Magnus L. Kirø May 13, 2014

Sentiment analysis of Tweets in correlation with financial investments

Work in progress,

to be completed by 1. jun 2014.

https://github.com/magnuskiro/master

Masters Thesis,

Artificial Intelligence Group
Department of Computer and Information Science
Faculty of Information Technology, Mathematics and Electrical Engineering

NTNU – Trondheim
Norwegian University of Science and Technology

Abstract

Background: As Twitter has become a global microblogging site, its influence in the stock market has become significant. This makes tweets an interesting medium for gathering sentiment. A sentiment that might influence trends in the stock market.

Motivation: If twitter can be used to predict trends in the stock market the casual investor would gain an advantage over the day-trader or the modern trading algorithms.

Another interesting aspect is the role of twitter in sentiment analysis. And how twitters role as a data source influences trends in the stock market.

Methods and experiments: Twitter is used as the data source. It provides easy access, lots of data, and many possibilities to utilise the available metadata.

To improve and verify the sentiment classification and trend comparisons we use a variation of methods. Simple statistical methods, such as counting positive and negative words. More advanced methods such as part of speech and other NLP related magic.We also explore the use of mete data such as location and language tags.

Results: Rough results of my research.

Conclusion: All OK ? No?

Acknowledgements

Acknowledgements goes here.

TODO Arivd + Pinar

Metadata

Metadata ?

Typically repositories, links to code, file downloads, websites etc. Maybe contact info.

Contents

List of Figures

Chapter 1

Introduction

1.1 What

TODO write what this thesis contains and what is the goal of the thesis.

What has been done and what was going to happen. What is this thesis about? What are we doing? What are the goals of this thesis? What is the setting for this thesis, the circumstances and environment of the work.

1.2 Why, Motivation

TODO write why I want to do this and why we want to look at these specific points.

Why we do this and the motivation we have for doing this. Why is this work done? Why do we benefit from this? Why do I want to do this? Why is this relevant for others?

1.3 Research questions

TODO rewrite if necessary. TODO change to subsections.

How do we determine the sentiment of a tweet?

Can we extract knowledge from tweets to find a sentiment?

We will look at the usefulness of tweets as a way to extract sentiment.

Which parts of a tweet is useful for the classification of a tweets sentiment?

Which methods are best to classify tweets?

How do we best find the sentiment of tweets?

How can twitter be used to aggregate a trend?

Can we build a trend based on information from tweets?

Can Twitter as a microblogging site be used as a data source in aggregation of trends.

Credibility, what sort of credibility level has to be attained to certify the quality of the trend prediction.

Which parts of twitter are most useful to generate a trend?

How does trends on twitter compare to technical analysis in the stock market?

Technical analysis compared with the tweet trend.

We will look at possible applications for the sentiment in the stock market.

Which twitter sources are most suitable for predicting the stock market trend?

In finance, the moving average is a result of technical analysis. This and other trend defining qualities of financial data is used to compile trends.

Twitter has data such as the amount of tweets posted today, the location where tweets are posted from, and which users has posted. Aggregated, these data become represents a trend.

Previously researchers have managed to predict direction of the market the next few days based on the volume of tweets.

We are interested in the correlation between trends on twitter and the moving average in finance. Hopefully this will give some insight of how the sentiment on Twitter influences the sock market.

1.4 Findings

TODO briefly outline what we have found in this thesis. What we figured out in this thesis.

1.5 Outline

TODO write where stuff are in this thesis. Should be short. The outline of the document and the description of what which part is about.

Chapter 2

Background and Previous Work

TODO write a summary of the newest techniques and inventions in the field of twitter research related to finance.

2.1 Twitter

Twitter is a social and information network. It's a real-time service for sharing and gathering small messages. These messages can represent everything form a persons opinion of ice cream, to the latest changes in the financial market or pictures from a Mars rover.

At the core of Twitter you have the Tweet. The Tweet is the 140 character message. These small pieces of information combined are the life line of Twitter. Tweets lets you communicate with other users, share photos and post all kinds of information. The small size of the tweets are not a hindrance for the flow of information. [1]

The fast growing messaging service handles 1.6 billion search queries every day. As of 2012 the 500 million users would generate 3.2 queries each on any given day. 340 million tweets were posted every day. [2]

Most medium and large companies have a presence on Twitter today. Posts can contain any type of information, from promotional content to service status to financial reports. [Jubbega, 2011, p8] says that 77 of the Fortune 100 companies have a twitter account.

Companies use twitter for feedback and customer relations. Questions can be asked with a specific hashtag. Or with an at sign to target a specific user. This makes it easy to filter the messages, and therefore easier to get in contact with the customer. Best Buy demonstrated the successfulness of twitter in

[1] About Twitter: https://twitter.com/about

[2] Wikipedia: http://en.wikipedia.org/wiki/Twitter

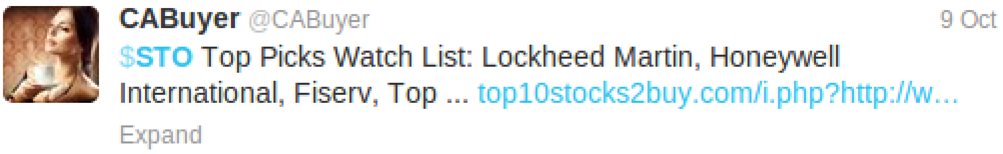
customer relations by answering questions with a specific hashtag. In 2009 they had answered nearly 20 thousand questions using twitter. [Li and Li, 2013, p1] Market Intelligence is also a major aspect of the microbloggin sphere.

Twitter represents one of the largest and most dynamic datasets of user generated content. Along with Facebook twitter data is in real time. This has major implications for anyone who are interested in sentiment, public opinion or customer interaction. [Speriosu et al., 2011]

A typical tweet contains about 11 words and provides an opinion or state of mind or a piece of information. Tweets can contain hashtags: something, user: @username, or other adaptations of prefixes such as $STO which represents a stock. The different prefixes or tags ($, #, @) easily distinguishes the content of the tweet. This also makes it easier to search and classify the content of tweets. Examples of tweets can be found in figure:2.1 and figure:2.2.

The retrieval of tweets seems like a challenge and impractical with a web scraper. But Twitter has made this easy by providing an API [3]. With the API you can write tweets and update the status of a user. But the best part of the API is that it provides search capabilities. To get a certain subset of all tweets, we can use the search function and view only the tweets we want.

On the front page of twitter we have the search function at the top right of the page. The search provides the ability to specify which types of tweets you want. And gives you the opportunity to find the information you are looking for.



CABuyer @CABuyer 9 Oct

$STO Top Picks Watch List: Lockheed Martin, Honeywell International, Fiserv, Top ... top10stocks2buy.com/i.php?http://w...

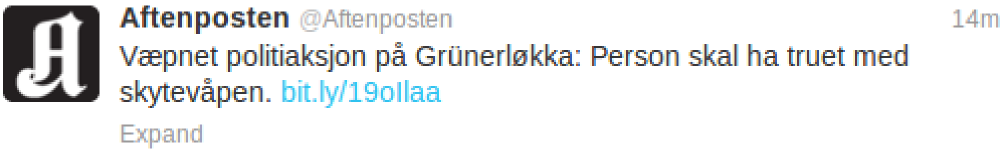
Expand

Figure 2.1: Typical tweet from Twitter.

2.2 Sentiment

Opinion mining on the web is not a new phenomenon. But in resent years it has become much more attractive to traders in the financial world. The usage of Twitter and other social media platforms is increasing. This means

[3] API: Application programming interface



Aftenposten @Aftenposten 14m

Væpnet politiaksjon på Grünerløkka: Person skal ha truet med skytevåpen. bit.ly/19oIlaa

Expand

Figure 2.2: Typical tweet from Twitter.

a surplus of raw data with easy access. Companies all over the world has started to use the social networks to their benefit. The use of information from social media has become part of the trend, although there are some drawbacks and shortcomings. Noise and garbage is one of them. The difficulty of the huge amount of available data is that it's difficult to find only the information relevant for your use. Even if you're right 80% of the time, the last 20% can prove devastating. [Stevenson, 2012]

Sentiment broadly refers to a persons state of mind. Based on the state of mind the person will do optimistic or pessimistic choices. A positive state of mind leads to optimistic judgements of future events, and a negative state of mind leads to pessimistic. [Doukas et al., January 10, 2010, p4]

The users may have different roles and intentions in different communities in the microblogging sphere, [Java et al., 2007]. A users intentions and its reasons for participation might be a factor in the sentiment analysis.

2.2.1 What is Sentiment Analysis

There are two main categories of approaches to sentiment analysis. The first is to use a classifier. The classifier can use methods such as naive Byes, maximum entropy or support vector model [Li and Li, 2013] This is typically a method where it would be natural to use machine learning of evolutionary algorithms to increase the classification correctness over time. The other is to use linguistic resources, such as corpora of negative and positive words. The developed linguistic resources are used to classify the sentiment of the text [Li and Li, 2013].

Li and Li has created a framework for sentiment analysis. The system consists of four main steps and is tested with experiments on twitter. First they do topic detection, identifying and extracting the topics mentioned in the tweet. Secondly opinions are classified. The polarity of the opinion is decided and the users impression is captured. Third. Credibility is assessed. This creates a better summarization of the expresser's credibility. Fourth, step one, two, and three is aggregated to reflect the true opinion and point of view. Combining the first three steps in the fourth results in a truer reflection of the expresser's opinion. [Li and Li, 2013]

One way of classifying tweets is to use predefined lexicon of positive and negative words. Consumer confidence and fluctuations of voting polls can be tracked in this way [Connor et al., 2010].

The work of [Diakopoulos and Shamma, 2010] describes a methodology for better understanding of temporal dynamics of sentiment. The system uses visual representation to achieve this. This is investigated in the reaction to debate video. Further [Diakopoulos and Shamma, 2010] detects sentiment pulse and controversial topics with the help of visualisation and metrics. [Diakopoulos and Shamma, 2010] used crowdsourcing[4] to classify batches of tweets. This was accomplished with Amazon Mechanical Turk, a crowdsourcing site[5].

[Barbosa and Feng, 2010] explores the problem of noise in biased and noisy data. They focus on noisy labels and add features to the tweets to increase the classification properties of the tweets. To filter out tweets that don't project a sentiment tweets are classified as subjective or objective. The subjective tweets are classified as positive or negative.

Classification of tweets can be generalised by using features. Features are elements such as unigrams, bigrams, and part-of-speech tags. An abstract representation of a tweet would be beneficiary to the classification. In this abstract representation [Barbosa and Feng, 2010] propose to use characteristics about how tweets are written and meta-information about the words in tweets. Meta-features and tweet syntax features are further features that can improve classification. Meta-features are information about the tweet, such as location, language, and number of retweets. The tweet syntax features are things such as hashtags, retweet, reply, links, punctuation and emoticons [Barbosa and Feng, 2010].

Another approach to the sentiment challenges with twitter is explored by [Becker et al., 2013]. They explore techniques for contextual polarity disambiguation and message polarity classification. Constrained and supervised learning is used to create models for classification. They describe a system that solves these tasks with the help of polarity lexicons and dependency parsers. Expanded vocabulary is one of the main aspects of their success, as they say in their findings: "We hypothesize this performance is largely due to the expanded vocabulary obtained via unlabeled data and the richer syntactic context captured with dependency path representations." [Becker

[4] Crowdsourcing is the practice of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers. http://en.wikipedia.org/wiki/Crowdsourcing

[5] Amazon Mechanical Turk (AMT): https://www.mturk.com/mturk/

et al., 2013]

In contrast to [Becker et al., 2013], [Speriosu et al., 2011] has used distant supervision and labeled propagation on a graph based data structure. The data structure represents users with tweets as nodes. And tweets with bigrams, unigrams, hashtags, etc as subnodes of the tweets. A label propagation approach rivals a model supervised with in-domain annotated tweets and outperforms the noisily supervised classifier and a lexicon-based polarity ratio classifier. [Speriosu et al., 2011]

2.2.2 Sentiment analysis in Finance

[Brown and Cliff, 2004, p2] writes the following on over-reaction of investors: "*He(Siegel (1992)) concludes that shifts in investor sentiment are correlated with market returns around the crash. Intuitively, sentiment represents the expectations of market participants relative to a norm: a bullish (bearish) investor expects returns to be above (below) average, whatever "average" may be."*. In the light of resent changes in the financial world and the use of sentiment from social media, the notion that opinions and sentiment of investors and market actors affect the market is not a new observation.

Use of sentiment can predict changes and momentum in the market. Bad news in an optimistic period creates cognitive dissonance in the small investors. This impacts the market by slowing down the selling rate of loosing stocks. [Doukas et al., January 10, 2010, p29] Further we can see that optimistic sentiment has a 2% monthly average return. While the investor sentiment is pessimistic we see a drastic reduction in returns. Down to 0.34%,[Doukas et al., January 10, 2010, p5]. After optimistic periods it is indicated that the monthly return is reduced to -0.49%. On the contrary there is no equivalent change after a pessimistic period, [Doukas et al., January 10, 2010, p6-7]. Momentum profits are only significant when the sentiment is optimistic, [Doukas et al., January 10, 2010, p29].

Hope and fear is used by [Zhang et al., 2011] to decide the movement of the market. The sentiment is aggregated to be hopeful or fearful. This basically focuses on positivity and negativity of the sentiment of that particular day. The daily sentiment is then compared to the market indicators of the same day to create a prediction of the market. [Zhang et al., 2011] finds that calm times give little hope or other emotions. Little turmoil results in few fluctuations in the market. And opposite, lots of emotions(hope, worry, fear), gives speed to the market.

[Brown and Cliff, 2004, p3] indicates that the sentiment does not cause subsequent market returns. For a short-term marketing timing this is bad news. However with the changes in social media over the last decade how is the situation today? With the microblogging sphere of today we can easily see the correlation of sentiment and the market indicators, [Jubbega, 2011]. But does the sentiment cause changes in the market-return? [Brown and Cliff, 2004, p3] also says that optimism is associated with overvaluation and subsequent low returns.

[Brown and Cliff, 2004, p] concludes that aggregated sentiment measures has strong co-movement with changes in the market. He also indicates that sentiment doesn't appear to be a good trading strategy. This, in the view of [Zhang et al., 2011], indicates a leap in sentiment research and what is possible with the microblogging of today.

2.3 Finance and Trading

The management of assets or liabilities and the management of funds over a period of time is called Finance. In finance the valuation of assets are time dependant. The same asset is not worth the same now and in a few minutes. Assets are priced based on expected returns and risk level. The three sub categories of finance are: personal, corporate and public. [6]. These categories describes very different parts of the financial world.

Trading is the action of buying or selling financial instruments. Financial instruments can be stocks, bonds, derivatives or commodities [7]. Trades takes place in markets, stock markets, derivatives markets or commodity markets.

Technical analysis in finance.

2.4 The Trend

The trend is the general opinion of the masses. As defined by the Free Dictionary: "The direction and momentum of a market, price, economy, or other measure. For example, if the price of a security is going mainly downward with only a few gains, it is said to be on a downward trend. Identifying and predicting trends is important finding the right moment to buy and sell securities. Trends are especially important in technical analysis, which recommends buying at the bottom of a downward trend and selling at the top of an upward trend." [8]

[6] Wikipedia:http://en.wikipedia.org/wiki/Finance

[7] Wikipedia:http://en.wikipedia.org/wiki/Trader_(finance)

[8] Dictionary description of trend: http://financial-dictionary.thefreedictionary.com/Trend

It's often talk about the fashion trend or the music trend when regular people talk about the trend. Or just the general direction of which a subject or subculture are moving.

Trends work in much the same way as opinions. An opinion is uttered then others start to think the same thing or feel the same way. The first group of people that move in the same direction are called trend setters. They are the people that show others how this trend works and what this trend is about.

On twitter we have lots of subcultures that all express themselves on their specific topic. Whether it's technology, art, finance or any other thing. In the sense of twitter we can take a step back and look at the content of messages and from there see if we can find common topics that people talk about, this being the topic of a subculture or a subspace of twitter. To get the trend we have to look at the content of the messages in a subspace. Given that the trend is the collective general collective opinion of the subspace we can look into this an see if we can find certain topics or areas of interest that aggregates to a trend.

When looking for twitter and trends there are few of far between those who work on it. No material or indication is found to suggest that trending on twitter is researched in regards to sentiment analysis of tweets.

Chapter 3

Data, retrieval and structure

This section describes the data sources, methods for acquisition, and the structure of the used data. 3.1 describes twitter and the mined tweets. 3.2 describe the different lists of words used in the classification process. And last the finance data is described in 3.3.

For each section the structure, characteristics, metadata and usage are described.

TODO chapter outline

3.1 Tweets

A tweet is a massage posted on twitter. The message can be up to 140 characters long and in many ways it resembles the well known SMS[1].

Tweets are posted to the users profile. When other people posts a previously posted tweet again it is called a retweet.

All users can follow other users on Twitter. Tweets from users you follow will appear in you stream of tweets on the main page of twitter.

3.1.1 Tweet Structure

Structure

There are a lot of metedata in the tweets. In fact most of the data in a tweet object is metadata.

The data we acquire from Twitter is in the JSON data format. JSON or *JavaScript Object Notation* is an open standard format that uses human-

[1] Wikipedia on Short Messaging Service:https://en.wikipedia.org/wiki/Short_Message_Service

readable text to transmit data objects consisting of attribute–value pairs[2].

A positive thing about the JSON data format is that we can directly evaluate it in python. By using literal evaluation in python the tweet object is interpreted as a dict. This makes the use of a tweet easy.

For an example of the data structure of a tweet, see appendix: D.

Content

A tweet is an astonishing compilation of data about who, where and when a tweet was posted.

As for the content we have the text, the message itself. With the content we have fields for all the links, all the emoticons, and all hashtags that are present in a tweet.

Every tweet is posted by a user. All the data of a user is also present for each tweet. Here we have data on follower count, profile images, friend count, time zone, and many other profile related items.

For the sharing of a single tweet we have data fields such as favorite_count and user_mentions. We also have favorited, retweet_count.

In addition we have the location of a given tweet. Where the tweet was posted, the name of the place, the coordinates of the tweet, the country, and the id of this place.

See all the different metadata types in appendix: D.

3.1.2 Twitter API

The twitter API is a convenient way for lots of people to access data from twitter. Tweets, streams, timelines, profiles and more are available through the api.

To provide easy easy access and conformity to industry standards, the api provides data in the JSON format.

While the api does not give access to 100% of the data from twitter, it gives a good representation of the tweets from the last 7 days.

Setup

To get access to the api there are a few requirements. You have to have a twitter account. You have to register the application you are going to use the api with, thereby getting authentication keys. Then you have to used the keys authenticate with twitter before you access the api.

For a simple guide to this we have http://datascienceandprogramming.wordpress.com/2013/05/14/twitter-api/ as a good example.

[2] Wikipedia on JSON:https://en.wikipedia.org/wiki/Json

The 4 authentication tokens you get from Twitter, app_key, app_secret, oauth_token,a nd oauth_token_secret, is used with the Twython library[3] as described below.

The simplest example of use of the api and the twython library can be described as follows:

- Authenticate towards twitter.
- Execute search query on twitter.
- Print ID's of all retrieved tweets.

The code of the example is as follows:

```
twitter = Twython(APP_KEY, APP_SECRET, OAUTH_TOKEN, OAUTH_TOKEN_SECRET)
results = twitter.search(q='Search query', count='15')

for status in results['statuses']:
    print status['id']
```

For more advanced use we have generators, and lots of parameters and api endpoints to use. Endpoints and search parameters will be described under section 3.1.2, *API endpoints options*. The twyhon framework and its advanced usage can be explored more in the code and in the documentation of the framework [4].

Restrictions

The api has some access restrictions. Or rather rate limitations. This is to be expected, as unlimited access would cripple the api.

Twitter limits request of a particular kind at 180 requests per 15 minutes. Which means that we can do 180 searches spread evenly over the time interval. Or we could do 180 requests as fast as possible and then wait. The rate limits can be explored thoroughly in the twitter documentation[5]

As for the practical implications of the limitations. They are not a problem in our case. We get more then enough data. By using a generator and pour out tweets we get 1000 tweets in about 60 seconds. But then we have to wait 15 minutes. This is suboptimal. As we will crash the program at access denied from the api.

[3] https://pypi.python.org/pypi/twython/

[4] https://twython.readthedocs.org/en/latest/usage/advanced_usage.html

[5] Twitter: https://dev.twitter.com/docs/rate-limiting/1.1

API endpoints

Twitter has made a lot of different endpoints available. The endpoints are divided into these categories: *Timelines*, *Tweets*, *Search*, *Streaming*, *Direct Messages*, *Friends & Followers*, *Users*, *Suggested Users*, *Favorites*, *Lists*, *Saved Searches*, *Places & Geo*, *Trends*, *Spam Reporting*, *OAuth*, and *Help*.

Of these categories we mainly use *OAuth*, *Help* and *Search*. Listing the endpoints used with parameters we get this list:

API endpoints:

Category	Parameter	Description
search	-	Used for acquisition of tweets. A query can take the following parameters:
	q	A UTF-8, URL-encoded search query of 1,000 characters maximum, including operators. Queries may additionally be limited by complexity.
	count	The amount of tweets acquired in each request. Standard = 15, max = 100.
	geocode	Get tweets close to these coordinates.
	lang	The language we want tweets in. Very limited by the number of people that speaks that language.
	locale	Language specific. If the query is in Norwegian we get tweets in Norwegian.
	result_type	mixed, recent or popular. This is the general mix of tweets returned in a search. Recent is the newest tweets, while popular are tweets that are retweeted a lot and tweets from users with many followers.
	until	Gets tweets before the given time.
	since_id	We get tweets posted after the given time.
	max_id	We get tweets with ids lower then the one given.
	include_entities	This parameter has no practical application to us.
	callback	An optional place where twitter can post tweets back to us.
authenticate	-	How we login and get access to the api.
	oauth_token	It is the authentication code or password to access the api.
rate_limit_status	-	The way we know if we are still inside the request limitations.
	resources	The elements that we want the status of. Currently that is search and help.

Searching

To perform a search we use the parameters described in the *API endpoints* table in section 3.1.2.

Mainly we need a query. The query is a normal search string where twitter will find tweets containing all words in the string.

```
query = "Finance Increase"
```

Further we can expand the search by using logic.

```
query = "Finance OR Investment AND Economy OR Growth"
```

The specific search used to compile the tweetset the dictionaries are based on is the following. Where we get all tweets containing one of the words: *Finance*, *Investment*, *Economy*, and *Growth*.

```
query = "Finance OR Investment OR Economy OR Growth"
```

Adding other parameters such as count and language we can further improve our search. To execute such a search we get python code like the following, where 'results' is a data structure containing tweets.

```
query = "Finance OR Investment OR Economy OR Growth"
results = twitter.search(q=query, count='15', language='no')
```

Mining optimization

The acquisition of tweets, or the mining operation went quite well. We got lots of tweets. But we ran into a problem. Retweets. In some cases we got up to 90% retweets in a mining session. Many of the tweets being essentially the same one, only retweeted multiple times.

When we are using tweets to create dictionaries, we do not want duplicate data that we do not need. Retweets are after all mostly duplicate data. So we removed the retweets and got a lot more different tweets.

As nearly all retweets start with 'RT' we can easily sort them out. Then we get a query like this.

```
query = "Finance OR Investment OR Economy OR Growth AND -RT"
```

When using the twython framework and its cursor function we get a continuous stream of tweets. A problem with this is that twython's cursor basically executes multiple searches. Thus yielding the same tweets multiple times, unless you change the search. So we change the search to accommodate this and use the *max_id* parameter.

```
query = "Finance OR Investment OR Economy OR Growth"
results = twitter.cursor(
                twitter.search,
                q=query,
                count="100",
                lauage=language,
                max_id='the id of the last tweet')

for result in results:
    print result
```

API Caveats

The main caveats of the twitter API is the request limitations and the limitations of the search engine.

As twitter says themselves: '*Please note that Twitter's search service and, by extension, the Search API is not meant to be an exhaustive source of Tweets. Not all Tweets will be indexed or made available via the search interface.*' and '*The limitations of the search engine of twitter indexes only about a weeks worth of tweets.*'[6]

Although this is not a big problem, coding and data acquisition could have been simpler. The solution for the week limitation was to broaden the search to include more words. This resulted in a more varied dataset, but also more tweets. We initially wanted to analyse tweets related to finance, so this was a bit of a negative point.

Further caveats of the request limitations. This means that we have to mine tweets over time. We should really have set this up on a server and mined tweets with a cron job every 15 minutesWikipedia on Cron: https://en.wikipedia.org/wiki/Cron.

3.1.3 Tweet sets

We ended up with two datasets to be used. The obama tweetset[7], which is a set of tweets containing around 1300 tweets about obama and the election of 2008/2009. And a self compile dataset, referenced as the kiro dataset, based on the words: *Finance*, *Investment*, *Economy*, and *Growth*.

[6] Twitter: https://dev.twitter.com/docs/faq#8650

[7] Neal Caren of University of North Carolina. Tweet file: http://www.unc.edu/~ncaren/haphazard/obama_tweets.txt

Search terms

The kiro dataset is based on only four words. This is a very limited part of twitter and in no way representative for it's full content. Neither does the search terms represent a full range of finance words.

An improvement would use a wide variety of finance words to mine tweets. This would improve the relation to finance and the relevance of the dictionaries afterwords.

Structure The self compiled datsets has one tweet per line. This is the JSON data object that is automatically imported into python.

As for the obama tweet set we only have the text, where we have one tweet text per line.

Caveats

Obama tweets is not ideal for sentiment analysis. To much political nonsense to comprehend. A political statement is not positive or negative. It is positive in the eyes on some and negative in the eyes of others. And there are people that think it is neutral because they do not care. Retweets are also present, so the actual data we get out of the tweet set is limited.

The kiro dataset has a lot retweets and neutral tweets in it. Therefore we have only used positive and negative tweets later, ignoring all the neutral ones. This gives us more relevant data to work with and less noise. Although we should have used the neutral tweets to improve the dictionaries.

3.1.4 Trend Data

When mining larger sets of tweets, the rate limitations and week limitation is a challenge. But solved by broadening the mining. The mining in itself is quite easy. Just execute a search and store all the new unique tweets.

To get a broad search we have a list of search terms. The terms are mostly usernames, but also some hashtags and other words. A drawback with the search terms are that they might not resemble the area of tweets we want to get. That are the finance related terms.

The trend tweets are stored in files named with the search term. Each term having it's own file with tweets. Then we sort the tweets by date and get files containing tweets for an individual date.

All the used search terms a stored in the file: '_search-terms' [8]. Most of the search terms are based on an article from 'Teknisk Ukeblad' where

[8] Search term file: https://github.com/magnuskiro/master/blob/master/code/trend/_search-terms

they list the twitter handles that are most significant in the oil industry. Since Norway has a lot of oil, the financial market and the trend is greatly dependent on it. And we can see if there are any relations between the compiled trend and the value of the stock exchange.

3.1.5 Problems, Shortcomings, and Possible Improvements

Retweets are a source of concern. The retweets does not give much in the sense of new and unique data. But they can provide a significance in sharing and importance of a given tweet. Retweets should be investigated more thoroughly in the future.

A shortcoming of the data mining is the search terms. Are the terms representative? Do we get good data or not? Are there other terms that are better suited to get accurate results? There are to many questions to ask about the data to rely on them to much. A wide array of tests and analysis should be done to remove this factor as a problem.

Another interesting point to consider is whether or not the choice of finance as the are of focus was smart. Is this area more or less difficult then other areas to navigate? We think that finance related tweets and news are more objective and has a firmer answer. And in total have less emotion and less room for interpretation.

3.2 Dictionaries

TODO introduction to dictionaries, of corpus whatever the name.
TODO the purpose of the dictionary
TODO use of the dictionaries.

3.2.1 Downloaded Dictionaries

TODO describe the distinctions of dl dict

Obama
TODO describe obama dictionary.

Loughran McDonald
TODO describe this dictionary Tim Loughran and Bill McDonald has a set of dictionaries available from the websites of University of Notre Dame [9].

List of Dictionaries:

- negative words General list of negative words. No particular category. Used for basic
- positive words This dictionary contains a small set of positive words. There are no general category for the words. The words are not directly related to finance.
- Uncertainty words
- litigious words
- modal words strong
- modal words weak

3.2.2 Compiled Dictionaries

The compiled dictionaries are based on two manually labeled tweet sets. My own, the kiro dataset, and the obama tweet set.

TODO ref the used datasets.

Details about the process of manually classifying tweets can be found in section 4.1.

TODO describe the dictionary compilation.

List of dictionaries:

TODO describe the different dictionaries

- Obama original, Monogram

  description

- LoughranMcDonald, Monogram

  description

- Obama original and LoughranMcDonald, Monogram, combined

  description

[9] TODO fiks tekst: nd.edu: http://www3.nd.edu/~mcdonald/Word_Lists.html

- Kiro, Monogram, self compiled

  description

- Obama, Monogram, self compiled

  description

- Kiro, Bigram, self compiled

  description

- Obama, Bigram, self compiled

  description

- Kiro, Trigram, self compiled

  description

- Obama, Trigram, self compiled

  description

3.2.3 Error analysis, removal of duplicate words

When creating the different dictionaries we remove duplicates from the positive and negative dictionary set. Words that are present in both the positive and negative dictionary is removed. By doing this we remove words that has no significance in the classification. But we also risk removing words with significance.

When looking at the duplicate words from the monogram dictionary based on the kiro dataset we found some errors. As a selection of words found, we have *dangerous, bad, go, inc, let, up, or, need, good, if, no, are, and, of, on, the, is, as.* Here we can see that the words *good* and *bad* are represented. Which is not good. By removing the words from the dictionaries we have removed significant words in further classification, thus reducing correctness of the algorithm. This is one of the drawbacks of the monogram dictionaries.

When looking at the removed duplicate words for bigram and trigrams we found no indication of the same problem. As the uniqueness of bigrams and trigrams are a lot greater we end up with very few duplicates and only duplicates that has no significance to the over all classification. Although we might have other unknown problems.

Most stop words and other insignificant words are removed with the removal of duplicate words. The same thing cannot be said about the bigram and trigram dictionaries. There we have no stop words present in themselves, but they are frequently part of other terms. For further improvements of classification with word counting and dictionary quality we should remove stop words, such as *as, is, on, off, and, or* etc, from the tweet/sentence before creating bi- and trigrams.

3.3 Finance Data

TODO obtaining the data(potential mining operations) TODO about the dataset, csv TODO potential problems

Chapter 4

Sentiment Classification

This section describes the experiments done. High level description and execution of experiments. Detailed execution and technical details in appendix.

Sentiment is described as ”an attitude toward something; regard; opinion.”[1] The sentiment is the perceived positivity of the message that the user tries to communicate. Sentiment is in many cases a personal thing, and can change from person to person or from setting to setting.

Some of the motivation for acquiring the sentiment of a tweet or a sentence, is that we can say something about a persons state of mind and from that predict behaviour. We want to use the sentiment to make smart decisions alter. As an example of usage it would be ideal to find a correlation between sentiment and stock exchange, thus making us able to increase revenue based with decisions based on the sentiment.

In this thesis we have two main ways of classifying tweets. Word counting and training a classifier. Both methods require dictionaries of positive and negative words, 3.2. In the classifier we use the dictionary to extract features from a tweet. And with the word counting we count the number of positive and negative words.

4.1 Manual Classification

TODO write about the obama tweet set. That most of the content is rubbish and very difficult to label positive or negative.

When labeling tweets manually there are a number of factors the complicates the process. Among them are the quality of the tweet, state of mind, language, and political affiliation.

[1] Sentiment - Dictionary.com: http://dictionary.reference.com/browse/sentiment?s=t

The quality of the tweet describes the content in many ways. Does the tweet contain links or hashtags? Are users mentioned?

State of mind for a person dictates that persons actions in many cases. This also has an effect on labeling tweets. A positive state of mind classifies more tweets negative as positive then others.

Political affiliation plays a huge role when labeling the obama tweet set. Do you like obama? Do you like Romney? Neither? This matters. The tweetset in itself is pro obama. So while labeling this tweet set we put aside political influence and looked at the core of the sentiment.

Note that all this happens in the brain during 3 to 60 seconds while reading the tweet text and is a very rough description of the thoughts happening. Labeling a tweet follows this algorithm in many ways:

Step	Thought	Description
1	Have we seen this tweet before?	Skip it or use previous classification.
2	#Hashtags or links present?	Some hashtags are automatically positive or negative. Also remove noise such as users and links.
3	Sarcasm?	If sarcasm is present, put up a warning flag saying that it is the opposite of step 4.
4	Special words?	Find a word that triggers positive or negative impression.
5	Done	Label tweet as positive or negative.

Result files

When classifying with manually we create files with the results. These files are comma separated files with three fields.

- Sentiment: Positive, neutral or negative. Represented by 1, 0 or -1.
- Tweet id, if it exists, else 'id'. It is a long number.
- The tweet text. In some cases sanitized.

4.2 Word count classification

TODO description of method. Simply put we count the positive vs negative words.

4.2.1 Classification

Polarity
TODO calculating the polarity.

The polarity of a given tweet is based on the difference in the amount of positive verses negative words.

pos/totw - neg/totw

Threshold
Threshold is the ratio of positive vs negative words that has to be present for a tweet to be either positive of negative.

The percentage of positive words minus the percentage of negative words gives the polarity value, or the positivity(how positive a tweet is) of a tweet. When actually deciding if a tweet is positive or negative we look at the polarity value. If the polarity value is above the threshold (polarity ¿ threshold) the tweet is classified as positive.

Examples of classification follows:
Example tweets:

- t1 = "good that he was decreasing badly"
- t2 = "he was good for increase"
- t3 = "good or bad"

Classification of t1:

- pos = 1 / 6 = 0.16666
- neg = 2 / 6 = 0.33333
- polarity = pos - neg = -0.1667
- threshold of 0 gives negative classification
- threshold of 0.1 gives negative classification
- threshold of -0.2 gives positive classification

Classification of t2:

- pos = 2 / 5 (to av fem ord) = 0.4
- neg = 0 / 5 = 0

- polarity = pos - neg = 0.4 - 0.0 = 0.4
- threshold = 0.4 positive
- threshold = 0.5 negative
- threshold = -0.1 positive

Classification of t3:

- pos = 1 / 3 = 0.3333
- neg = 1 / 3 = 0.3333
- polarity = pos - neg = 0
- threshold = 0 positive
- threshold = 0.1 negative
- threshold = -0.1 positive

TODO write this in full. Threshold average acc == 0.1 best. Further we found the best average threshold is 0.1. From the table under we have the threshold value, and the average classification accuracy among the 18 entries for each threshold.

-0.1 avg: 0.631616666667 -0.2 avg: 0.616144444444 -0.3 avg: 0.60595 -0.4 avg: 0.598822222222 -0.5 avg: 0.588833333333 -0.6 avg: 0.571155555556 -0.7 avg: 0.542322222222 -0.8 avg: 0.508316666667 -0.9 avg: 0.488105555556 0.0 avg: 0.647972222222 0.1 avg: 0.651616666667 0.2 avg: 0.65115 0.3 avg: 0.643016666667 0.4 avg: 0.63055 0.5 avg: 0.612277777778 0.6 avg: 0.593483333333 0.7 avg: 0.571272222222 0.8 avg: 0.545783333333 0.9 avg: 0.530727777778

4.2.2 Results

TODO write results from the classification of the different dictionaries. Dictionaries based on their own dataset naturally scores the best. When cross classifying we see that the bigram dictionaries score the best. With the trigram dictionaries nearly as good as the bigram dictionaries.

TODO write about the variations based on the datasets.

Threshold variations

By varying the threshold we hoped to find an optimal point of which we could separate tweets based on polarity. From the following graphs, figure 4.1, we can see no clear distinction of one value being better than the other ones.

In figure 4.1 we list the results of the experimentation with the threshold. Table 4.2.2 lists the dictionaries and dataset used for which graphs in figure 4.1. 'kiro dataset' and 'obama dataset' columns tells which dataset that was classified in which graph.

Dictionary name and description	kiro dataset	obama dataset
Obama original, Monogram	1	10
LoughranMcDonald, Monogram	2	11
Combined Obama original and LoughranMcDonald, Monogram	3	12
Kiro, Monogram, self compiled	4	13
Obama, Monogram, self compiled	5	14
Kiro, Bigram, self compiled	6	15
Obama, Bigram, self compiled	7	16
Kiro, Trigram, self compiled	8	17
Obama, Trigram, self compiled	9	18

4.2.3 Drawbacks

TODO write about drawbacks.

Dictionaries

Word positioning The dictionaries are based on the manually labeled tweets, so we can't create bi and tri-grams based on the position of a word in a tweet. Rather there is no way of automatically decide if a single word is positive or negative.

Threshold TODO write how many tweets that end to 0 when threshold is 0. pos-words == neg-words −¿ 0 but still the accuracy is quite high.

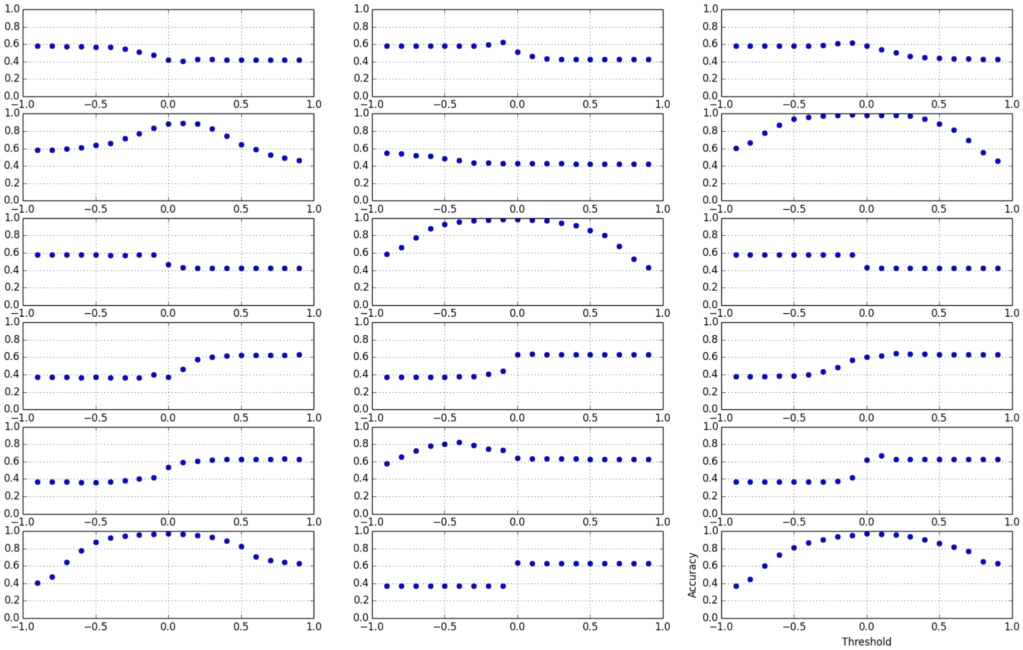


Figure 4.1: The graphs plot the different variations of threshold. Counting is columns first; top left is 1, top mid is 7, top right is 13.

Although we can se that in line 6 we have very few cases that this happens. From that we can can conclude that with the right choice of dictionary we don't have the problem of the threshold value.

0.0 null cases, 0.0 : 234 of 997 null cases, 0.0 : 543 of 997 null cases, 0.0 : 178 of 997 null cases, 0.0 : 53 of 997 null cases, 0.0 : 14 of 997 null cases, 0.0 : 7 of 997 null cases, 0.0 : 446 of 997 null cases, 0.0 : 28 of 997 null cases, 0.0 : 931 of 997 null cases, 0.0 : 335 of 1365 null cases, 0.0 : 854 of 1365 null cases, 0.0 : 345 of 1365 null cases, 0.0 : 233 of 1365 null cases, 0.0 : 37 of 1365 null cases, 0.0 : 462 of 1365 null cases, 0.0 : 52 of 1365 null cases, 0.0 : 1221 of 1365 null cases, 0.0 : 92 of 1365

4.3 With Classifiers

TODO write about drawbacks.

4.3.1 SVM

With both datasets. Using the self compiled monogram dictionaries.

Results from svm testing. which kernel works best?

4.3.2 Naive Bayes

With both datasets. Using the self compiled monogram dictionaries.

Results from testing with different dictionaries.

4.4 Comparison of classifiers

TODO highlights of the classifiers
TODO common denominators. commonalities.
TODO comparing the results of the classifiers.

4.5 Biased Mind

The datasets of manually labeled tweets are biased based on a persons personal opinion and state of mind in the moment of classification. Therefore we have to keep in mind that all further results are based on the assumption that everyone agrees on the manual labeling. This is of course a big potential source of errors to be explored more in 9.

4.6 Comments

TODO improvements
TODO drawbacks
TODO future work

4.7 Conclusions

TODO summarize the stuff we have learned shortly.
TODO mention future work.

Chapter 5

Trending

TODO outline this chapter.

5.1 The trend is your friend

TODO why do we want a trend? What is a trend. How do I define it? How do I use it in this context? And how does it work?

5.2 Trends on Twitter

TODO trends on twitter already. TODO what the trends are and what to use them for. TODO how we created the trend graph TODO Our own trend. compiled

How we can find trends on twitter and how we can use them.

5.3 Trending in Finance

TODO how can we see trends in finance if any at all? TODO hoped use of the trend compilation.

How trends are in finance. How we find them and what we use them for.

5.4 Comparing the trend and the moving average

TODO comparing the twitter trend, compiled trend and finance input. TODO any correlations among the trends? TODO what we get out of the trend we have created TODO if the trend created have any strong points? TODO What went wrong? TODO lessons learned of the trend?

Comparing a found trend on twitter with a found trend in finance. Are there correlations?

Chapter 6

The Code

TODO consider of this should be an appendix. TODO write the chapter introduction.

The main phases:

Data retrieval

Sentiment classification

Trend aggregation

Finance comparison

6.1 Description

TODO the purpose of the code and the level of completeness. Description of the prototype. It's purpose, and what it does.

6.2 Outline

TODO folder structure and what files are relevant.

6.3 Technology

TODO write down the technologies and tools used. The technology used, frameworks etc.

Python, twython

Twython (python bassert) Rameverk for tilkobling og integrasjon mot twitter apiet. Se https://github.com/ryanmcgrath/twython/tree/master/examples for eksempler. https://github.com/ryanmcgrath/twython

6.4 Functionality

TODO what functions do we have TODO how stuff works. How the system works and under which conditions.

6.5 Issues

TODO code shortcomings and potential improvements. Problems in the implementation and the general solution.

6.6 Usage, howto

TODO running of the code and what output to expect. TODO the shortcomings of windows and mac. How the prototype is used. User manual.

Chapter 7

Results and Discussion

All our results are in ones and zeroes. And further we discuss why there are only zeroes. And how that affects the outcome and future endeavors for the pirates we are.

Chapter 8

Conclusion

We worked hard, and achieved very little.

Chapter 9

Future Work

All the things I didn't have time to do my self.

9.1 data

Twitter API Look at the different endpoints and see how see how we can use them smartly. Twitter: https://dev.twitter.com/docs/api/1.1

search terms for the tweet sets the tweet sets should be expanded to use a wide range of finance words. then create dictionaries from the sets.

implications of special twitter content @users, hashtags, links? Does any of them matter much?

Trend Data The analysis of the trend data. There we ahve over 30k of tweets waiting to be analysed.

Bibliography

Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. 2010. Coling 2010: Poster Volume, pages 36–44, Beijing, August 2010.

Lee Becker, George Erhart, David Skiba, and Valentine Matula. Avaya: Sentiment analysis on twitter with self-training and polarity lexicon expansion. 2013. Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013), pages 333–340, Atlanta, Georgia, June 14-15, 2013. c 2013 Association for Computational Linguistics.

Gregory W. Brown and Michael T. Cliff. Investor sentiment and the near-term stock market. *Journal of Empirical Finance*, 11(1):1 – 27, 2004. ISSN 0927-5398. doi: http://dx.doi.org/10.1016/j.jempfin.2002.12.001. URL http://www.sciencedirect.com/science/article/pii/S0927539803000422.

B. Connor, R. Balasubramanyan, B. Routledge, and N. Smith. From tweets to polls: linking text sentiment to public opinion time series. 2010. URL http://www.cs.cmu.edu/~nasmith/papers/oconnor+balasubramanyan+routledge+smith.icwsm10.pdf.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing*. Association for Computational Linguistics, 2013.

Nicholas A. Diakopoulos and David A. Shamma. Characterizing debate performance via aggregated twitter sentiment. 2010. URL http://dmrussell.net/CHI2010/docs/p1195.pdf.

Qiming Diao and Jing Jiang. A unified model for topics, events and users on Twitter. In *Proceedings of the 2013 Conference on Empirical Methods in*

Natural Language Processing, pages 1869–1879, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/D13-1192.

John A. Doukas, Constantinos Antoniou, and Avanidhar Subrahmanyam. Sentiment and momentum. January 10, 2010. Updated May 20, 2011. Available at SSRN: http://ssrn.com/abstract=1479197 or http://dx.doi.org/10.2139/ssrn.1479197.

A. Java, X. Song, T. Finin, and B. Tseng. Why we twitter: Understanding microblogging usage and communities. 2007. 9th WebKDD and 1st SNA-KDD workshop on web mining and social network analysis, 2007.

Annika Jubbega. Twitter as driver of stock price. Master's thesis, BI Norwegian School of Management, 2011.

Yung-Ming Li and Tsung-Ying Li. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1):206 – 217, 2013. ISSN 0167-9236. doi: http://dx.doi.org/10.1016/j.dss.2013.01.023. URL http://www.sciencedirect.com/science/article/pii/S0167923613000511.

Jianfeng Si, Arjun Mukherjee, Bing Liu, Qing Li, Huayi Li, and Xiaotie Deng. Exploiting topic based twitter sentiment for stock prediction. 2013. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 24–29, Soa, Bulgaria, August 4-9 2013. c 2013 Association for Computational Linguistics.

Michael Speriosu, , Nikita Sudan, Sid Upadhyay, and Jason Baldridge. Twitter polarity classification with label propagation over lexical links and the follower graph. 2011. Proceedings of EMNLP 2011, Conference on Empirical Methods in Natural Language Processing, pages 53–63, Edinburgh, Scotland, UK, July 27–31, 2011. c 2011 Association for Computational Linguistics.

Timm O. Sprenger and Isabell M. Welpe. Tweets and trades: The information content of stock microblogs. December 2010.

Alexandra Stevenson. The social media stock pickers, Oct 23 2012. URL http://search.proquest.com/docview/1114502067?accountid=12870. Copyright - Copyright Financial Times Ltd. 2012. All rights reserved.; Last updated - 2012-10-23.

Xue Zhang, Hauke Fuehres, and Peter A. Gloor. Predicting stock market indicators through twitter "i hope it is not as bad as i fear". *Procedia - Social and Behavioral Sciences*, 26(0):55 – 62, 2011. ISSN 1877-0428. doi: http://dx.doi.org/10.1016/j.sbspro.2011.10.562. URL http://www.sciencedirect.com/science/article/pii/S1877042811023895. ¡ce:title¿The 2nd Collaborative Innovation Networks Conference - COINs2010¡/ce:title¿.

Appendix A

Processed Articles

A.1 Article template

file:*filename.pdf* citation:[]

* What did they use tweets for?
* What do they do?
* Event detection. Is the tweet about merging?
* How is learning present? * Is the approach statistical of NLP? * Where can this article be useful later?
* What does this article give answers to?

A.2 A Unified Model for Topics, Events and Users on Twitter

file: EMNLP192.pdf citation:[Diao and Jiang, 2013]

* What did they use tweets for?

Modelling topics, events and users in a unified way.

* What do they do?

LDA-like topic model, Recurrent Chinese Restaurant Process(discover events), Event-topic affinity vectors to model association (events–¿topics), Detecting meaningful events, Grouping events by topic. Tweet separation, topic(personal life)/event(global events)-tweet.

* Event detection. Is the tweet about merging?

Online and offline detection. Online= early detection of major events, efficiency is the main focus. Offline, focusses on getting all the relevant tweets. Don't assume every tweet is linked to an event. LDA?

* How is learning present?

* Is the approach statistical of NLP?

* Where can this article be useful later?

With event detection. Tweet separation. Financial tweets.

* What does this article give answers to?

A.3 Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data

file:*twitter-pos.pdf* citation:[Derczynski et al., 2013]

A.4 Tweets and Trades: The Information Content of Stock Microblogs

file:*SSRN-id1702854.pdf* citation:[Sprenger and Welpe, December 2010]

* What did they use tweets for?

"We find the sentiment (i.e., bullishness) of tweets to be associated with abnormal stock returns and message volume to predict next-day trading volume." [Sprenger and Welpe, December 2010]

* How are tweets used?

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

What twitter is used for, Twitter chapter.

Twitter incentives. [Sprenger and Welpe, December 2010, p4]

Description of bullishness, message volume and what it does etc.

[Sprenger and Welpe, December 2010, p52] suggest that stock microblogs can claim to capture key aspects of the market conversation.

Picking the right tweets remains just as difficult as making the right trades.

* What does this article give answers to?

Whether bullishness can predict returns. Whether message volume is related to returns, trading volume, or volatility. Whether the level of disagreement among messages correlates with trading volume or volatility. Whether and to what extent the information content of stock microblogs reflects financial market developments Whether microblogging forums provide an efficient mechanism to weigh and aggregate information

A.5 Exploiting Topic based Twitter Sentiment for Stock Prediction

file:*filename.pdf* citation:[Si et al., 2013]

* What did they use tweets for?

Predicting the stock market. Stock index time series analysis. daily one-day-ahead predictions.

* How are tweets used?

Dirichlet Process mixture model to learn the daily topic set. Vector regression. Topic-based prediction.

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

Twitter's topic based sentiment can improve the prediction accuracy. [Si et al., 2013, p28]

* What does this article give answers to?

A.6 Twitter as driver of stock price

file:*Twitter as driver of stock price-Jubbega.pdf* citation:[Jubbega, 2011]

* What did they use tweets for?

* How are tweets used?

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

General about twitter.

* What does this article give answers to?

A.7 Twitter Polarity Classification with Label Propagation over Lexical Links and the Follower Graph

file:*twitter polarity classification.pdf* citation:[Speriosu et al., 2011]

* What did they use tweets for?

Polarity classification. Positive/negative.

* How are tweets used?

With label propagation. Distant supervision. Graph based data structure. user–¿tweet–¿bigram/unigram/hashtag/etc.

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

Data section / sentiment /

Twitter section: What people uses twitter for.

Label propagation approach rivals a model supervised with in-domain annotated tweets and outperforms the noisily supervised classifier and a lexicon-based polarity ratio classifier. [Speriosu et al., 2011]

Twitter represents one of the largest and most dynamic datasets of user generated content.

* What does this article give answers to?

A.8 AVAYA: Sentiment Analysis on Twitter with Self-Training and Polarity Lexicon Expansion

file:*Sentiment Analysis on Twitter with Self-Training and Polarity Lexicon Expansion.pdf* citation:[Becker et al., 2013]

* What did they use tweets for?

Contextual Polarity Disambiguation and Message Polarity Classication * How are tweets used?

Constrained learning with supervised learning. Unconstrained model that used semi-supervised learning in the form of self-training and polarity lexicon expansion

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

Technical approach of models and sentiment analysis. State of the art on sentiment analysis with twitter.

* What does this article give answers to?

dependency parses, polarity lexicons, and unlabeled tweets for sentiment classification on short messages

We hypothesize this performance is largely due to the expanded vocabulary obtained via unlabeled data and the richer syntactic context captured with dependency path representations. [Becker et al., 2013]

A.9 Robust Sentiment Detection on Twitter from Biased and Noisy Data

file:*Robust Sentiment Detection on Twitter from Biased and Noisy Data.pdf* citation:[Barbosa and Feng, 2010]

* What did they use tweets for?
Sentiment analysis with focus on noise reduction.

* How are tweets used?
Noisy labels. Classifies tweets as subjective or objective. Then distinguishes the subjective into positive and negative tweets. Generalization of tweet classification. Meta-information. How tweets are written. More abstract representation.

* Where can this article be useful later?
Previous work, sentiment analysis, twitter, sentiment features. * What does this article give answers to?
It provides a better way to classify tweets.

A.10 Investor sentiment and the near-term stock market

file:*Investor sentiment and the near-term stock market.pdf* citation:[Brown and Cliff, 2004]

* Where can this article be useful later?
In the finance chapter for historic value and where we have come from.

[**?**, p2] on over-reaction of investors writes: " He(Siegel (1992)) concludes that shifts in investor sentiment are correlated with market returns around the crash. Intuitively, sentiment represents the expectations of market participants relative to a norm: a bullish (bearish) investor expects returns to be above (below) average, whatever "average" may be.". In the light of resent changes in the financial world and the utilisation of sentiment from social media, the notion that opinions and sentiment of investors and market actors affect the market is not a new observation.

[Brown and Cliff, 2004, p3] indicates that the sentiment does not cause subsequent market returns. For a short-term marketing timing this is bad news. However with the changes in social media over the last decade how is the situation today? With the microblogging sphere of today we can easily see the correlation of sentiment and the market indicators [TODO:Citation]. But does the sentiment cause changes in the market-return? [Brown and Cliff, 2004, p3] also says that optimism is associated with overvaluation and subsequent low returns.

* What does this article give answers to?

[Brown and Cliff, 2004, p] concludes that aggregated sentiment measures has strong co-movement with changes in the market. He also indicates that sentiment doesn't appear to be a good trading strategy. This, in the view of [Zhang et al., 2011] indicates a leap in sentiment research and what is possible with the microblogging of today.

A.11 Predicting Stock Market Indicators Through Twitter "I hope it is not as bad as I fear"

file:*Predicting Stock Market Indicators Through Twitter.pdf* citation:[Zhang et al., 2011]

* What did they use tweets for?

Gather hope and fear for each day using tweets. The sentiment indication of each day is compared to the marked indicators of the same day.

* How are tweets used?

Get the Positive/negative sentiment.

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

Address the question of intention of users on twitter. Good summary of things done in regards to twitter. (Might be a bit outdated, from 2010).

* What does this article give answers to?

That hope, fear and worry makes the stock go down the day after. Calm times, little hope, fear or worry, makes the stock go up.

A.12 Deriving market intelligence from microblogs

file:*Deriving market intelligence from microblogs.pdf* citation:[Li and Li, 2013]

* How are tweets used?

Companies use twitter for feedback and customer relations. Questions can be asked with a hashtag of to a specific user. This makes it easy to sort filter the messages, and therefore easier to get in contact with the customer. Best Buy demonstrated the successfulness of twitter in customer relations by answering questions with a specific hashtag. In 2009 they had answered nearly 20 thousand questions using twitter. [Li and Li, 2013, p1] Market Intelligence is also a major aspect of the microbloggin sphere.

* What did they use tweets for?

Sentiment classification. Topic detection, pos/neg classification.

* Event detection. Is the tweet about merging?

* Where can this article be useful later?

stateOf-twitter / state-sentiment / data /

* What does this article give answers to?

A.13 The social media stock pickers

file:*social_media_stock_pickers.pdf* citation:[Stevenson, 2012]

Opinion mining on the web is not a new phenomenon. But in resent years it has become much more attractive to traders in the financial world. Twitter and the social media's opinion is on the rise. This means a surplus of raw data with easy access. Companies all over the world has started to use twitter and readily available tweets to their benefit. Trading with social media is part of the trend. Although there are some drawbacks and shortcomings. Noise and garbage is one of them. It's difficult to accurately sort through all the data and get only the information relevant for your use. Even if your right 80% of the time, the last 20% can prove devastating. [Stevenson, 2012]

A.14 Sentiment and Momentum

file:*SSRN-id1479197.pdf* cition:[Doukas et al., January 10, 2010]

Not Twitter. Intra-day transaction data. Sentiment affects the profitability of price momentum strategies.

Use of sentiment can predict changes and momentum in the market. Bad news in an optimistic period creates cognitive dissonance in the small investors. This impacts the market by slowing down the selling rate of loosing stocks. [Doukas et al., January 10, 2010, p29]

Sentiment broadly refers to the state of mind a person has. Whereas negative of positive. Based on the current state of mind the person will do optimistic or pessimistic choices. A positive state of mind leads to optimistic judgements of future events. And a negative state of mind leads to pessimistic judgements. [Doukas et al., January 10, 2010, p4]

Further we can see that optimistic sentiment has a 2% monthly average return. While the investor sentiment is pessimistic we see a drastic reduction in returns. Down to 0.34%.[Doukas et al., January 10, 2010, p5] After optimistic periods it is indicated that the monthly return is reduced to -0.49%. On the contrary there is no equivalent change after a pessimistic period. [Doukas et al., January 10, 2010, p6-7] Momentum profits are only significant when the sentiment is optimistic. [Doukas et al., January 10, 2010, p29]

A.15 Is Trading with Twitter only for Twits?

Document Description: Blog post that describes the findings of the atricle [TODO art:ref].

The article has developed a strategy for trading stocks based on the bullishness of the tweet. [TODO glossary bullishness] Bullishness as I understand it is the same as the negativity of the tweet.

The article bases it's findings on three factors. The holding time of a stock (the time from you buy it until it's sold). The history of x days (how many of the past days are used to determine the tweet signal[TODO glossary tweet signal]). And the number of picks (how many stocks you hold a any given time).

It is also indicated that The main article has some good information about how tweets are built up. (Dollar-tagging for representation of a given stock, $AAPL)

Has a good figure of the system.

Indicates that the message volume and trade volume are related.

RefArticle: **??** Twitter mood Predicts the Stock Market.

Tags: buy/sell-signals, tweet signals, dollar-tagged, OpinionFinder, GPOMS,

A.16 From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series

The article uses polling data and two years of tweets as their data.

Basically a comparison of the opinion expressed on twitter and the opinion from phone enquiries.

Uses word counting to distinguish relevant tweets from the rest.

The twitter dataset is huge, typically billions of tweets.

Daily sentiment = positive tweets / negative tweets.

Appendix B

Tweet usage overview

Time series

Stock index time series analysis

Message volume

Message polarity, Bullishness/bearishness

Predicting the stock market

Predict next day-trading volume

Daily one-day-ahead predictions

Topic based prediction

Vector regression

Dirichlet Process mixture

Label propagation

Appendix C

Web resources

http://hum.csse.unimelb.edu.au/emnlp2013/papers.html

http://neuro.imm.dtu.dk/wiki/Twitter$_{s}entiment_{a}nalysis$

http://provalisresearch.com/products/content-analysis-software/wordstat-dictionary/sentiment-dictionaries/

http://www3.nd.edu/ mcdonald/Word$_{L}ists.html$

Appendix D

Tweet Data Structure

```
{
  u'contributors': None,
  u'truncated': False,
  u'text': u'WO2013149663A1 Estimating Anisotropic Resistivity Of A
Geological Subsurface $STO #G01V #G01V11 http://t.co/yyPFEJSdIj',
  u'in_reply_to_status_id': None,
  u'id': 390051769780142080,
  u'favorite_count': 0,
  u'source': u'<a href="http://w.pat.tc" rel="nofollow">TwittlyDumb</a>',
  u'retweeted': False,
  u'coordinates': {
    u'type': u'Point',
    u'coordinates': [
      5.7326363,
      58.9645836
    ]
  },
  u'entities': {
    u'symbols': [
      {
        u'indices': [
          77,
          81
        ],
        u'text': u'STO'
      }
    ],
    u'user_mentions': [
  ],
  u'hashtags': [
    {
      u'indices': [
        82,
        87
      ],
      u'text': u'G01V'
    },
    {
      u'indices': [
        88,
        95
      ],
      u'text': u'G01V11'
    }
  ],
  u'urls': [
    {
      u'url': u'http://t.co/yyPFEJSdIj',
      u'indices': [
        96,
        118
      ],
      u'expanded_url': u'http://w.pat.tc/WO2013149663A1',
      u'display_url': u'w.pat.tc/WO2013149663A1'
    }
  ]
},
u'in_reply_to_screen_name': None,
u'in_reply_to_user_id': None,
u'retweet_count': 0,
u'id_str': u'390051769780142080',
u'favorited': False,
u'user': {
  u'follow_request_sent': False,
  u'profile_use_background_image': True,
  u'default_profile_image': False,
  u'id': 163877216,
  u'verified': False,
    u'profile_text_color': u'333333',
    u'profile_image_url_https': u'https://si0.twimg.com/profile_images
/2309783804/355j4shhjrh4rqb5vsys_normal.jpeg',
    u'profile_sidebar_fill_color': u'DDEEF6',
    u'entities': {
      u'url': {
        u'urls': [
          {
            u'url': u'http://t.co/apqPEHN3aC',
            u'indices': [
              0,
              22
            ],
            u'expanded_url': u'http://w.pat.tc',
            u'display_url': u'w.pat.tc'
          }
        ]
      },
      u'description': {
        u'urls': [

        ]
      }
    },
    u'followers_count': 299,
    u'profile_sidebar_border_color': u'C0DEED',
    u'id_str': u'163877216',
    u'profile_background_color': u'C0DEED',
    u'listed_count': 8,
    u'profile_background_image_url_https': u'https://abs.twimg.com/
images/themes/theme1/bg.png',
    u'utc_offset': 32400,
    u'statuses_count': 247688,
    u'description': u'New patent information from WIPO.
IPC-based hashtags for realtime subject searching.',
    u'friends_count': 203,
    u'location': u'Tsukuba, Japan',
    u'profile_link_color': u'0084B4',
    u'profile_image_url': u'http://a0.twimg.com/profile_images/
2309783804/355j4shhjrh4rqb5vsys_normal.jpeg',
    u'following': False,
    u’geo_enabled’: True,
    u’profile_banner_url’: u’https://pbs.twimg.com/profile_banners/
163877216/1359154591’,
    u’profile_background_image_url’: u’http://abs.twimg.com/images/
themes/theme1/bg.png’,
    u’screen_name’: u’w_pat_tc’,
    u’lang’: u’en’,
    u’profile_background_tile’: False,
    u’favourites_count’: 10,
    u’name’: u’World Patents Mapped’,
    u’notifications’: False,
    u’url’: u’http://t.co/apqPEHN3aC’,
    u’created_at’: u’Wed Jul 07 14:08:23 +0000 2010’,
    u’contributors_enabled’: False,
    u’time_zone’: u’Tokyo’,
    u’protected’: False,
    u’default_profile’: True,
    u’is_translator’: False
  },
  u’geo’: {
    u’type’: u’Point’,
    u’coordinates’: [
      58.9645836,
      5.7326363
    ]
  },
  u’in_reply_to_user_id_str’: None,
  u’possibly_sensitive’: False,
  u’lang’: u’en’,
  u’created_at’: u’Tue Oct 15 09:49:23 +0000 2013’,
  u’in_reply_to_status_id_str’: None,
  u’place’: {
    u’full_name’: u’Stavanger, Rogaland’,
    u’url’: u’https://api.twitter.com/1.1/geo/id/dee2255bd015b52c.json’,
    u’country’: u’Norway’,
    u’place_type’: u’city’,
    u’bounding_box’: {
      u’type’: u’Polygon’,
      u’coordinates’: [
        [
          [
            5.5655417,
            58.884420999999996
          ],
          [
            5.8687141,
            58.884420999999996
          ],
          [
            5.8687141,
            59.0608787
          ],
          [
            5.5655417,
            59.0608787
          ]
        ]
      ]
    },
    u’contained_within’: [

    ],
    u’country_code’: u’NO’,
    u’attributes’: {

    },
    u’id’: u’dee2255bd015b52c’,
    u’name’: u’Stavanger’
  },
  u’metadata’: {
    u’iso_language_code’: u’en’,
    u’result_type’: u’recent’
  }
}
```