... 15
  4.3 AstraZeneca .................................................................................................................. 17
  4.4 ExxonMobil .................................................................................................................... 19
5 Analysis ............................................................................................................................... 21
  5.1 Correlation between Twitter sentiment and stock returns ......................................... 21
  5.2 Correlation between Twitter emotion and stock returns .............................................. 22
  5.3 A linguistic perspective ................................................................................................ 23
6 Discussion ............................................................................................................................ 24
7 Conclusion ............................................................................................................................ 25
References ............................................................................................................................... 26
1 Introduction

In contrast to computers, it is difficult for humans to always make rational decisions and not let feelings and emotions affect them. It has been noted that there are a lot of anomalies in financial markets that would not appear if people were rational and if stock prices were based on all available information about a company and nothing more. That is what the efficient market hypothesis by Fama (1970) suggests. The reality is clearly more complicated than that and to be able to predict future stock returns, human emotions need to be considered.

How to measure human emotions? A possible solution is to turn to social media and microblogs such as Twitter where millions of people share their thoughts and opinions about a variety of topics. That includes publicly traded companies. Assuming that emotions affect people’s financial decisions, it is possible that the changes that occur in the emotion and sentiment expressed about a company on Twitter correlate with the changes in the company’s stock price. If this is the case, emotional tweets could prove useful for future stock market prediction.

1.1 Background

Twitter is a social media platform with over 300 million active monthly users and several hundred million tweets generated every day (Twitter, 2018). A tweet is a short message of up to 280 characters. Java et al. (2007) looked at why people use Twitter and divided tweets into four categories: daily chatter, conversations, sharing information and reporting news. This is a lot of invaluable data for researchers in many different fields. Stock market prediction is one field that has been trying to take advantage of this data to increase prediction accuracy.

Pagolu et al. (2016) and Kordonis, Symeonidis and Arampatzis (2016) have tried to use Twitter sentiment analysis to predict stock market movements. Both papers found a strong correlation between Twitter sentiment and the changes in companies’ stock prices. However, the first paper only looked at Microsoft and the other one at top-16 technology stocks according to Yahoo! Finance. That could be a problem because Bing, Chan and Ou (2014) discovered that their model was better at predicting the prices of IT stocks compared to for example the manufacturing industry.
Some researchers have taken sentiment analysis further and divided tweets into more distinct categories. One of the most famous papers in the field by Bollen and Mao (2011) analyzed public mood on Twitter and found that calm had the strongest correlation with stock market movements whereas sentiment was not particularly predictive. Zhang, Fuehres and Gloor (2011) who measured collective hope and fear on Twitter concluded that all kinds of emotional outbursts had a negative correlation with stock market indices.

Instead of comparing the changes in public mood with the changes in stock market indices, Liu (2017) chose to look at individual companies and measure the changes in emotions expressed about them. Tweets were classified into the six basic Ekman emotions: anger, disgust, fear, joy, sadness, surprise (Ekman, 1992). No significant correlations were found and that might have been due to several reasons. Firstly, the corpus that was used for training the emotion classifier was general-purpose and not domain-specific. Secondly, the data used for analysis had a lot of noise and many irrelevant tweets. My thesis is going to try to tackle these problems to see if that results in any significant correlations being discovered.

1.2 Purpose
The purpose of this thesis is to compare the correlation between Twitter sentiment and stock returns with the correlation between Twitter emotions and the changes in stock prices for several companies. I would therefore like to propose the following research questions:

1. What kind of correlation, if any, is there between Twitter sentiment and stock prices?
2. What kind of correlation, if any, is there between the six Ekman emotions expressed on Twitter and stock prices?

The rest of the thesis is organized as follows. In Chapter 2 previous research on the topic will be introduced. In Chapter 3 the method and material for data collection will be explained. In Chapters 4, 5 and 6 results will be presented, analyzed, and the compatibility of the results and previous research will be discussed. Chapter 7 is going to wrap up the results.
2 Literature review

In this chapter an overview of existing research in the field about the English language will be given. Firstly, linguistic properties of Twitter language and ways of expressing emotions on Twitter will be discussed. Thereafter, research on Twitter sentiment and its applicability for stock market prediction will be discussed. The last two subsections will look at emotions and how researchers have been able to relate them to stock market changes.

2.1 Language and emotion on Twitter

There are many different genres of internet language that differ from each other in various ways. Twitter represents the language of microblogging and social media. Crystal (2011) analyzed a sample of 200 Twitter posts – tweets – on the topic language to try to make generalizations about linguistic properties of Twitter language. He found that a variety of shortening techniques were used, for example, contractions, logograms, abbreviations and elliptical sentences, although sentence-final periods were in most cases present. The different shortening techniques combined with nonstandard punctuation – using ellipsis dots and the dash - make it difficult to analyze tweets syntactically. Another common feature of tweets was the occurrence of what Crystal calls minor sentences such as yeah, lol, wow. On the other hand, some tweets included relatively complex sentences. Mohammad (2016) described the language used on social media as full of misspellings, creatively-spelled words, hashtags, emoticons and abbreviations.

Many of the strategies mentioned by Mohammad (2016) are used to express emotions and social media platforms such as Twitter are a rich source of emotional content. Nevertheless, the emotions expressed on the internet are difficult to recognize because the cues available in face-to-face communication such as facial expressions, body language and tone are absent online. Apart from emoticons, in recognizing emotions on social media one is mostly dependent on words. This is a problem because Twitter emotions are mostly not explicit, the emotion conveyed by a word can differ depending on the context and several emotions can be expressed in one message (Mohammad, 2016).

Hancock, Landrigan and Silver (2007) conducted a study on expressing and recognizing happiness and sadness through Instant Messaging and discovered that four strategies were used for expressing them: disagreement, negative affect terms,
punctuation and verbosity. Conversation partners had no difficulties in recognizing the two emotions and in doing that, they mostly relied on the partner’s use of negations and exclamation points. It gets more complicated when there are more emotions to distinguish between. Roberts et al. (2012) noted that the inter annotator agreement they reached even after an annotation standard for categorizing tweets into emotion categories had been created was somewhat low but consistent with other works on emotion annotation.

2.2 Twitter sentiment analysis
A number of papers have looked at sentiment in tweets and different methods have been used to create sentiment classifiers that could divide tweets into two or three categories according to their positivity, negativity or neutrality. Many researchers have tried to avoid involving human annotators because manually annotating many tweets is laborious and expensive.

Pak and Paroubek (2010), for example, used positive and negative emoticons to extract positive and negative tweets whereas newspaper headlines were used as examples of neutral tweets. Even Bifet et al. (2011) relied on emoticons in their tweet classification which was then used to illustrate the crisis in Toyota in 2010 when the company had to recall millions of cars due to problems with accelerator pedals. The emoticon approach could work for creating a general-purpose corpus of annotated tweets and it did work in the Toyota case, but it is questionable if tweets about publicly traded companies, especially tweets with a financial nature, commonly include emoticons.

Koulompis, Wilson and Moore (2011) chose to look at hashtags that in their opinion indicated message polarity combining that with a dataset containing tweets with either positive or negative emoticons. Considering that tweets about companies often have a neutral tone, it is highly likely that most of the hashtags, when there are any, show neutral sentiment as well.

Furthermore, Zimbra et al. (2018) compared 28 academic and commercial Twitter sentiment analysis tools. Although none of the tools achieved an average accuracy of over 72%, domain-specific tools outperformed the general-purpose ones by over 10% which shows the importance of having different training data for different purposes. The methods described above could only be used to create general-purpose databases.
2.3 Twitter sentiment and the stock market

There have been some attempts to relate Twitter sentiment changes to changes in the stock market. All the papers reviewed below found a correlation between sentiment and stock market movements and their accuracy gives a good starting point for comparing sentiment with emotion when it comes to changes in stock prices.

Kordonis, Symeonidis and Arampatzis (2016) created their training corpus using positive and negative emoticons, an approach described in the previous subchapter, combined with a sentiment lexicon that gives a valence for each word in English. For top-16 technology stocks according to Yahoo! Finance an average accuracy of 87% was achieved in predicting the future movement of a stock. This is a promising result which makes it interesting to try to find other more distinct categories that tweets could be classified into that would result in an even higher accuracy. A possible shortcoming with that paper is that only technology stocks were covered. Bing, Chan and Ou (2014) trained their own sentiment classifier which also used a sentiment lexicon to give words sentiment scores. Comparing the predictive accuracy of their model for companies in different industries, the IT industry had the highest predictive accuracy when compared to companies in the fields of finance, media, energy, manufacturing and medicine.

Even other papers on the topic have mostly concentrated on technology stocks. Pagolu et al. (2016) analyzed 250 000 tweets about Microsoft and its products over a one-year period. Unlike in the studies mentioned above, human annotators were used to create a training corpus because according to the authors, sentiment classification is extremely field specific. The accuracy of their model for stock price movement was about 70%. It is possible that letting humans annotate more tweets than the 3216 that were used, would have resulted in a significantly higher accuracy. Finding a balance between using a lot of data and using field specific data to train sentiment classifiers is an obvious challenge which makes it important to develop large field specific corpora.

2.4 Twitter emotion analysis

As with sentiment classification, different researchers have used different approaches to classify tweets into categories representing different emotions. In this case, even the emotions used as categorization labels have been different.

Ekman (1992) suggested that there were basic emotions such as anger, disgust, fear, joy, sadness and surprise. These emotions share nine characteristics among which
some are unique to each emotion and some, for example, quick onset, are shared by all of them. Subsequent research has often used the six Ekman emotions in tasks requiring categorizing text into emotional categories.

Bann and Bryson (2013) analyzed six different sets of basic emotions proposed in the existing literature. An emotional corpus of tweets was created by searching for tweets containing any of 21 different emotion words. Analyzing 21 000 tweets showed that Ekman’s set was the most semantically distinct of the six. A new set of semantically even more distinct basic emotions was proposed, consisting of the following: accepting, ashamed, contempt, interested, joyful, pleased, sleepy, stressed. Nevertheless, research on emotional tweets continues to use the Ekman set.

Roberts et al. (2012) created a corpus of emotional tweets annotated with the six Ekman emotions plus love believing love would be a commonly occurring emotion in informal text. To find examples for each emotion, they chose 14 topics which they thought would show these emotions. Humans were used to annotate the training data. Being limited to the 14 chosen topics, the collected corpus was not representative of all tweets. Secondly, it is not possible to use that approach to create a training corpus when the object of interest is all tweets about a specific company because it is not certain that a company evokes all these different emotions in people tweeting about it.

Mohammad (2012) collected around 21 000 tweets that included a hashtag for any of the six Ekman emotions which then formed the Twitter Emotion Corpus (TEC). This corpus has similar problems to the one described above considering the purpose of this thesis. It only represents a small part of all tweets and it is unlikely that tweets about companies and their products include hashtags of Ekman emotions. As for sentiment classification, having a domain-specific corpus to train the classifier on plays a vital role.

It is also the case that emotions are not evenly distributed across Twitter and there are significantly more tweets showing joy compared to those showing fear (Kim, Bak and Oh, 2012). It is possible that the emotion distribution would be different for a corpus of tweets about companies.

2.5 Emotional tweets and the stock market

Although most of the research concerning classifying tweets into emotions has used the six Ekman emotions as categories, the picture is more varied when it comes to relating emotions to the stock market.
Bollen and Mao’s (2011) famous paper on the topic looked at over nine million tweets over a period of almost ten months and measured the public mood. Six types of mood were considered: calm, alert, sure, vital, kind and happy. Even the public sentiment in tweets was considered. They found that the calm dimension correlated with the changes in the Dow Jones Industrial Average index whereas the other moods did not show a significant correlation. Interestingly, they did not find a correlation between the public sentiment and the changes in the DJIA.

Zhang, Fuehres and Gloor (2011) measured collective hope and fear on Twitter by counting tweets including mood words. Hope and happy were used to find positive tweets and fear, worry, anxious, nervous and upset were used for collecting negative tweets. Their inference from the data was that all kind of emotional outbursts, positive or negative, had a negative correlation with stock market indices such as the DJIA and S&P 500. It was speculated that people post more emotional tweets in times of economic uncertainty.

It is true that the stock market is the most psychological near the bottom and near the top when investors are the most irrational and let emotions lead their financial decisions (Boström, 2017). As Bollen and Mao (2011) put it: “It has often been said that stock markets are driven by fear and greed.” Gilbert and Karahalios’ (2010) study confirmed that observation. They used a dataset of LiveJournal blog posts to create the Anxiety Index and discovered that high anxiety levels had a negative correlation with the S&P 500 index.

The studies mentioned above measured collective mood and looked at how it correlated with stock market indices. There is a limited number of papers on the correlation between emotions and the changes in the stock prices for individual companies. Liu (2017) has written on the topic and compared the six Ekman emotions for predicting future stock returns.

In the study, Mohammad’s Twitter Emotion Corpus was used to train an emotion classifier. The corpus and its potential shortcomings were discussed in section 2.3. In addition, Liu used a word-emotion lexicon which has English words associated with each of the basic emotions (Liu, 2017). As with the Twitter Emotion Corpus, it is questionable if tweets about publicly traded companies include emotionally loaded words.
Enrique Rivera’s NASDAQ 100 dataset with circa one million tweets containing any NASDAQ 100 company ticker was used for the correlation analysis. No significant correlations were found between any of the Ekman emotions and the NASDAQ 100 companies (Liu, 2017). One possible explanation to this might have been the unsuitability of the corpus used to train the emotion classifier. Another reason might have been noise in the NASDAQ 100 dataset where the presence of a company ticker did not always mean that the tweet was about that company.

In the end, Liu was able to create a trading strategy that resulted in profit by using tweet volume to identify events and positive and negative keywords specifically related to earnings announcement tweets to identify sentiment, because the emotion classifier turned out not to be as accurate as hoped. It was suggested that a crucial step further would be to create labeled training sets of financial tweets which would be more accurate in classifying tweets using financial language compared to using general-purpose datasets or sets from other domains (Liu, 2017). To check if there really is no correlation between the Ekman emotions and changes in stock prices or if Liu’s result was due to poor data, my thesis will try to tackle the potential problems in the study discussed and use a corpus of relevant tweets for both training and analysis.

3 Method

This chapter describes the processes of data collection and classification that were used to find answers to the proposed research questions. In addition, potential problems with the chosen methods will be discussed.

3.1 Data

A Twitter developer account was created to get access to Twitter premium APIs. Thereafter, searchtweets, which is a Python wrapper for Twitter’s Premium and Enterprise search APIs, was used to extract tweets with the Search Tweets API about four publicly traded companies between September 17 and September 28, 2018. In total, 5985 tweets in English were extracted of which 2138 mentioned American Airlines, 1290 American Express, 730 AstraZeneca and 1827 were about ExxonMobil. The reason for choosing to look at a two-week period was that the use of a free sandbox account meant that the number of requests to the API was limited and having more companies to
compare was chosen over having more data about a specific company. Yahoo! Finance was used to collect historical stock prices for the analyzed companies that are listed on either The New York Stock Exchange (American Express, ExxonMobil and AstraZeneca) or The Nasdaq Stock Exchange (American Airlines).

Data was collected about the following four companies: American Express, American Airlines, ExxonMobil and AstraZeneca. These companies were chosen because they are publicly traded large companies that at most had a few hundred tweets per day between September 17 and 28. Because of the request limit, it was important that the number of tweets would be low enough and not exceed the limit. It was also speculated that these companies operate in branches that people are likely to have strong opinions on and post emotionally loaded content about. To collect tweets, the company name was used as a search query, although for ExxonMobil, only “Exxon” was used. Using cashtags was also considered but that turned out to be impossible, because cashtags are only available in the Twitter Enterprise APIs.

For each tweet, the timestamp and the tweet text were saved, and retweets were filtered out. In addition, all the tweets were manually assessed and irrelevant tweets as well as duplicates were removed. Irrelevant tweets were ads, sponsor ads, spam messages, job offers and tweets that were not about the company. The last category included tweets mentioning American Airlines Center, for example.

3.2 Tweet classification
Considering that the purpose of this thesis is to compare using tweet sentiment and emotion in predicting stock price movements, two different approaches were used to classify tweets. The first approach was using tweet sentiment which could be either positive, negative or neutral. The second approach was to classify tweets into the six basic Ekman emotions: anger, disgust, fear, joy, sadness and surprise plus a category named none for tweets not expressing any emotion.

Manually annotated tweets were used as training data for the emotion and sentiment classifiers. Forms with multiple choice questions were used for annotation and the annotators were asked to choose which sentiment or emotion they thought each tweet expressed about the company it mentioned. Each tweet was annotated by five people and the emotion or sentiment that was chosen by the majority was assigned to the tweet. Tweets with no majority agreement were discarded: 10.5% of tweets surveyed for
sentiment and 19.3% of tweets surveyed for emotion did not get a majority for one category. All tweets were preprocessed by removing usernames, website addresses and punctuation marks, except for question and exclamation marks and several full stops next to each other. These punctuation marks were considered to possibly be relevant for tweet emotion. Thereafter, Google’s AutoML Natural Language program was used to train the classifiers. The AutoML Natural Language program developed by Google lets one create custom machine learning models and can be used by anyone, with or without previous machine learning experience.

3.2.1 The Sentiment Model
327 human annotated tweets were used to train The Sentiment Model of which 115 were negative, 107 neutral and 105 had a positive sentiment. Table 1 shows examples of tweets with different sentiment.

Table 1. Tweet examples for each sentiment.

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Sentiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>#Exxon joins the growing number of #oilandgas companies who are passionately supporting the fight against #climatechange</td>
<td>Positive</td>
</tr>
<tr>
<td>surprise...you thought we believed what you believe...really we're just shills for google and exxon</td>
<td>Negative</td>
</tr>
<tr>
<td>Exxon outlines possible small-business resource center for north Baton Rouge firms</td>
<td>Neutral</td>
</tr>
</tbody>
</table>

37 random text items from the dataset were used to test the trained model, which achieved a precision of 77.1% and a recall score of 73%. The higher the precision, the fewer false positives, and the higher the recall score, the fewer false negatives there are.

The model classified 82.4% of the negative test items correctly. The same values for neutral and positive tweets were 77.8% and 54.5%, respectively. This is to be kept in mind when analyzing correlation results. The average precision – which is a metric for model accuracy between 0 and 1 - for this model was 0.86. In the rest of the thesis, this model will be referred to as The Sentiment Model.
3.2.2 The Emotion Model

513 human annotated tweets were used to train The Emotion Model of which 110 did not express any emotion, 100 expressed surprise; and joy, disgust and anger had 101 examples each. Table 2 shows examples of tweets with different emotions. The emotions fear and sadness were discarded because neither of them had over 10 examples among the manually annotated tweets which means that the proportion of tweets expressing these emotions would almost always be 0%.

Table 2. Examples of tweets for each emotion.

<table>
<thead>
<tr>
<th>Tweet</th>
<th>Emotion</th>
</tr>
</thead>
<tbody>
<tr>
<td>@AmericanExpress Tried on many occasions now to get my new American Express platinum card added to my online account. On my first attempt, your rep transferred me to Quicken Loans. wtf? Second, third, and fourth attempts also failed for different reasons. I regret going with amex</td>
<td>Anger</td>
</tr>
<tr>
<td>AA CEO States he is against the use of AA Aircraft to deport illegals. I am not sure I want to support an Airline that chooses to defy the Laws of the USA. #IllegalAliens #BuildTheWall #MAGA</td>
<td>Disgust</td>
</tr>
<tr>
<td>My future is bright so long as @AstraZeneca keeps making Crestor. 😎</td>
<td>Joy</td>
</tr>
<tr>
<td>A guy on my American Airlines flight is in full Delta Airlines uniform...is this even legal?</td>
<td>Surprise</td>
</tr>
<tr>
<td>80% of the medications in our #pharmacy are generics purchased by NOVA ScriptsCentral and 20% are donated brand-name pharmaceuticals provided through RxPartnership of Richmond, #Virginia, Direct Relief and AstraZeneca. #WednesdayWisdom #Healthcare</td>
<td>None</td>
</tr>
</tbody>
</table>

Furthermore, less than 100 of the surveyed tweets were annotated with the emotions disgust (32 tweets) and surprise (24 tweets). To collect 100 examples for these two emotions, which is the minimum limit required by Google’s AutoML Natural Language, Mohammad’s Twitter Emotion Corpus (2012) was used. As explained in section 2.3, emotion hashtags were used to create that corpus. For this thesis, the hashtags were removed, and human annotators were used to validate the emotions expressed. If
the emotion assigned by most annotators matched the removed hashtag, the tweet was added to training data.

53 random text items from the dataset were used to test the model which had a precision score of 66.7% and a recall score of 64.2%. This means that The Emotion Model was less reliable in classifying tweets correctly compared to The Sentiment Model. The amount of test items that this model classified correctly was 60% for surprise, 72.7% for anger, 63.6% for joy, 66.7% for none and 55.6% for disgust. The average precision metric for this model was 0.68. In the rest of the thesis, this model will be referred to as The Emotion Model.

3.3 Shortcomings
The more relevant data there is to train machine learning models, the better they learn to distinguish between distinct categories. To train The Sentiment Model and The Emotion Model, the minimum amount of training items required by Google was targeted for reasons related to both time and cost of collecting the necessary training data. Also, it was impractical to collect all the necessary examples for the emotions of disgust and surprise from the extracted Twitter data because that might have meant letting humans annotate all of them. Instead, the approach described in the previous section was used.

Another possible problem with the data used for calculations in this paper is that the stock markets are closed on weekends which means that there is no stock price data for these days. To tackle that problem, an approach used by Goel and Mittal (2011) among others was used. The last known stock price for Friday was used as x and the closing price for the following Monday as y. Thereafter, the formula \((x+y)/2\) was used to fill in stock price data for Saturdays. For Sundays, the calculated closing price for Saturday was used as x instead. This might have influenced the correlation values presented in chapter 4. Due to that, it is important in future studies to look at a longer period so that more actual closing values are considered.

Lastly, there might be a difference between looking at all tweets about a company and only looking at the financial ones. The last category can be filtered out by using cashtags but those are only available when using Twitter Enterprise APIs. That means it was too costly to make that comparison for this thesis, but it could be worth considering in future studies.
4 Results

In this chapter, the distribution of tweets for all four companies analyzed is covered. In addition, correlation values are shown for each sentiment and emotion with stock price returns one, two and three days after the day tweets were counted. To find the strength of linear correlation between pairs of sentiment or emotion percentages and stock returns, an online Pearson correlation calculator was used (Stangroom, 2018). The percentages of each sentiment or emotion on day \(d\) were used as X values and the changes in stock prices one (\(d+1\)), two (\(d+2\)) and three (\(d+3\)) days after were used as Y values. Correlation coefficients that are statistically significant at \(p < 0.1\) – which means that the risk of the calculator finding a correlation when there is none is less than 10% - are in bold.

4.1 American Airlines
The number of tweets mentioning American Airlines ranged between 119 and 249 per day during the period September 17 to September 28, 2018. In total, 2138 tweets about American Airlines were classified by both trained models.

The Sentiment Model classified 16% of the tweets as positive, 50.7% as negative and 33.3% as neutral (Figure 1). For individual days, the proportion of positive tweets was between 9.9% and 22%, the proportion of negative tweets was between 39.3% and 61.2% and neutral tweets made up between 24.3% and 39.3% of all tweets.

![Figure 1. Percentage of tweets with different sentiment – American Airlines.](image)

Table 3 below shows the correlation between the percentage of positive, negative and neutral tweets for each day considered and the stock return one (\(d+1\)), two (\(d+2\)) and three (\(d+3\)) days after that. As can be seen, all the correlations are weak. Due to the small

---

<table>
<thead>
<tr>
<th>Table 3</th>
<th>Correlation Table</th>
<th>p-value</th>
<th>Significance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>Negative</td>
<td>Neutral</td>
<td>Positive</td>
</tr>
<tr>
<td>0.51</td>
<td>0.32</td>
<td>0.64</td>
<td>0.07</td>
</tr>
</tbody>
</table>
number of tweet sentiment and stock return pairs used to calculate the correlations, none of them are statistically significant at $p < 0.1$. This means that to confirm these correlations, a longer period must be looked at.

However, an interesting thing to note is the correlation between the proportion of neutral tweets and stock returns the next day. This is the highest of the calculated correlations and has a $p$ value of 0.158429 which means that it is worth taking a closer look at with more data.

### Table 3. Correlation between tweet sentiment and stock returns – American Airlines.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.0132</td>
<td>0.164</td>
<td>0.1271</td>
</tr>
<tr>
<td>Negative</td>
<td>-0.2652</td>
<td>-0.2635</td>
<td>-0.1671</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.4342</td>
<td>0.2876</td>
<td>0.1606</td>
</tr>
</tbody>
</table>

The Emotion Model classified 39.5% of all tweets about American Airlines with the label anger, 6% got the label disgust, 20.5% were classified as joy, 2% of the tweets were labeled as surprise and the label none was assigned to 32% of the tweets (Figure 2).

![Emotion by proportion - American Airlines](image)

Figure 2. Percentage of tweets expressing emotions - American Airlines.

For individual days, the proportion of tweets that were classified as anger by the model ranged between 29.4% and 55.4%, the proportion of tweets labeled with joy was between 16.4% and 28.1% and the proportion of tweets showing no emotion ranged from
20.2% to 40.9%. The relatively uncommon labels disgust and surprise changed from 3.2% to 9.2% and 0.4% to 4.0%, respectively. It is possible that this is because of the small number of field-specific examples that were used to train The Emotion Model to recognize disgust and surprise.

Table 4 below shows the correlation between the proportion of each emotion for each day and the stock return one, two and three days after. As in Table 3, the correlations are weak and not statistically significant at p < 0.1. Anger seems to be the most promising subject for an in-depth analysis with more data.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>-0.3812</td>
<td>-0.3716</td>
<td>-0.3807</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.1393</td>
<td>-0.051</td>
<td>-0.1945</td>
</tr>
<tr>
<td>Joy</td>
<td>0.1651</td>
<td>0.298</td>
<td>0.2385</td>
</tr>
<tr>
<td>Surprise</td>
<td>-0.0965</td>
<td>0.176</td>
<td>0.3397</td>
</tr>
<tr>
<td>None</td>
<td>0.3175</td>
<td>0.2403</td>
<td>0.2967</td>
</tr>
</tbody>
</table>

4.2 American Express
In total, 1290 tweets mentioning American Express were collected and classified. The number of tweets per day from September 17 to 28 was between 68 and 147.

47.6% of all tweets about American Express were assigned the neutral label by The Sentiment Model, 32.2% of tweets were classified as positive and 20.2% as negative (Figure 3). For one day, there were at least 21.9% and at most 50.3% positive tweets; minimum 10.2% and maximum 27.4% negative tweets and between 32% and 58.3% neutral tweets. Compared to American Airlines, American Express had a significantly lower proportion of negative tweets whereas there was an increase in the percentage of both positive and neutral tweets.

Figure 3. Percentage of tweets with different sentiment – American Express.
Table 5 below shows the correlation coefficients for pairs of sentiment proportion and American Express’ stock returns. Most of the correlations are weak and statistically insignificant. However, there are two coefficients in Table 5 showing moderate correlation and being statistically significant at p < 0.1. It looks like the percentage of positive and neutral tweets is moderately correlated to American Express’ stock returns two days after the tweets were published. Negative tweets that were the most uncommon ones for American Express show no such correlation.

Table 5. Correlation between tweet sentiment and stock returns – American Express.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>-0.3386</td>
<td>-0.5945</td>
<td>-0.3326</td>
</tr>
<tr>
<td>Negative</td>
<td>0.0274</td>
<td>-0.1456</td>
<td>-0.2046</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.3172</td>
<td>0.6857</td>
<td>0.4657</td>
</tr>
</tbody>
</table>

Of the 1290 tweets classified, The Emotion Model assigned 23.6% the label anger, 2.7% the label disgust, 29.2% the label joy, 4.1% the label surprise and 40.3% the label none (Figure 4).

For individual days, there were between 17.3% and 29% tweets expressing anger, between 0% and 4.8% tweets showing disgust, between 22.6% and 38% expressing joy, between 0.8% and 11.8% showing surprise and from 25.2% up to 54.8% not expressing
any emotion. Compared to tweets about American Airlines, those mentioning American Express expressed less anger and disgust and more joy, surprise or no emotion at all.

Table 6 shows the correlation coefficients for pairs of emotion percentage and American Express’ stock returns one, two and three days after the tweets were published. The correlations are weak and not statistically significant, although the proportion of disgust and stock return the day after are worth further analysis.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>0.0715</td>
<td>-0.2566</td>
<td>-0.0759</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.4806</td>
<td>-0.0265</td>
<td>0.0635</td>
</tr>
<tr>
<td>Joy</td>
<td>-0.049</td>
<td>0.048</td>
<td>0.1236</td>
</tr>
<tr>
<td>Surprise</td>
<td>-0.1591</td>
<td>-0.0847</td>
<td>-0.3519</td>
</tr>
<tr>
<td>None</td>
<td>-0.0406</td>
<td>0.1503</td>
<td>0.0954</td>
</tr>
</tbody>
</table>

4.3 AstraZeneca

Both The Sentiment Model and The Emotion Model classified 730 tweets mentioning the pharmaceutical company AstraZeneca. The number of tweets about AstraZeneca per day varied from 21 to 114.

The Sentiment Model classified 55.5% of the 730 tweets as positive, 11.9% as negative and 32.6% as neutral (Figure 5). For individual days, the percentage of positive tweets ranged from 31.3% to 72%, the percentage of negative tweets from 2.6% to 34.4% and the proportion of neutral tweets was between 17.3% and 44.9%.

![Sentiment by proportion - AstraZeneca](image)

Figure 5. Percentage of tweets with different sentiment – AstraZeneca.
When compared to American Airlines and American Express, AstraZeneca had a significantly higher proportion of positive tweets than the other two. The proportion of neutral tweets was very similar for American Airlines and AstraZeneca, but the proportion of positive and negative tweets showed opposite patterns.

Table 5 shows the correlation coefficients between pairs of sentiment proportion and AstraZeneca stock returns. There is a moderate and statistically significant correlation at $p < 0.1$ between the proportion of negative tweets and AstraZeneca’s stock returns three days later. The type of tweets showing correlation (negative) is the most uncommon one of all three.

**Table 7. Correlation between sentiment and stock returns – AstraZeneca.**

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>0.1503</td>
<td>-0.0646</td>
<td>-0.354</td>
</tr>
<tr>
<td>Negative</td>
<td>0.0133</td>
<td>0.3014</td>
<td>0.5627</td>
</tr>
<tr>
<td>Neutral</td>
<td>-0.2493</td>
<td>-0.2909</td>
<td>-0.183</td>
</tr>
</tbody>
</table>

The Emotion Model classified 2.5% of all tweets about AstraZeneca as anger, 1.9% as disgust, 32.7% as joy and 1.5% as surprise; 61.4% of the tweets did not express any emotion (Figure 6).

![Emotion by proportion - AstraZeneca](image)

**Figure 6. Percentage of tweets showing emotions – AstraZeneca.**

For each day, there were between 0% and 9.4% tweets labeled with anger, between 0% and 5.9% labeled with disgust, from 16.3% to 46.5% expressing joy and
from 49.1% to 75.5% not expressing any emotion. Compared to American Airlines and American Express, the proportion of tweets labeled with anger was much lower and the proportion of tweets not showing emotions formed a clear majority.

Table 8 shows the correlation values between percentages of different emotions for each day considered and AstraZeneca’s stock return one, two and three days later. All the calculated correlations for AstraZeneca are weak and statistically insignificant.

Table 8. Correlation between emotion and stock returns – AstraZeneca.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>-0.1866</td>
<td>0.1721</td>
<td>0.2203</td>
</tr>
<tr>
<td>Disgust</td>
<td>0.1986</td>
<td>-0.1196</td>
<td>-0.2705</td>
</tr>
<tr>
<td>Joy</td>
<td>-0.0373</td>
<td>0.1424</td>
<td>0.042</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.0625</td>
<td>-0.2143</td>
<td>-0.2759</td>
</tr>
<tr>
<td>None</td>
<td>0.0434</td>
<td>-0.1305</td>
<td>-0.0003</td>
</tr>
</tbody>
</table>

4.4 ExxonMobil

1827 tweets mentioning ExxonMobil between September 17 and September 28, 2018, were classified by both trained models. The number of tweets per day mentioning ExxonMobil varied from 104 to 277.

The Sentiment Model classified 10.5% of all tweets about ExxonMobil as positive, 51.8% as negative and the remaining 37.7% as neutral (Figure 7). This is relatively similar to American Airlines, although American Airlines had a slightly higher proportion of positive and a slightly lower proportion of neutral tweets. For individual days, the percentage of positive tweets about ExxonMobil varied between 6.2% and 14.6%, the percentage of negative tweets was between 39.5% and 59.2% and the proportion of neutral tweets was at least 30.4% and at most 54.3%.

![Figure 7. Percentage of tweets with different sentiment – ExxonMobil.](image-url)
Table 9 below shows the correlation coefficients for the proportion of positive, negative and neutral tweets about ExxonMobil and the company’s stock returns one, two and three days after the tweets were published. There is a moderate and statistically significant negative correlation between the percentage of positive tweets and ExxonMobil’s stock returns regardless of if the stock return one, two or three days later is considered. It is the most uncommon type of tweets that show this kind of correlation.

Table 9. Correlation between sentiment and stock returns – ExxonMobil.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>-0.525</td>
<td>-0.5673</td>
<td>-0.5969</td>
</tr>
<tr>
<td>Negative</td>
<td>-0.0826</td>
<td>-0.1745</td>
<td>-0.0985</td>
</tr>
<tr>
<td>Neutral</td>
<td>0.269</td>
<td>0.3569</td>
<td>0.3095</td>
</tr>
</tbody>
</table>

The Emotion Model classified 40.3% of all tweets about ExxonMobil as anger, 4.4% got the label disgust, 8% were classified as joy, 2.7% as surprise and the remaining 44.6% of the tweets got the label none (Figure 8).

Figure 8. Percentage of tweets expressing emotions – ExxonMobil.

For each separate day, the percentage of tweets expressing anger was between 24.8% and 48.4%, the proportion of tweets showing disgust was between 0.7% and 8%, the number of tweets labeled with joy varied from 5.1% to 10.4%, the proportion of surprised tweets was between 0% and 4.6% and the proportion of tweets not expressing
emotions was between 37.6% and 63.6%. Thus, tweets that had the label anger or none formed an absolute majority of all tweets about ExxonMobil.

Table 10 shows the correlations between percentages for all five emotion categories analyzed and ExxonMobil’s stock returns the next day, the day after that and three days after the tweets were published. As for the other three companies considered, the correlations are weak and statistically insignificant but as for American Express, the emotion of disgust should be analyzed with more data.

<table>
<thead>
<tr>
<th></th>
<th>d+1</th>
<th>d+2</th>
<th>d+3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>-0.0989</td>
<td>-0.2235</td>
<td>-0.2303</td>
</tr>
<tr>
<td>Disgust</td>
<td>-0.4543</td>
<td>-0.4275</td>
<td>-0.247</td>
</tr>
<tr>
<td>Joy</td>
<td>-0.326</td>
<td>-0.0461</td>
<td>-0.2152</td>
</tr>
<tr>
<td>Surprise</td>
<td>0.1682</td>
<td>0.2095</td>
<td>0.2036</td>
</tr>
<tr>
<td>None</td>
<td>0.2405</td>
<td>0.283</td>
<td>0.2756</td>
</tr>
</tbody>
</table>

5 Analysis

In this chapter, the found correlation results between Twitter sentiment and stock returns will be compared and analyzed. Thereafter, the possibility of using Twitter emotion analysis in predicting the stock market will be covered. Lastly, the found results will be discussed from a linguistic perspective.

5.1 Correlation between Twitter sentiment and stock returns

The Sentiment Model achieved an average precision of 0.86 and classified more than 75% of both neutral and negative test items correctly. With these metrics, The Sentiment Model has the potential to successfully compete with the best Twitter sentiment analysis tools reviewed by Zimbra et al. (2018) and shows that using human annotators to classify training data is justified, although it is costly and time-consuming. Furthermore, the result confirms that domain specific sentiment analysis tools outperform the general-purpose ones.

The correlation coefficients for different sentiments in combination with stock returns after different number of days differed from company to company. Nevertheless, for three out of the four companies analyzed, some moderate and statistically significant
correlations were found. American Airlines was the only company with no significant
correlations between tweet sentiment and stock return.

American Express showed a correlation between positive and neutral tweets and
stock returns two days after. Most of the tweets about American Express were either
positive or neutral. AstraZeneca and ExxonMobil showed a different pattern where the
most uncommon sentiment type was the one that correlated with stock returns. In
AstraZeneca’s case, the proportion of negative tweets correlated with stock returns three
days later. For ExxonMobil, there was a correlation between the percentage of positive
tweets and stock returns for all three days. Although The Sentiment Model had some
difficulty recognizing positive tweets and that might have influenced the results, the fact
that statistically significant correlations were also found with neutral and negative tweets
speaks for the existence of a correlation between Twitter sentiment and stock returns.

All the papers reviewed in section 2.2 found a correlation between tweet sentiment
and stock returns. Thus, the statistically significant correlation results found in this thesis
confirm that observation. More data is needed to get more reliable results, but the
correlation coefficients presented in the previous chapter indicate that the type and
strength of correlation depend on the characteristics of each individual company. That
might be one of the reasons why Bing, Chan and Ou (2014) found a higher predictive
accuracy for IT companies compared to companies in other fields.

5.2 Correlation between Twitter emotion and stock returns
The Emotion Model achieved an average precision of 0.68 and was thus less reliable in
classifying tweets into categories than The Sentiment Model. Annotation results when
gathering training data for The Emotion Model confirmed Kim, Bak and Oh’s (2012)
observation that emotions are not evenly distributed across Twitter and that there are not
that many tweets showing fear. In the collected dataset about companies, fear and sadness
appeared extremely rarely and were therefore discarded. Most tweets were annotated with
the labels anger, joy and none. However, these labels had in most cases different
proportions to the labels positive, negative and neutral and can therefore not be considered
equivalent.

All the correlation coefficients calculated for all four companies were weak and
statistically insignificant, but some pairs of emotion and stock returns are worth analyzing
with more data. These pairs were American Airlines and anger, American Express and
disgust and ExxonMobil and disgust. In these cases, if more data confirmed the presented correlation coefficients, these could possibly be used in stock market prediction.

To compare tweet sentiment and emotion, the first one showed stronger correlation to changes in the stock price. This contradicts the findings of Bollen and Mao (2011) who found a correlation between the dimension of calm and the changes in the DJIA but no correlation between tweet sentiment and the same index. It is possible that measuring collective emotions and comparing them to stock market indices gives different correlation results compared to looking at individual companies.

Due to the small amount of data used in this thesis, it is not possible to claim with 100% certainty that there is no correlation between tweets expressing Ekman emotions and changes in stock prices. It does, however, match Liu’s (2017) result who did not find any significant correlations. In this thesis, even after noise from the collected corpus was removed and mostly domain specific tweets used for training The Emotion Model, no statistically significant correlations were found.

5.3 A linguistic perspective
As described in chapter 2.1, microblog language is different from other genres of writings. The messages on Twitter are short, use different shortening techniques, nonstandard punctuation and creative language and are therefore difficult to analyze syntactically which also makes it challenging for machine learning models. In addition, emotions expressed on Twitter lack paralinguistic cues characteristic to emotions in face-to-face conversations.

Thus, it is not surprising that most of the approaches used for automatic sentiment and emotion classification discussed in chapter 2 relied heavily on words: tweets for the different categories were found by searching for specific words or hashtags or by using sentiment lexicons that give sentiment scores for English words. Considering that emotions are mostly not explicitly stated in tweets and that the same words can be used to express different emotions in different contexts (Mohammad, 2016), this possibly explains the relatively poor performance of general-purpose sentiment and emotion classifiers. Creating domain-specific annotated corpora goes a long way in solving the issue of context, as Liu (2017) showed when using keywords specifically related to earnings announcements to create a profitable trading strategy. Even so, it seems to be an
inevitability that the ways of expressing emotions on the internet are more limited compared to real life.

Using words, punctuation and emoticons might suffice to express a limited number of emotions; especially when the only distinction that must be made is between one positive and one negative emotion like in Hancock, Landrigan and Silver’s (2007) study. It might not be enough when a larger number of emotion categories needs to be distinguished between. Bann and Bryson (2013) did find the Ekman set of basic emotions to be the most semantically distinct of all the analyzed basic emotion sets but the differences could be hard to recognize when reading tweets. This is supported by the fact that nearly one fifth of the tweets surveyed for emotion did not achieve majority agreement, whereas the same figure for tweets surveyed for sentiment was about one tenth. This suggests that the found results for Ekman emotions were not necessarily due to lack of correlation but instead because increasing the number of categories made it harder to accurately recognize and label them.

6 Discussion

According to the results that were presented in chapter 4, there is no statistically significant correlation between tweets expressing any of the Ekman emotions and changes in stock prices. Pairs of tweet sentiment and stock returns showed stronger correlations. In general, these results are consistent with what other studies on the topic have found. There might be several reasons for these results.

Firstly, fewer categories usually means that the categories are more distinct from each other than a set of more categories. That in turn means that it is easier for machine learning models to learn to distinguish between the labels. The fact that The Sentiment Model achieved a higher average accuracy than The Emotion Model supports this hypothesis. In this case, the reason no correlations are found might be due to the difficulties machine learning models have when learning to assign labels and not due to the nonexistence of a correlation.

Secondly, it might be the case that there are emotions that correlate with stock returns but that these are not the Ekman emotions. Bollen and Mao (2011) found a correlation with calm. Gilbert and Karahalios (2010) found a correlation with their
constructed Anxiety Index, although they looked at LiveJournal blog posts and not at
tweets. The fact that the basic Ekman emotions continue to be the most widely used ones
in research does not have to mean that these are the categories that show the strongest
correlations.

From a linguistic point of view, the fact that one fifth of tweets surveyed for
emotion did not achieve majority agreement and The Emotion Model’s poorer
performance compared to The Sentiment Model speak for the importance of non-verbal
cues one has when trying to understand the other’s emotion in face-to-face conversation.
It appears that words, punctuation and emoticons can be used to express some differences
in emotion, but people do seem to have difficulties expressing and recognizing the same
amount of different emotions online as they do in real life. Thus, the number of basic
emotions online seems to be smaller than offline, possibly only consisting of positive,
negative and neutral sentiment that can be successfully conveyed with the means
available.

There are many possibilities for further research on the topic. The first step further
would be to use the method presented in this study but collect more domain specific
training data and look at a longer period than two weeks to get more reliable correlation
coefficients. Furthermore, since different companies and companies in different fields
seem to differ in which and how strong correlations they show, studying causality would
add important puzzle pieces to studying correlation between tweets and the stock market.
Lastly, comparing a corpus of all tweets about a company and only the financial ones
including cashtags might give different results and is therefore worth looking into.

7 Conclusion

A common feature for all four companies analyzed - American Airlines, American
Express, AstraZeneca and ExxonMobil – was that the correlations between Ekman
emotions and their stock returns were weak and statistically insignificant. All but
American Airlines showed correlation between sentiment and stock return, but the type
of sentiment and which day was looked at differed. Nevertheless, statistically significant
correlations were found.
Although more data is needed to confirm the correlations found, the results indicate that there is a stronger correlation between tweet sentiment and stock returns compared to tweets that express the basic Ekman emotions and changes in stock prices. More distinct categories did not result in stronger correlations and there could be several reasons for that. Thus, tweet sentiment seems to be a better stock market predictor than emotions. Considering previous studies written on the topic, this is a plausible result.

References


