Magnus L. Kirø April 29, 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 meta data.

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.

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 this section.

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 this section.

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

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

What we figured out in this thesis.

1.5 Outline

The outline of the document and the description of what which part is about.

Chapter 2

Background and Previous Work

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 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

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

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

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.

Figure 2.1: Typical tweet from Twitter.

Figure 2.2: Typical tweet from Twitter.

[3] API: Application programming interface

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 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 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.

[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/

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]

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

[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

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

The data. What used, how and why. Acquired data how and from where.

System specification and solutions.

This section describes the data sources used in this thesis.

We describe the methods for acquiring data in each source.

Further, the structure, meta data, and characteristics of the data is described.

3.1 Tweets

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.

3.1.1 Tweet Structure

There are a lot of meta data in the tweets. In fact most of the data in a tweet object is not the tweet text itself.

Tweets directly usable in python

3.1.2 Twitter API

Description of the api and which options we have with the search.

Simple guide to access the twitter api: http://datascienceandprogramming.wordpress.com/201
api/

This is a list of all the API calls that is used in this thesis.

Search

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.

Mining optimization

3.1.3 Tweet sets

We are aiming to use multiple sets of 10k tweets.

This many tweets in each set. etc.

Three sets based on three different search terms. Split the original three sets into 10 subsets. Take one subset from each superset and manually classify them.Then automatically classify the 27 other sets.

3.1.4 Biased Data

Due to the necessity of a search term in the query, we only get tweets that are related to the given terms.

Further more the datasets of manually labeled tweets are biased based on my personal opinion and state of mind in the moment of classification.

3.1.5 Trend Data

3.1.6 Problems, Shortcomings, and Possible Improvements

The potential problems and shortcomings of the data.

3.2 Dictionaries

3.2.1 Downloaded Dictionaries

Tim Loughran and Bill McDonald has a set of dictionaries available from the websites of University of Notre Dame [1].

List of Dictionaries:

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

- 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

There are a lot of words that are not classified yet. Those words should be stored and classified later. This to improve the classifiers potential to correctly classify tweets.

3.2.2 Compiled Dictionaries

Monogram, obama 1 and 10
Monogram LoughranMcDonald 2 and 11
Monogram, combined Obama and LoughranMcDonald 3 and 12
Kiro, Monogram, self compiled 4 and 13
Obama, Monogram, self compiled 5 and 14
Kiro, Bigram, self compiled 6 and 15
Obama Bigram, self compiled 7 and 16
Kiro, Trigram, self compiled 8 and 17
Obama Trigram, self compiled 9 and 18

3.3 Finance Data

* obtaining the data(potential mining operations) * about the dataset * structure * 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.

TODO what is sentiment TODO why do we get it TODO how do we use it

* Experiment with the time frame of the prediction of the trend. * Typically the variation of time. What's the longest into the future that we can predict the trend?

4.1 Word count classification

TODO description of method. Comparison of positive vs negative words.

A basic approach to classification is to count words. Positive and negative words. To do this it is necessary have dictionaries to provide the classification of simple words. Section:3.2 describes the dictionaries used. Some of them are quite simple, while others are more extensive.

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

TODO write about the threshold variations.

TODO write results from the classification of the different dictionaries.

TODO write about drawbacks.

4.2 With Classifiers

TODO write about drawbacks.

4.2.1 SVM

With both datasets. Using the self compiled monogram dictionaries.

Results from svm testing. which kernel works best?

4.2.2 Naive Bayes

With both datasets. Using the self compiled monogram dictionaries.

Results from testing with different dictionaries.

4.3 Comparison of classifiers

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

4.4 Comments

TODO improvements TODO drawbacks TODO future work

4.5 Conclusions

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

Chapter 5

Trending

5.1 The trend is your friend

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

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

5.3 Trending in Finance

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

5.4 Comparing the trend and the moving average

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

Chapter 6

The Prototype

The main phases:

Data retrieval

Sentiment classification

Trend aggregation

Trend and mean average comparison

6.1 Description

Description of the prototype. It's purpose, and what it does.

6.2 Architecture

Top level design of how the software works and communicates through the layers.

6.3 Technology

The technology used, frameworks etc.

Python, javascript. Twython, Flask, angularjs, d3.

6.4 Frameworks

Frameworks I know about that might be useful. TODO: englishify this section.

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

Flask (python bassert) Minimalt rammeverk for webapplikasjoner. Dette gjør det enkelt å lage et API som leverer ferdir klassifiserte tweets og tidssegmenter. http://flask.pocoo.org/

AngularJS Generelt godt rammeverk for frontend på web. Kan kanskje brukes sammen med d3 for å presentere dataene. http://angularjs.org/

D3 Javascript rammeverk bassert på data. Skal visst være bra å bruke til å tegne grafer og slikt Jeg tenker at det kan være gunstig i presentasojnen av data og sammenlinkningene av moving average og den kalkulerte twitter trenden. http://d3js.org/

6.5 Structure

Code structure and which files are where.

6.6 Environment

How the system is run and under which conditions.

6.7 Issues

Problems in the implementation and the general solution.

6.8 Usage, howto

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.

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'
  }
}
```