Univerza v Ljubljani
Fakulteta za računalništvo in informatiko

Peter Gabrovšek

Analiza relacij med valutnim trgom in družabnimi omrežji

MAGISTRSKO DELO

MAGISTRSKI PROGRAM DRUGE STOPNJE
RAČUNALNIŠTVO IN INFORMATIKA

Ljubljana, 2017

University of Ljubljana
Faculty of Computer and Information Science

Peter Gabrovšek

Analysis of Relations Between Currency Market and Social Networks

MASTER'S THESIS

THE 2nd CYCLE MASTER'S STUDY PROGRAMME
COMPUTER AND INFORMATION SCIENCE

Ljubljana, 2017

Acknowledgments

I would like to express my gratitude to my thesis supervisor Assoc. Prof. Dr. Marko Robnik-Šikonja. The advice and guidance of Prof. Robnik-Šikonja was very valuable especially in the final stages of writing the thesis so that it could be concluded successfully.

I would like to thank Assoc. Prof. Dr. Igor Mozetič for bearing with me on a weekly basis and for constant support and guidance throughout the duration of the research.

My special thanks are dedicated to Prof. Dr. Andrej Bauer. I am gratefully indebted for his valuable help and the deeds that he extended to me when I needed them the most.

I would like to thank my colleagues Miha, Sašo, Darko, Petra and Jasmina at JSI for supporting me in my endeavours.

I would like to express my thanks and appreciation for the financial support from the H2020 FET project DOLFINS (grant no. 640772), and the Slovenian Research Agency (research core funding no. P2-103), which made my life so much easier.

I would like to thank my friends Saša, Žiga, Nina, Boštjan, Gašper, and Blaž (in random order) for brightening up my free time throughout the years of my study.

Finally, I must express my profound gratitude to my parents and my brother Gregor, who provided me with unfailing support throughout the years of study and through the research and writing of this thesis.

Peter Gabrovšek, 2017

”Every overdeveloped muscle has a major intellectual achievement on its conscience”

— Professor Abdullah Nightingale

Contents

Povzetek

Naslov: Analiza relacij med valutnim trgom in družabnimi omrežji

Družabna omrežja nezadržno rastejo, kot tudi njihov vpliv na vsa področja življenja. V naši nalogi smo raziskovali in analizirali odnose med valutnim trgom (imenovanim tudi Forex) in družabnimi omrežji. Posebej smo analizirali vsebine, ki zadevajo valutni tečaj EUR/USD zaradi velikega trgovalnega obsega tega para. Družabno omrežje Twitter smo analizirali v obdobju treh let.

Zbrali smo podatke o gibanjih valutnega tečaja, podatke o dogodkih in tvite. Tvite smo anotirali na podlagi vsebine oziroma pričakovanj avtorjev glede gibanja tečajev. Z analizo uporabnikov Twitterja in tvitov v povezavi z EUR in USD smo odkrili skupine uporabnikov z različnimi obnašanji. Na podlagi ugotovitev smo razvili model za klasifikacijo uporabnikov v skupine. Ustvarjen model predstavlja osnovo raziskave in ugotovitev.

Razvili smo spletno aplikacijo za prikaz podatkov, pregled podatkov in prikaz rezultatov analiz. Aplikacija nam omogoča hitro analizo in razhroščevanje le-te.

Dogodki povezani z EUR in USD imajo velik vpliv na gibanje tečajev. Analizirali smo povezave med gibanji tečajev na Forexu in sentimentom tvitov v času okoli dogodkov. Analizirali smo uspešnost napovedovanja različnih skupin uporabnikov v času dogodkov (npr. izjave vplivnih finančnih ustanov v EU in ZDA). Običajno se to izrazi v povečanem številu tvitov. Skupine uporabnikov se razlikujejo v moči napovedovanja valutnih gibanj. Nekatere skupine uporabnikov so pri napovedovanju valutnih gibanj celo uspešnejše

od poklicnih analitikov, kar potrjuje uspešnost našega modela.

Ključne besede

študija dogodkov, valutni trg, Twitter, umetna inteligenca, podatkovno rudarjenje

Abstract

Title: Analysis of Relations Between Currency Market and Social Networks

Social media are gaining an unprecedented momentum as well as impact on many areas of our life. In this thesis we investigate and analyse the relationship between foreign exchange market (also called Forex) and social media. Specifically, we analyse topics concerning EUR/USD exchange rate because of the large trading volume of the currency pair. We analysed Twitter in the span of three years.

We gathered the data on market movements, events, and Twitter posts. We annotated the tweets with authors' expectation of the Forex movement. By analysing tweets and users tweeting about EUR and USD, we discovered groups of users that behave differently and devised a model for classifying users into these groups. The model is the basis of our research and findings.

We developed a web application for visualisation and browsing the data and results of the analyses. This application enabled fast analysis and debugging of it.

Events connected to EUR and USD have high influence on market movements. We studied the relations between Forex movements and Twitter sentiment around the time of events. We analysed the performance of different user groups around the events (i.e. financial announcements in the USA and EU), which usually result in significant increase of Twitter volume. Predictive performance of the user groups varies in terms of describing market movements. Certain user groups give better results than professional analysts which shows efficiency of our user classification model.

Keywords

event study, Forex, currency market, Twitter, artificial intelligence, data mining

Razširjeni povzetek

Naloga analizira povezave med družabnimi omrežji (Twitter) in valutnimi gibanji (EUR in USD). Analiza je sestavljena iz zbiranja, čiščenja in prikazovanja podatkov, študije dogodkov in analize različnih skupin uporabnikov omrežja Twitter ter njihovih lastnosti.

Motivacija za naše delo so obetavni rezultati predhodnih analiz relacij med trgom delnic in družabnimi omrežji. Naš cilj je bil nadaljevanje in izboljšava predhodnih odkritij ter aplikacija metod na največjem in najbolj priljubljenem finančnem trgu - Forexu. Naš namen je bil vzpostavitev mehanizmov za analizo in napovedovanje valutnih gibanj.

I Kratek pregled sorodnih del

Forex je v podatkovni znanosti zelo raziskano področje. Raziskovalci uporabljajo metode strojnega učenja [1], kot so nevronske mreže[2, 3], metode podpornih vektorjev [4, 5] in druge [6]. Določene raziskave uporabljajo tudi študije dogodkov [7], statistične prijeme [8] in genetske algoritme [9].

Za analizo Forexa obstajajo številne metode. Vrsta raziskav uporablja tehnične analize [2], ki temeljijo na preteklih valutnih gibanjih. Še posebej obetavne rezultate kaže kombinacija Forexa in z njim povezanih novic. S prihodom družabnih omrežij so finančne novice (ki vsebujejo tudi napovedi gibanj finančnih trgov [10, 11, 12]) hitreje in bolj pogosto posredovane javnosti.

Analize sentimenta finančnih trgov lahko razdelimo na dva dela. Prvi tip analiz uporablja gibanje finančnih trgov za učenje klasifikacijskih modelov [13], drugi tip analiz pa s pomočjo ročno označenih besedil išče povezave med sentimentom novic in gibanji finančnih trgov [12, 14, 15, 16].

Zaradi velikih vplivov določenih dogodkov na finančne trge se število objav v družabnih omrežjih v času dogodkov močno poveča. Študije dogodkov v povezavi z borznimi trgi [10, 14] so bile osnova za številne raziskave. Več raziskav s študijo dogodkov je bilo opravljenih v povezavi z borznimi trgi kot s Forexom, kar nam je dalo dodaten motiv za nalogo.

II Metode analize odnosov med kompleksnimi sistemi

Analizirali smo medsebojne vplive med dvema kompleksnima sistemoma, Forexom in Twitterjem. Za analizo smo pridobili podatke o gibanju valut, o dogodkih in tvite povezane s Forexom, vse troje za obdobje treh let. Zaradi velike količine podatkov smo ustvarili spletno aplikacijo za vizualizacijo podatkov in analiz. Z analizo teh podatkov smo ustvarili model za klasifikacijo uporabnikov Twitterja v različne skupine. Opravljena je bila študija dogodkov oziroma njihov vpliv na finančne trge glede na različne skupine uporabnikov.

II.I Pridobivanje in obdelava podatkov

Za analizo smo podatke pridobili podatke v obdobju treh let (2014-2016) iz različnih virov. Podatke o valutnih gibanjih na minutni resoluciji smo pridobili na spletni strani *histdata.com*. Podatke o dogodkih povezanih z EUR in USD smo zbrali na spletni strani *Forex Factory*. Tvite nam je posredovalo podjetje Sowa Labs. Del tvitov so ročno označili. Tviti so bili razporejeni v tri skupine glede na pričakovano gibanje valutnega trga (pozitivno, nevtralno in negativno).

S pomočjo že označene podmnožice tvitov smo naučili model za avtomatično klasifikacijo tvitov. Model smo zgradili z metodo podpornih vektorjev. Model je bil validiran z 10-kratnim prečnim preverjanjem. Rezultati validacije modela so pokazali vidoko klasifikacijsko točnost. Po podrobnejši analizi smo ugotovili, da razlog za to leži v visokem deležu robotskih tvitov. Roboti imajo zelo omejen besednjak, kar omogoča enostavno določitev sentimenta.

V primerjavi z ostalimi analizami smo mi uporabili bistveno večji nabor ročno označenih tvitov za učenje našega modela.

II.II Arhitektura sistema

Obvladovanje podatkov je zaradi njihove velikosti in raznolikosti otežkočeno, kar je bil povod za razvoj spletne aplikacije za vizualizacijo in analizo podatkov. Spletna aplikacija je sestavljena iz več delov: podatkovne baze, zaledja, spletne strani in sistema za analizo podatkov.

Spletno stran smo razvili z uporabo modernih tehnologij. S *HTML5* in *CSS3* smo oblikovali spletno stran. Dinamičnost spletne strani smo zagotovili z *Javascriptom*, *jQueryjem* in *Angularjem*. Za vizualizacijo podatkov je bila uporabljena spletna knjižnica *Highcharts*.

Zaledni sistem je bil implementiran v programskem jeziku *Python*. Zaledni sistem povezuje podatkovno bazo (*PostgreSQL*) in spletno stran (angl. frontend).

Spletna stran omogoča podroben vpogled v gibanja valut, v dogodke povezanih s Forexom, sentiment tvitov in podrobnosti uporabnikov družabnega omrežja Twitter ter njihove tvite.

II.III Model za klasifikacijo uporabnikov

Pri analizi uporabnikov Twitterja smo jih razdelili v skupine glede na njihovo aktivnost, vplivnost, besednjak itd. Z upoštevanjem teh lastnosti smo oblikovali štiri glavne skupine uporabnikov: trgovalni roboti, relevantna podjetja, relevantne posameznike in spam uporabnike. Na podlagi lastnosti posameznih skupin smo zgradili klasifikacijski model. Model štejemo za najpomembnejši dosežek naloge. Model daje dobro osnovo tudi za nadaljnje študije, saj bi z njegovo razširitvijo lahko definirali še natančnejše skupine uporabnikov. Uporabili bi ga lahko tudi na drugih področjih.

Nadaljnja analiza tvitov je potrdila pravilnost našega modela. Upoštevali smo lastnosti tvitov posameznih skupin, ki se razlikujejo po statistikah retvitov, količini dnevnega objavljanja, količini objav v času dogodkov itd. Analiza je pokazala očitne razlike med posameznimi skupinami uporabnikov. Skupine relevantnih podjetij in posameznikov so bile velikokrat retvitane, kar nakazuje na njihovo relevantnost.

II.IV Študija dogodkov

Iz množice dogodkov povezanih z EUR in USD smo izbrali samo tipe dogodkov z veliko frekvenco in vplivom na Forex. Večina dogodkov je bila v povezavi z USD, dogodki povezani z EUR so bili v manjšini, kar je verjetno posledica razpršenosti objav po posameznih evropskih državah, medtem ko so objave v ZDA centralizirane.

Za namen študije dogodkov smo izračunali njihovo polarnost (pozitivni, nevtralni ali negativni) s pomočjo sentimenta tvitov v času dogodkov. Rezultati polarizacije se med posameznimi skupinami uporabnikov Twitterja zelo razlikujejo. Kljub velikemu vplivu na Forex v analizo nismo zajeli govorov voditeljev centralnih bank EU in ZDA, ker jih je težko numerično ovrednotiti. Ti dogodki so zanimivi za nadaljnje raziskave.

Študija dogodkov pokaže njihov vpliv na valutna gibanja in družabna omrežja. Povprečili smo kumulativne abnormalne donose dogodkov z isto polarnostjo. Krivulje donosov so se razlikovale, ko smo polarnost dogodkov določili z uporabo sentimenta tvitov različnih skupin uporabnikov, kar potrjuje pravilnost klasifikacijskega modela.

III Sklep

V nalogi smo izdelali model za klasificiranje uporabnikov Twitterja v različne skupine. Izkazalo se je, da imajo različne skupine različno napovedno moč glede gibanja valut, kar pomeni, da se določenim skupinam splača slediti, drugim pa izogniti. Model predstavlja dobro osnovo za uporabo na drugih področjih. Kljub dobrim rezultatom je še v začetni obliki, zato so možne še številne izboljšave in nadgradnje.

Razvita spletna aplikacija je dober in uporaben pripomoček za pregled in prikaz gibanja valut, dogodkov in tvitov. Aplikacija je nastala kot stranski produkt naloge, kljub temu pa ni omejena le na Forex in se jo lahko uporabi tudi na preostalih finančnih področjih.

Chapter 1

Introduction

This thesis analyses the relation between social networks (Twitter) and the currency market. The analysis consists of gathering and cleaning data, visualisation of data, event analysis and differentiation among the types of users.

The motivation for this thesis are previous studies[10] that analyse the relations between stock markets and the social media, which are showing promising results. Our aim was to extend and improve previous findings to the biggest and most popular financial market - Forex. Our intention was to establish mechanisms for analysing and predicting movements of financial markets.

1.1 Foreign Exchange Market

Foreign exchange market, also known as "Forex", is a global decentralised market for trading with currencies. By means of trading volume, it is by far the biggest market in the world. Daily trading volume exceeds 5 trillion dollars.

Forex has some properties, which distinguish it from other markets:

- Huge trading volume.
- Geographical dispersion as it allows trading anywhere in the world.

- Operating continuously 24 hours a day with exception of weekends.
- Leverage can be used to enhance one's profit.

Forex is attractive to many people since it is not taxable in some countries (one of them is Slovenia at the time of writing this thesis) unlike most other financial instruments.

When talking about Forex, some terms should be explained for easier orientation:

- Forex **brokers** are companies that offer trading with currencies to individuals and companies.
- **Leverage** is a technique involving the use of borrowed funds for trading with currencies (or any other asset in general). It increases the value of investments up to 500 times.
- **Spread** is the difference between bid and ask price. **Swap** is the amount of money, that is charged to investors over midnight.
- Brokers and other institutions provide **analyses** as a service to the investors. There are many types of analyses: fundamental analysis, technical analysis, wave analysis, ...

  Some analyses try to predict future movements based solemnly on historical price movements, while other analyses take into account other factors, such as financial events, global events, social media, ...

1.2 Twitter

Social media are nowadays deeply anchored in our society and are still gaining momentum. Methodically studying the social media can give us answers to many questions in various fields, such as politics, culture, trends, and financial markets.

Twitter is currently one of the most used social networks. Every day, around 100 million users post 340 million posts or messages called tweets. Tweets are 140 characters long at the most and are mainly public. The content on Twitter is very diverse, however in this thesis, we are interested in content related to euro (EUR) and US dollar (USD), only.

Although social networks provide a lot of useful data, there are also useless and misleading posts (trash), providing misleading and manipulative content, which we try to detect and avoid.

1.3 Related Work

Forex is a well-researched area in the data science. Researchers use machine learning methods [1], such as neural networks [2, 3], support vector machine [4, 5], and others [6]. There are studies involving event study methodology [7], statistics [8], and genetic algorithms [9].

There are various approaches to the analysis of the Forex market. A lot of research has been done on technical analysis [2], which means studying historical price movements. The combination of Forex and news shows promising results. With arrival of social networks, the financial news is coming sooner and more frequently to individuals. They contain information about future movements of financial markets [10, 11, 12].

Sentiment analysis of the market can be divided into two parts. The first one uses financial data to teach the models for labelling the texts [13] and the second one uses labelled news to search for relation between the sentiment and future financial market movements [12, 14, 15, 16].

Information can also be extracted from events, which cause increased volume in social network posting and in other news sources. There was more research using event study done on stock market [10, 14] in comparison to Forex.

1.4 Thesis content

In Chapter 2, we describe the collection and preparation of the financial, event and Twitter data. We adapted time zones to UTC, annotated tweets, excluded redundant data, and stored the data in the database.

In Chapter 3, we describe the development of the web application which enables browsing and visualisation of the data. The chapter also contains the description of the system.

In Chapter 4, we devise a user classification model, which is the core of the thesis. We analysed different aspects of the data, including event study. Chapter 5 concludes the work and represents ideas for the future work.

Chapter 2

Data

In this thesis, we analyse three types of data. **Twitter data** is needed to extract sentiment of the currency market of our interest, as well as analyse different groups of users and their reliability. **Financial data** is needed to test our models and to see how the market behaves in different situations. Finally, **event data** is needed to measure the impact of events on the currency market and to compare our models with the financial analysts.

The used data is gathered over the span of three years (2014-2016). We focus on two currencies, EUR and USD, since they have the largest share of the Forex market. The volume of tweets for these two currencies is large, therefore the analysis should be reliable.

The data is stored in the PostgreSQL database. Analysing data in the database takes considerably less time than analysing the data stored in files. A lot of data pre-processing is done directly in the database.

The data is diverse since it is obtained from different sources and in different formats.

Figure 2.1 shows the offset of different time zones and shifts, which occurs with daylight saving time (DST). All the data was converted from their own time zone to UTC to avoid daylight saving time related problems.

Figure 2.1 shows London and New York Forex sessions, which are two out of four most important ones. The other two are Sydney and Tokyo. Session represents the activity time of particular stock exchange. These four stock exchanges together ensure the 24 hour ongoing trading.

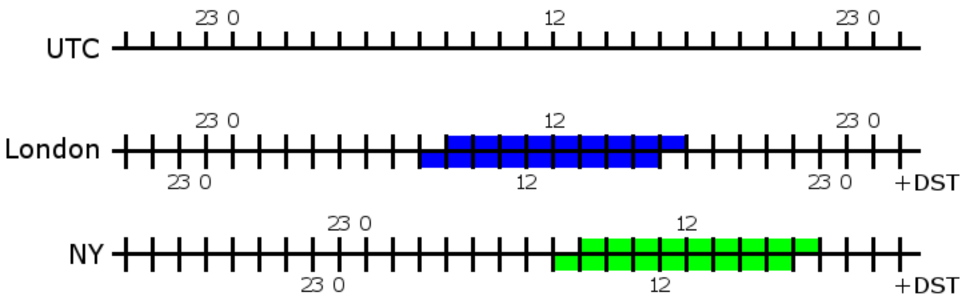

Figure 2.1: Time line, daylight saving time and Forex sessions.

2.1 Financial Data

Our main focus is the analysis of movements of the exchange rate between EUR and USD. The historical movement of a currency is usually described with one minute resolution with timestamps, open and close prices, and highest and lowest prices on that interval. This graphical representation is called a candle, as shown in Figure 2.2.

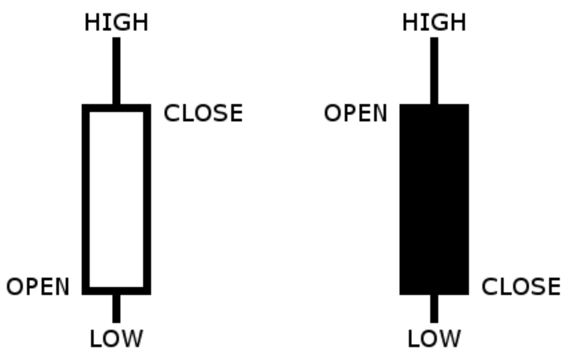

Figure 2.2: Example of bullish candle (left) - open price is lower than close price - and bearish candle (right) - open price is higher than close price.

The historical data of accumulated price movements with one minute resolution can be freely obtained on the internet. On the other hand, the precise historical data can only be purchased at a high price from ECB[1].

Depending on the orientation of the candle, we call it either a bullish or a bearish. The bullish candle has the opening price lower than the closing price, which means that on the given time interval, the price has risen. On the other hand, bearish candle has the opening price higher than the closing price, which means that the price went down in the given time interval. Both candles have high and low price, which represent the highest and lowest value that the price reached in that interval.

This financial data is aggregated to larger resolutions. Typical resolutions in Forex are one minute (m1), five minutes (m5), 15 minutes (m15), 30 minutes (m30), one hour (h1), four hours (h4), one day (d1), one week (w1), one month (mn) and one year (y1). In our thesis, we use resolutions up to 1 day.

Sometimes (but very rarely), currency data includes trading volume information, but this data usually does not represent actual trading volume, because every broker has its own volume information, and may not match global trading volume. For example, certain brokers operate in Europe only. The volume increases during working hours, so the movement after working hours does not match the volume of the global Forex movement. A hypothetical example is shown in Figure 2.3.

The financial data needed for this thesis was retrieved from histdata.com[2]. Timestamps of this data are in ET (Eastern Time Zone), which was converted to UTC.

The span of the gathered data starts on 1/1/2013 and ends on 31/12/2016. For the purpose of the event study analysis, we need a financial market model which requires data from before the analysed period.

[1]https://www.ecb.europa.eu/home/html/index.en.html
[2]http://www.histdata.com/download-free-forex-historical-data/

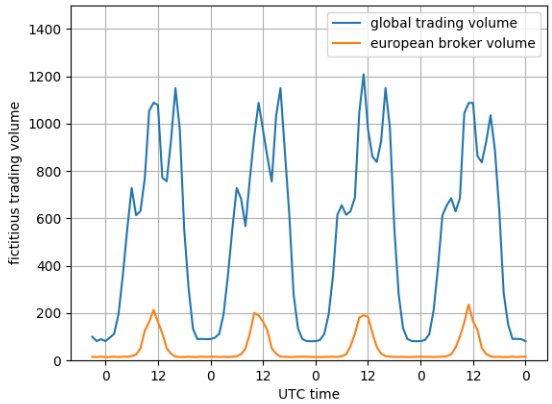

Figure 2.3: Comparison of global Forex trading volume and trading volume of European broker.

2.2 Twitter Data

Tweets related to currencies EUR and USD were acquired by the Sowa Labs[3] company. All tweets are from 14/2/2014 to 12/12/2016. Tweets were gathered through Twitter search API, where the query contained: "eurusd", "usdeur", "eur" or "usd".

For every currency and currency pair, there was a separate data source, which were then merged. The process of merging demands special care, since sentiment of tweets concerning EUR is usually opposite to the sentiment of tweets talking about USD. Prior to further processing of the tweets, the sentiment was converted accordingly. For example, positive sentiment of tweet mentioning currency pair USDEUR is converted to negative, because we are interested in the pair EURUSD, which has the opposite meaning of the sentiment.

All tweets are stored in PostgreSQL. They have the following attributes: tweet id, currency, time of creation, text, retweet count, language, retweet reference, user id, label (sentiment), group (bot, company, individual, ...).

[3]http://www.sowalabs.com/

2.2.1 Annotated Tweets

A subset of tweets was annotated by Sowa Labs. We received two sets of tweets: the training and the application set. The training set contains manually annotated tweets. Tweets were annotated by annotators, who were previously instructed on the process. Annotators used three possible labels: positive, neutral or negative. Labels represent annotators' opinion about the tweet's expectation about the price movement (rise, stay, fall) [17].

The annotation was done by about 20 annotators. Some tweets were annotated by more than one annotator to calculate inter annotator agreement and to react to wrong annotations.

2.2.2 Sentiment/Stance Model

We created a prediction model using 44,000 manually annotated tweets. The model was created with the help of the Latino[4] library, which uses double plane SVM algorithm. The model is verified by 10-fold cross validation. The results of validation are shown in Table 2.1. Surprisingly, the obtained model was very accurate. The reason for that will be explained in the later text (tweets from trading robots are the main reason).

Measure	**Value**
Accuracy	0.811 ± 0.014
Precision	0.814 ± 0.012
Recall	0.811 ± 0.014
F1	0.810 ± 0.014

Table 2.1: Results of the validation of tweet labelling model.

The procedure of labelling tweets is shown in Figure 2.4. Obtained model was applied to non-annotated tweets, so we ended up with 14,679,466 labelled tweets.

[4]https://github.com/LatinoLib/LATINO

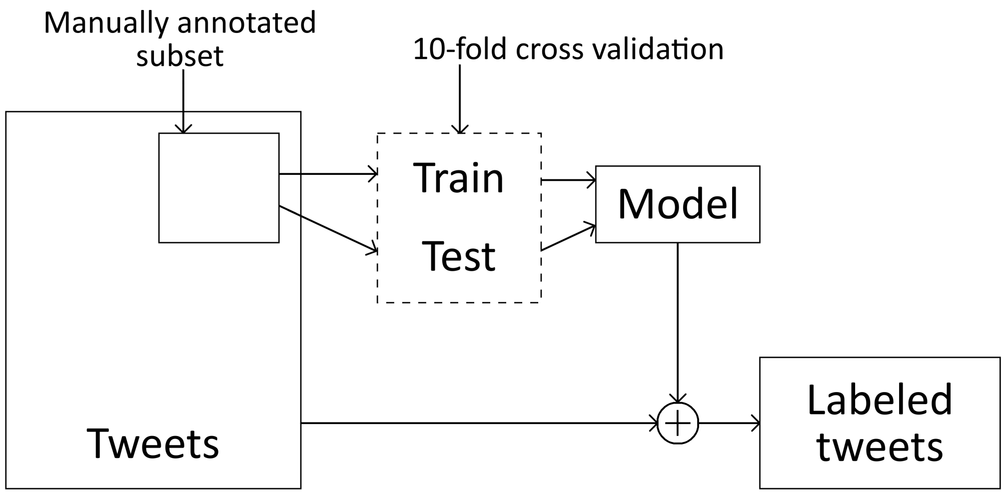

Figure 2.4: The process of labelling the tweets with the help of manually annotated tweets.

2.2.3 Tweet Processing

Tweets with calculated sentiment are aggregated on common Forex resolutions. Separate tweets are used for analysis of users to detect user relevance and possibly harmful and deceiving users. Users with high relevance (relevant companies and individuals) are used for further analysis.

We extracted user ids from tweets. We also calculated certain properties of users: user id, number of tweets, how many times users retweeted (and the ratio), how many times all user's tweets were retweeted (and the ratio), Hirsch index (defined in subsection 4.1.1), user group (bot, company, ...) and length of activeness counted in days.

We checked if tweets and users still exist by checking the response status. If the response status was 200, the tweet/user still exists, if it was 404, it does not. If the response was neither 200 nor 404, checking the existence of the tweet/user was postponed. The task of checking the existence of 14 million tweets was very time consuming since Twitter does not allow too many connections at the time, so the computer could run only five processes simultaneously.

2.3 Event Data

Events, which are related to Forex, are financial announcements provided by governments or central banks [18, 19, 20]. Events are especially interesting to study because their dates are known in advance and we can await them prepared. The Twitter and the trading volumes increase around the time of the events.

We found no service or web page with historical data for longer periods of time available for simple downloading. On average, there is usually a year worth of data available, instead of the needed 3 years. Finally, data was obtained from Forex Factory[5] using some hacking and trickery.

We gathered the details of the events in the time span from 1/1/2014 to 31/12/2016. We aligned the events with the UTC time.

Event data was obtained for all the currencies, but only events concerning EUR and USD were analysed. We excluded events of the type ALL, that concern all the currencies. Events of this type are globally important meetings: G20, G7, IMF, WEF, OPEC meetings, as well as the Jackson Hole Symposium.

Every event has a different impact. Concerning Forex Factory, there are four different impact types: high, medium, low, holiday. We retrieved all of them but in our analysis we used high impact events only. In a span of three years, there were between 730 and 770 (depending on sufficient amount of twitter data of different groups of Twitter users) high impact events considered, which we find sufficient for our analysis.

[5] https://www.forexfactory.com/calendar.php

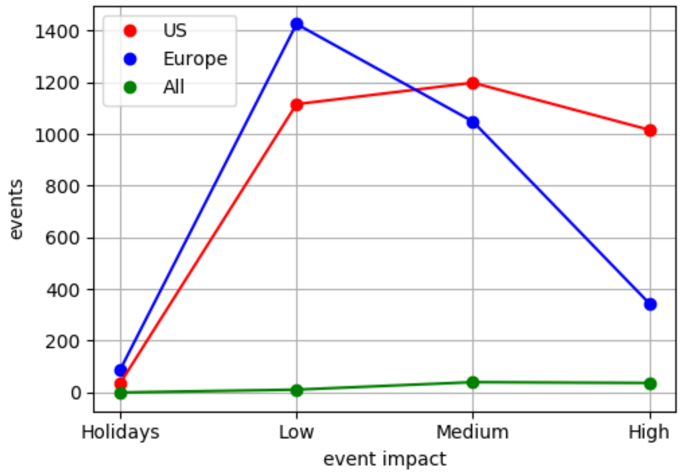

Figure 2.5: Number of events with respect to impact and region (currency).

Chapter 3

System Architecture

We implemented a web application that enables convenient browsing, displaying and analysing of the data.

The web application consists of two parts: back-end and front-end. The back-end communicates with the front-end and retrieves data from the database. The front-end receives response from the back-end and displays the content. The analysis is implemented in Python.

3.1 Database, Back-end and Front-end

The data is stored in the PostgreSQL database. The same database is used for both the analysis and the data visualisation.

Back-end is implemented in *Python 3*. Back-end retrieves the data from the database with *psycopg2* library. The data is sent to front-end with the help of *http.server* library. Libraries *JSON* and *pandas* are used to format the data to be usable in the front-end.

Front-end visualises the data. *HTML5* and *CSS3* are used for web page framework and the design. *JavaScript*, *jQuery*[1] and *Angular*[2] are used for communication with back-end and the dynamics of the web page. The pur-

[1]https://jquery.com
[2]https://angularjs.org

pose of the *Highstock*[3] and *datetimepicker*[4] is to select and display the time series of currency movements, Twitter sentiment and the events.

The schema of our application is shown in Figure 3.1.

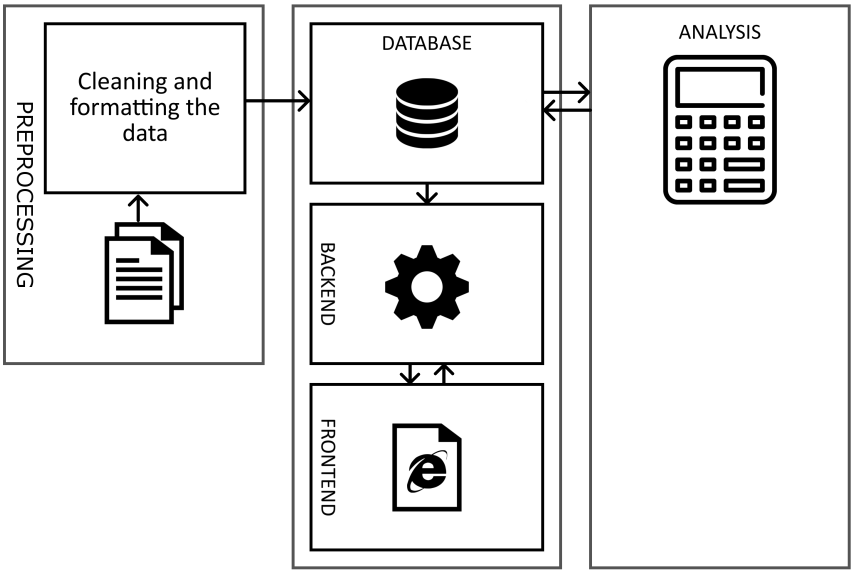

Figure 3.1: Schema of the application.

3.2 Use of Web Page

The web page consists of two subpages. The first subpage visualises the currency movements, events and the Twitter sentiment. The second subpage shows the details of the selected user and their tweets.

The following sections describe the use-case scenario of the advanced user.

3.2.1 Currency & Events

On the first subpage, we can browse EUR/USD currency pair movements. We can choose the time of our interest and the web page will display an

[3]https://www.highcharts.com/products/highstock
[4]https://github.com/xdan/datetimepicker

interval around it, as shown in Figure 3.2. Vertical red lines represent high impact events.

The web page enables us to choose the resolution for the currency graph, with which we would like to display the data. When we want to see the movements of the currency ratio around a specific event, we use fine resolution, e.g. m1, m5, m15. In case that we would like to see the global trend, we use a higher resolution, e.g. h1, h4, d1.

Figure 3.2: Currency movement and events in selected time window.

The currency movement is followed by volume and sentiment distribution of the tweets on the displayed interval (Figure 3.3). Sentiment can be viewed for different user groups defined in section 4.2.

The Twitter sentiment shows stacked volume of the negative, neutral, and positive tweets so we can see the relations among them. Together they represent a full Twitter volume.

At the bottom of the web page, there are details of the events (Figure 3.4) that are visible in Figure 3.2.

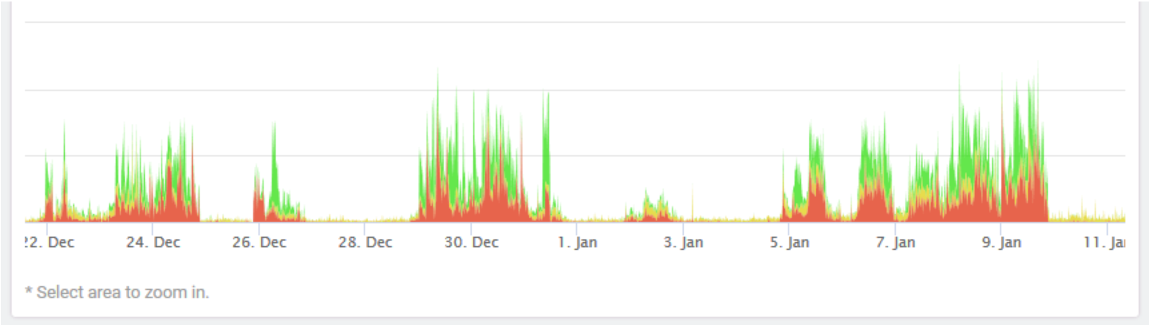

Figure 3.3: Twitter sentiment for all users in selected time interval.

Events

Load/Refresh

Details of events displayed on chart.

#	Time	Currency	Title	Actual	Forecast	Previous
1	2014-12-23 14:30:00+01:00	USD	Core Durable Goods Orders m/m	-0.4%	1.1%	-1.1%
2	2014-12-23 14:30:00+01:00	USD	Final GDP q/q	5.0%	4.3%	3.9%
3	2014-12-23 16:00:00+01:00	USD	New Home Sales	438K	461K	445K
4	2014-12-24 14:30:00+01:00	USD	Unemployment Claims	280K	291K	289K
5	2014-12-30 16:00:00+01:00	USD	CB Consumer Confidence	92.6	94.6	91.0
6	2014-12-31 14:30:00+01:00	USD	Unemployment Claims	298K	287K	281K
7	2015-01-02 16:00:00+01:00	USD	ISM Manufacturing PMI	55.5	57.6	58.7
8	2015-01-06 16:00:00+01:00	USD	ISM Non-Manufacturing PMI	56.2	58.2	59.3

Figure 3.4: Details of events on selected interval.

3.2.2 Twitter Users

Statistics of the users and tweets that were retrieved are shown in Figures 3.5 and 3.6. Details of the selected user are shown in Figure 3.5. We display the number of tweets retrieved, the number of retweets, how many times their tweets were retweeted, the Hirsch index (defined in subsection 4.1.1), active days, and whether Twitter account was still active at the time of checking (July 2017). There is also a link to their Twitter profile.

At the most 100 of their tweets are displayed. An example can be seen in Figure 3.6. We display the creation time of the tweets, their content, how many times they were retweeted (if at all), their sentiment and whether tweets were deleted (do not exist anymore).

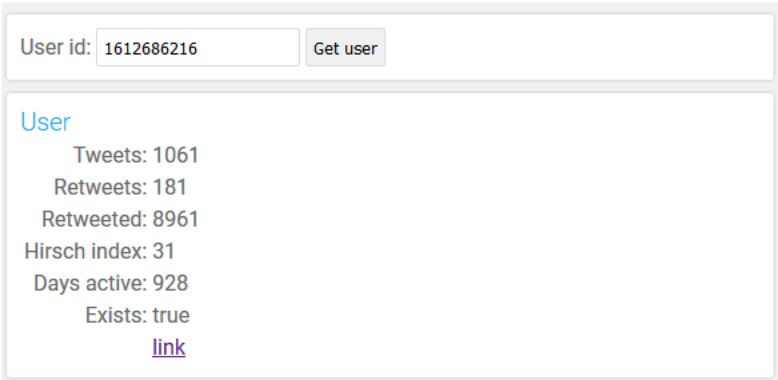

Figure 3.5: Details of the selected user.

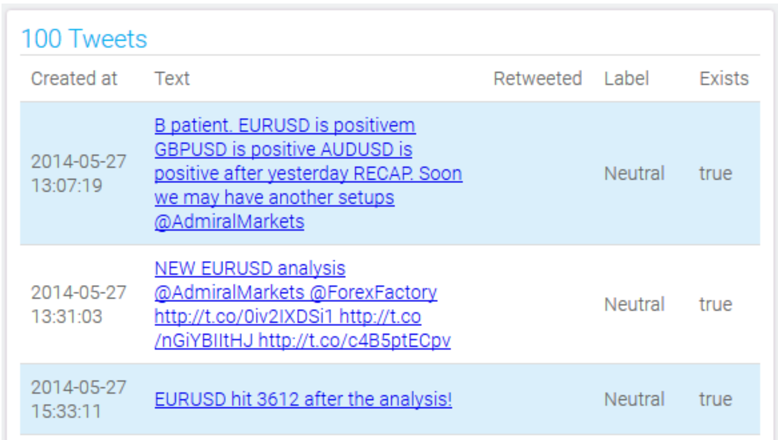

Figure 3.6: Tweets (max. 100) of the selected user.

3.3 Analysis

For the analysis we use *Python 3*. To retrieve and prepare the data from the database we use libraries called *psycopg2*, *pandas* and *re* (regular expressions). The *scipy* library is used to create the market model for the cumulative abnormal return. To plot the results of the analysis we use the *matplotlib* library.

For the purpose of the analysis, we use *Python* and some common methods of data science. Details of the analysis are explained in the next chapter.

Chapter 4

Analyses

In this section, we analyse Twitter users and their tweets. Based on the findings, we then analyse the relationship between the users and the currency movements. We end the chapter with analysis of relevant events.

4.1 Twitter Users

In order to analyse the tweets, we have to understand the behaviour of the Twitter users. Inspection of the tweets revealed sizeable content variety. We speculate that this occurs because of different types of users.

After an in-depth inspection, we identified four major user groups. According to our findings, we devised a simple classification model. The model automatically classifies users into groups with respect to their properties. The groups are analysed separately in section 4.5.

Although this model is very simple (see section 4.2), it gives promising results. In the future, the user groups are worthy of examination and further development.

4.1.1 User Properties

For the purpose of devising a classification model which classifies users into different groups (see section 4.2), we have to look into Twitter user properties in detail.

Beside unique user id, provided by Twitter, every user also has the following properties:

- The **number of tweets** (*tweets*) tweeted by the Twitter user. The distribution of users with respect to the number of tweets is shown in Figure 4.1.

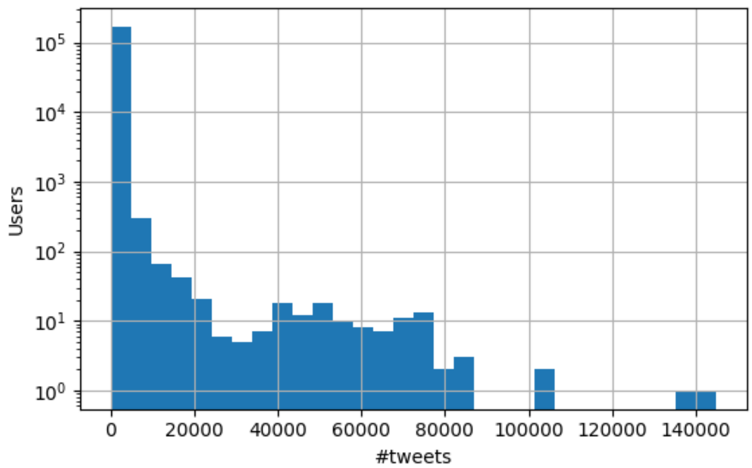

Figure 4.1: Distribution of users with respect to the number of tweets.

- **Days active** property is defined as following:

$$d_{active} = max(date_{user}) - min(date_{user}) + 1$$

  The property gives the time span of the users' tweeting. This property helps to calculate the average tweet and retweet rate. It does not reveal for how many days the user was actually active (it also includes days without tweeting).

- **Tweet rate**. Definition:

$$t_{rate} = \frac{tweets(user)}{d_{active}}$$

This property reveals how many tweets per day the user produced on average.

- **Retweets** ($retweets$), how many tweets of each separate user are retweets.
- **Retweets ratio**:

$$retweets_{ratio} = \frac{retweets}{tweets}$$

The ratio between *retweets* and *tweets* of the user.

- **Retweeted** ($retweeted$), how many times tweets of a particular user were retweeted. The number of retweeted tweets can be higher than the number of users' own tweets.
- **Retweeted ratio**:

$$retweeted_{ratio} = \frac{retweeted}{tweets}$$

The ratio between *retweeted* and *tweets* of the user.

- **Hirsch index** [21] describes the influence of the user and is defined as:

$$\text{h-index}(user) = max_i min(RT(i), i)$$

where RT is a function that returns the number of times the tweet i was retweeted, and i is the index of the tweet in the ascending sequence of tweets sorted by retweet count. In short, a Hirsch index of five means that a user has five tweets that were retweeted at least five times.

- **Existence**, whether a user is still active or not.
- **user type** is one of the following:

trading robot, relevant company, relevant individual, spam.

The classification of the user types is explained in section 4.2.

4.2 User Classification

By collected Twitter user properties (subsection 4.1.1) we can devise a model to classify users into five groups. Four of them are important for our analysis, and the last one contains the undefinable users. Not all the user properties are used, remaining properties can be used in further development of our model.

Tweets of the four user groups are later used to analyse Forex events. The goal is to establish the relation between Twitter sentiment of each group and Forex market movements caused by the events. The users are not limited to be members of a single group only. Groups Trading robots and Spam have more users in common than other groups.

4.2.1 Forex Trading Robots

A large proportion of all the tweets have a similar format. For example:

- Bought 0.08 Lots $EURUSD 1.39223 SL 1.38308 TP 1.39608
- Sold 0.08 Lots $EURUSD 1.38391 SL 1.39399 TP 1.37899
- Closed Sell 1.7 Lots EURUSD 1.29617 for +4.9 pips, total for today +114.5 pips #best #forex
- Closed Buy 1.2 Lots EURUSD 1.29602 for -5.8 pips, total for today +114.3 pips #trade #results #forex

It is obvious that such type of tweets belong to Forex trading robots. These robots should not be confused with the tweetbots [22, 23] which main purpose is to tweet periodically to advertise or otherwise deliver content to the Twitter sphere.

The Forex trading robots are programs that follow currency movements and trade with currencies. We speculate that this is a group of commercially available robots which advertise themselves by tweeting about their results.

It is likely that users of Forex robots get some sort of an incentive by allowing trading robots to tweet on their Twitter accounts about their trades.

We recognise a tweet as robotic ($t_{robotic}$) when it starts with phrases shown in Table 4.1. Such phrases are commonly used in the Forex trading. Rule to classify a user as a Forex trading robot is when more than 75% of its tweets are robotic:

- $t_{robotic_{rate}} > 0.75$

We manually checked 2000 random tweets classified as robotic but we found that only four of them were not created by robots. The accuracy of this classification is then 99.8%.

Starts with	**Count**	**Share [%]**
Bought	156,553	1.1
Sold	154,264	1.1
Closed Buy	4,770,414	32.5
Closed Sell	4,833,488	33.0
Buy Limit	1,537	0.0
Sell Limit	2,627	0.0
Buy Stop	2,760	0.0
Sell Stop	2,703	0.0
Total:	**9,924,346**	**67.6**

Table 4.1: Beginning phrases of twitter bots and corresponding number of tweets.

Out of all the studied users, 4,580 belong to this group. They represent around 2.7% of all our users. This is a small percentage compared to the volume of tweets they produce. But it is not a surprise, since we believe that these Twitter accounts are typically machines and not humans.

4.2.2 Relevant Forex Companies and Analysts

We want to identify a group of reliable and relevant Twitter users with at least moderately large activity, so one could in principle follow them in the future.

Rules to classify the users into this group are:

- $d_{active} > 30$
- $t_{rate} > 0.5$
- $retweeted_{ratio} > 0.25$

The model classified 195 users (0.12%) into this group. They produced 210,733 tweets, which is 1.44% of all the tweets in our analysis. Users in this group are mostly Forex brokers and their analysts as identified upon manual checking.

The content of gathered tweets of this group are predictions and reports of these users. Although there is a very small number of users in this group, the relation between the sentiment of their tweets and the Forex movement around events is significant, actually of the highest quality of all the groups.

4.2.3 Relevant Individuals

Relevant individuals are recreational Forex traders or retired forex analysts. They are trading but not professionally.

They tweet about their observations on unusual currency movement, references to technical analyses and also some forecasts regarding Forex. The performance of this group is similar to relevant companies but with lower influence and relevance.

Relevant individuals have the below stated properties, but should not fall into the group of relevant Forex companies and analysts:

- $d_{active} > 30$
- $retweeted_{ratio} > 0.05$

The model classified 6,660 (3.9%) users into this group. They produced 810,511 tweets (5.5%).

4.2.4 Spam/scam/advertisement

Spam/scam/advertisement. This group of users provide malicious or non-informative content related to Forex. They usually have a very large amount of tweets in a short period of time.

Many users of this group also belong to the trading robot group.

Spam/scam/advertisement:

- $tweets > 1000$
- $retweeted_{ratio} < 0.01$

The model classified 869 (0.5%) users into this group. They produced 10,266,836 tweets (69.9%), 1,597,782 (10.9%) of which do **not** belong to users categorised as Forex trading robots.

4.2.5 Other

This group covers users that were not classified in any of four major groups by our model. This group includes users with the following properties:

- They produce small amount of tweets.
- They mostly retweet other users.
- Unpopular users with tweets, that were rarely retweeted or not at all.
- Users that usually tweet about topics unrelated to Forex.

We found 157,823 users in this group producing only 2,215,224 tweets. This group contains a large amount of users that could belong to other groups but failed to be classified due to the simplicity of our model.

4.3 Tweets

In this section we analyse tweets and their properties. We analyse 14,679,466 tweets which were collected in the time span of three years by Sowa Labs.

4.3.1 Tweet Properties

Below are listed properties of tweets used in our analysis:

- **Twitter id** (id_{tweet}) is a unique identifier of every tweet, as provided by Twitter.
- **Created at** is the timestamp of tweet's creation.
- **Text** is the content of the tweet.
- **Retweet count** (*retweeted*) is the number of times the tweet was retweeted within our data set.
- **Retweet reference** ($retweet_{ref}$) is the id of the original tweet which was retweeted.
- **User id** is the id of the Twitter user, the author of the tweet.
- **Label** (*sentiment*) denotes the sentiment of the tweet. This property was not provided by Twitter but was annotated by Sowa Labs company as described in subsection 2.2.1.
- **Existence** (*exists*) denotes whether the tweet still existed in July 2017.

4.3.2 Sentiment Distribution

The distribution of the sentiment in our set of tweets is different compared to the previous state-of-the-art articles [10]. The proportion of positive and negative tweets is unusually high (Figure 4.2) in comparison to previous analyses [21, 24]. We notice that the distribution of the sentiment among groups varies considerably. This is a confirmation that separating the users into groups is necessary.

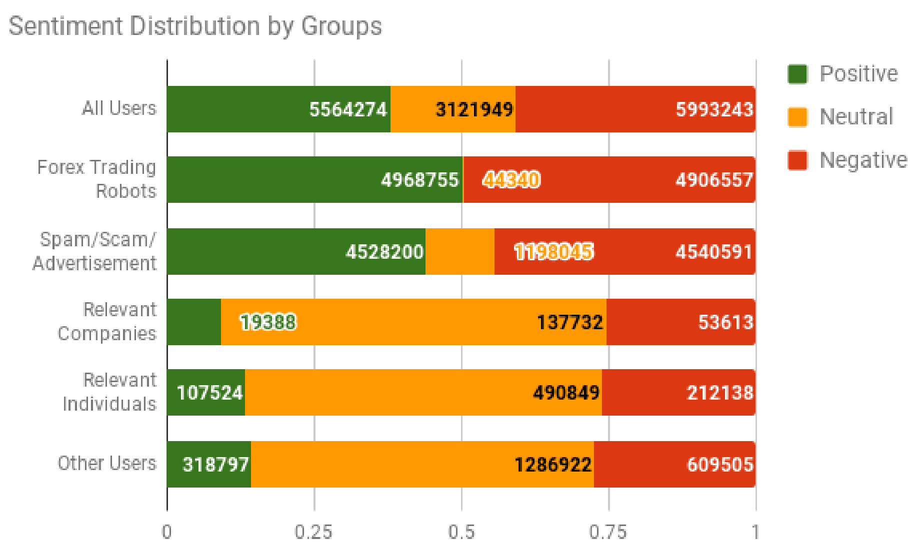

Figure 4.2: Relative sentiment distribution of user groups (bars are supplemented with tweets count).

The most unusual sentiment distribution of the tweets belonging to the group of Forex trading robots (see subsection 4.2.1). This set of tweets contains an almost negligible amount of tweets with a neutral sentiment. Upon further inspection, it turns out that this is caused by the specific vocabulary used in these tweets. As shown in Table 4.1, the vocabulary of these tweets is specific and leaves almost no option for sentiment to be neutral. For example:

- Tweets beginning with "Bought" or "Closed sell" express the expectation that the price will move in the favour of EUR, so the exchange rate EURUSD will rise.

- For tweets beginning with "Sold" or "Closed buy" the situation is the opposite: tweet expects the rate to drop.

Group of Spam/Scam/Advertisement users behaves similarly to the group of Forex trading robots. Obviously, the majority of the users belong to both groups.

On the other hand, the distribution of the groups Relevant companies, Relevant individuals and Other users match the sentiment distribution in previous studies.

Figure 4.2 shows that the volume of negative tweets in the last three groups is much higher than the volume of the positive ones. The reason is the predominant down trend of the EURUSD exchange rate during the period that our thesis covers.

4.3.3 Retweeting

We analyse two different aspects of the retweeting activity. One aspect is when the user retweets and the other is when the tweets of the user are retweeted. Out of all the analysed tweets, 247,237 tweets were retweeted and they were retweeted 492,648 times (on average 1.99 times per each retweeted tweet).

	Retweets		**Retweeted**	
Forex trading robots	707	(0%)	39,426	(0.4%)
Robots without spam	513	(0%)	20,101	(1.6%)
Relevant Companies	9,563	(4.5%)	205,059	(97.3%)
Relevant individuals	46,059	(5.7%)	206,488	(25.5%)
Spam/Advertisement	29,090	(0.3%)	25,776	(0.3%)
Spam without robots	28,896	(1.8%)	6,451	(0.4%)
Other users	429,873	(19.4%)	42,683	(1.9%)
Total	515,033	(3.5%)	492,648	(3.4%)

Table 4.2: Retweet count for the analysed groups.

The column "Retweets" in Table 4.2 shows that users of different groups behave differently. Forex trading robots do not retweet. Spam/Scam/Advertisement retweet rarely (they are either tweetbots [22, 23] or they advertise and do not bother to retweet).

The column "Retweeted" was used as an attribute in the user classification model, hence it should not be used as reference for determination of the quality of the model. It is interesting that the value distribution of columns "Retweeted" and "Retweets" coincide.

4.3.4 Daily Twitter Volume

Figure 4.3 shows average daily Twitter volume. There are global movements, where peaks coincide with opening times of largest forex sessions:

- peak around minute 100 of a day, approximately 1:30 (UTC), coincides with the opening time of the Tokyo session.
- peak around minute 400 of a day, approximately 7:30 (UTC), coincides with the opening time of the London session.
- peak around minute 800 of a day, approximately 13:00 (UTC), coincides with the opening time of the New York session.

We also notice peaks which are not part of a global Twitter volume movement. These peaks happen on regular intervals, mostly hourly, at the top of the hour. The peaks are especially noticeable for Spam users (Figure 4.4). Presumably these tweets belong to tweetbots which post periodically.

4.3.5 Twitter Volume Around Events

We are especially interested in tweets around the time of events. Increase of the Twitter volume (apart from regular 24 hour cycles) seen in Figures 4.5 and 4.6, around events tells us that Twitter users are following events, which confirms that it is worth focusing on them.

The increase of volume around the time of events is consistent with each group, the least pronounced change of volume of tweets occurs in the Spam group.

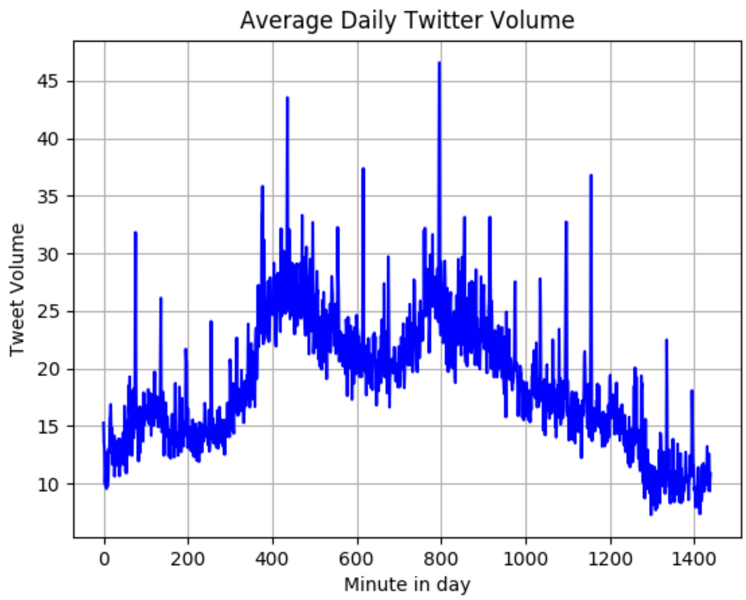

Figure 4.3: Average daily Twitter volume.

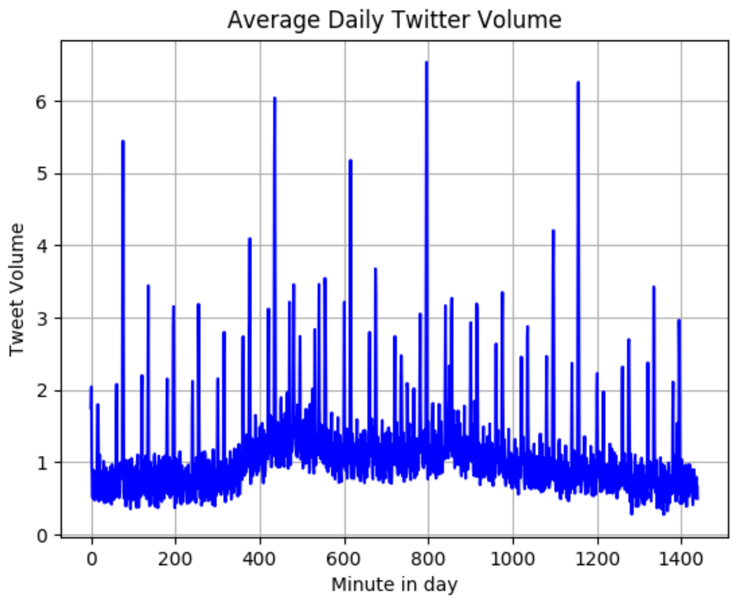

Figure 4.4: Average daily Twitter volume of spam users.

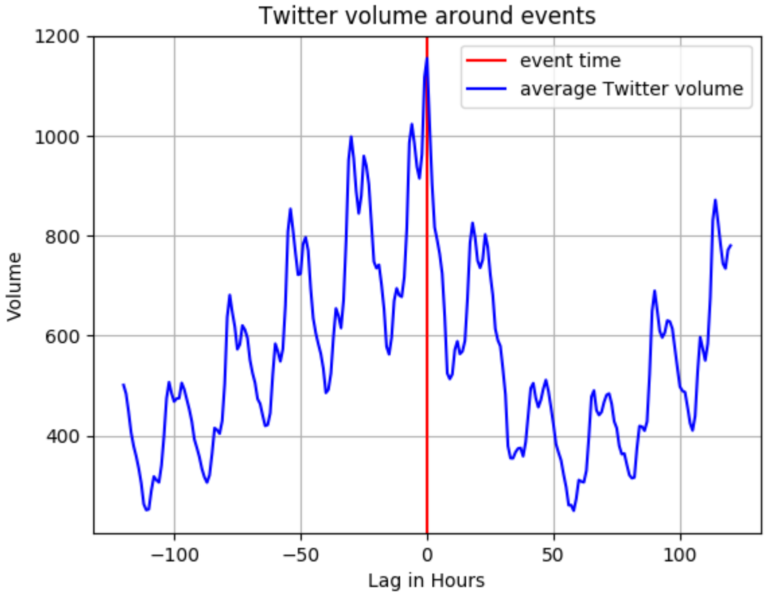

Figure 4.5: Average Twitter volume around the time of events of all users.

4.3.6 Deleting tweets (and Users)

There is an interesting phenomenon, where certain tweets and users are being deleted after a while. There are two main types of motives to delete tweets, harmless and fraudulent. We are interested in the fraudulent ones. We speculate that after events, certain users delete their tweets with the false predictions and keep the true predictions. This way, they want to improve their reputation.

The analysis shows that there is a very small amount of such users, so we did not bother to remove them from our analysis. The community typically recognises such users and ignores them (or even reports them).

4.4 Events

Due to the large number of events, we only consider the high impact events [19] that happen at least once per month. Other events would probably not

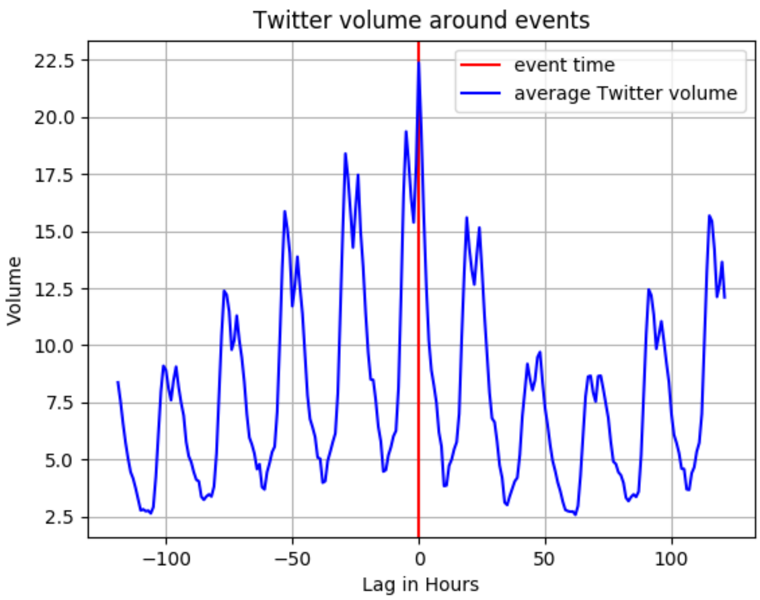

Figure 4.6: Average Twitter volume of relevant companies and analysts around the time of events.

add much to the value of the analysis. The list of the analysed events as listed on Forex Factory is shown in Table 4.3.

Every event has the following properties:

- *forecast* tells us the expectation of the analysts.
- *actual* tells us the actual state after the event.
- *previous* tells us the state of previous event of the same type, this property does not enter our analysis and is listed for the sake of completeness only.

Properties of different event types can have different formats. They can be expressed as percentage, numerical values (in thousands, millions, ...) or non-numerical values (e.g. talks by the ECB [25] chief Draghi, or FED chief Yellen). We omitted events with non-numerical values, although they can have a large impact on the market. They tend to be complex to analyse.

Event Name	**Related Currency**
Crude Oil Inventories	USD
Non-Farm Employment Change	USD
CB Consumer Confidence	USD
ISM Non-Manufacturing PMI	USD
ADP Unemployment Claims	USD
Core Retail Sales m/m	USD
Core CPI m/m	USD
Building Permits	USD
Unemployment Rate	USD
German Ifo Business Climate	EUR
Retail Sales m/m	USD
Prelim UoM Consumer Sentiment	USD
PPI m/m	USD
Philly Fed Manufacturing Index	USD
ISM Manufacturing PMI	USD

Table 4.3: Analysed event types and currencies related to them.

Due to different formats and meanings, every event has its own rules to be classified as positive/neutral/negative. In general, the interpretation of events' falls into two groups:

- Events "Unemployment rate" and "ADP Unemployment Claims" have a positive effect on the currency if $actual < forecast$ and vice versa.
- Other events in Table 4.3 have the opposite set of rules: the event has a positive effect on the currency if $actual > forecast$ and vice versa.

We also consider classifying a part of events as neutral. As usual in event studies [26] we take a 2.5% threshold:

$$\left| \frac{actual - forecast}{forecast} \right| < 0.025$$

4.5 Event Study

The core of our analysis is the event study. We compare the signal in the social media (Twitter sentiment) about market movements with the measures used by the professional analysts.

In our analysis we use the following measures:

- **Market return** is a quantity which is computed as the price of the asset changes during a given period of time. Market return is defined as follows:

$$r = \frac{p_1 - p_0}{p_0}$$

  where p_0 represents the price at the beginning of the interval (opening price) and p_1 represents the price at the end of the interval (closing price).

- Event study methodology analyses activities that deviate from the normal behaviour. For this purpose we have to calculate the return of abnormal activity (**abnormal return**) which usually happens after the event. This is done by subtracting the normal price movement (market model) from the abnormal movement.

- **Market model** is created by using the data of the market movement before the event. We use a simple linear regression of the market movement 30 days prior to the event. The slope coefficient k of the linear regression is then used for subtracting it from the market movement after the event.

  After obtaining the market model for the event, we subtract it from the price movement (i.e. on one minute resolution) after the event to get the abnormal price:

$$p_{a_i} = p_i - k * i$$

  where p_i is actual price at the time i after the event and p_{a_i} is abnormal price at the same time.

Once we have abnormal price movement, we can calculate abnormal return:

$$r_{a_i} = \frac{p_{a_{i+1}} - p_{a_i}}{p_{a_i}}$$

- **Cumulative abnormal return** (CAR) is a quantity usually used in the event studies. CAR measures aggregated returns over longer period of time, called the event window. We calculate CAR which computes the abnormal return from the event at the given time:

$$car_i = \sum_{j=0}^{n} r_{a_j}$$

Examples of CAR are shown in Figure 4.9.

- **Twitter sentiment score** measures how positive or negative were the tweets in a selected time frame. Its value is in the interval $[-1, 1]$.

$$sentiment\_score(T) = (+1)\frac{pos + 1}{n + k} + (0)\frac{neu + 1}{n + k} + (-1)\frac{neg + 1}{n + k}$$

where pos, neu and neg represent the number of positive, neutral and negative tweets in the time interval T, n represents the number of all the tweets in T ($n = pos + neu + neg$), k is a number of different types of sentiment (in our case three, positive, neutral and negative). This is a part of the Laplace correction (along with a "+1" in the numerator of the fractions) used to avoid zero and low volume problems.

The formula for our case:

$$sentiment\_score(T) = \frac{pos - neg}{n + 3}$$

4.5.1 Analyses

In this section we present the results of our event study. The analyses are straightforward because all the prepared tools are ready to use.

Sentiment Analysis

Our goal is to compare the sentiment of Twitter users and Forex Factory[1] analysts. The sentiment of the analysts is defined in section 4.4 as the difference between the forecast of the analysts and the actual event outcome.

The sentiment of Twitter users in our analysis is the sentiment score of tweets one hour after (nowcast) or before (forecast) the event. We have chosen the time span of one hour because smaller time interval would not contain sufficient amount of tweets, while longer time intervals would not contain specific enough information about the event.

Our main focus is nowcasting because the sentiment of the analysts is known only after the event so the comparison between analysts and Twitter users can be done with nowcasting only.

Categorising Events

We analyse the sentiment of the analysts and Twitter users for three categories of the events: positive, neutral and negative. We make several different analyses: events are categorised according to the types of entities (analysts or Twitter users), among different Twitter users, and whether we analyse tweets before or after the events.

When categorising the events according to the sentiment of the analysts we use the event classification method defined in section 4.4. For the categorisation of the events according to the Twitter users we set a threshold for the sentiment score so that one third of the events fall into each group. Sentiment score distribution and thresholds for users are shown in Figure 4.7.

Using the categorisation of the events and the sentiment of different entities, we get the result in form of CAR curves shown in Figure 4.8. The analyst curves can be separated around the events, but soon the effect of the event waters down. On the other hand, events categorised by the sentiment

[1] http://www.forexfactory.com

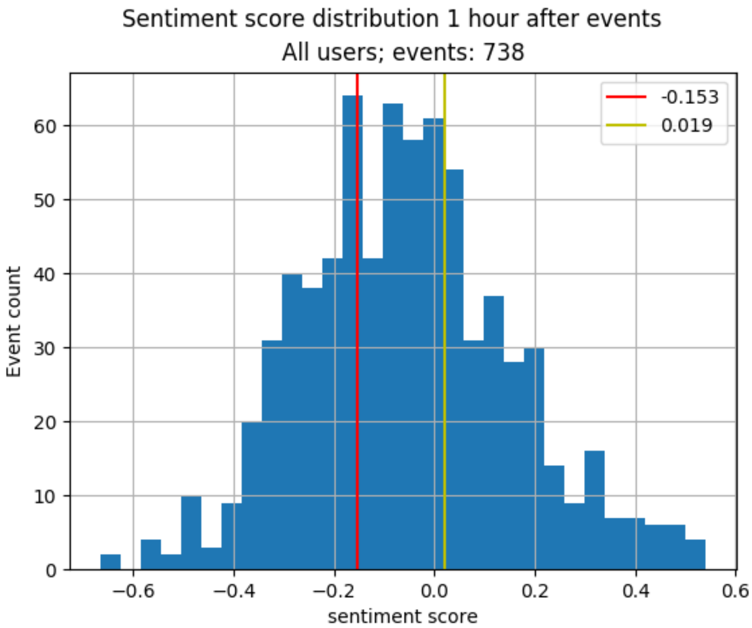

Figure 4.7: Sentiment distribution for tweets one hour after events.

of the tweets, display larger movements but in the wrong direction, events classified as positive give negative returns and vice versa.

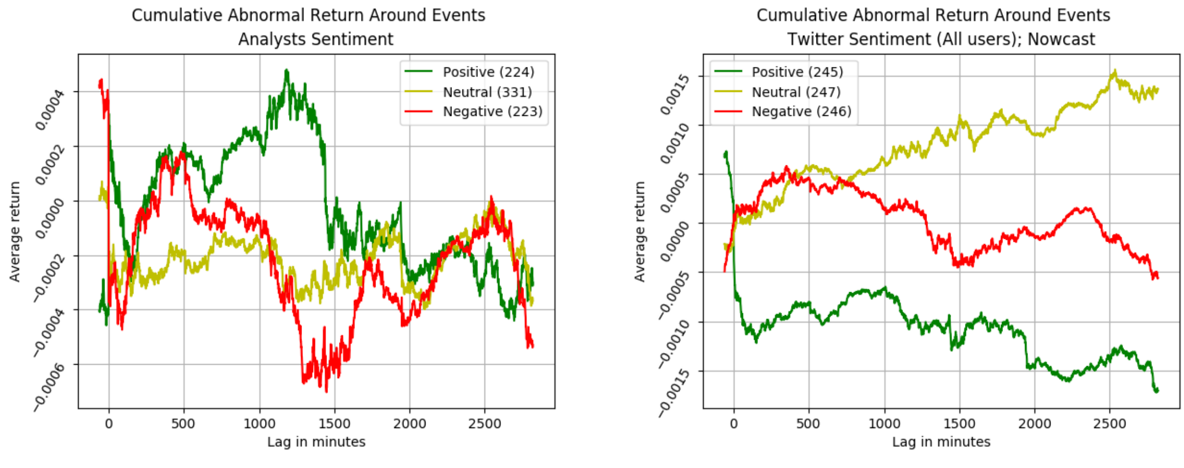

Figure 4.8: CAR for events according to the sentiment of the analysts (left) and Twitter users (right).

Twitter Groups

The analysis of different Twitter user groups demands the calculation of different thresholds for the distribution of sentiment score for each group separately. We do not use universal thresholds (obtained from all tweets) because every group of users has its own characteristics (as determined in Figure 4.2) thus should be treated differently.

The CAR curves for events of classification according to different groups of Twitter users are shown in Figure 4.9. Movements follow the correct direction when events are categorised by more relevant users, but when categorised with sentiment from less relevant users the movement is either chaotic (group Spam) or follow a wrong direction (group Trading robots). This results confirm correctness of our user classification model.

Nowcast vs. Forecast

We compare Twitter sentiment before and after the event. The results are shown in Figure 4.10. As we can see, no group shows promising results in forecasting. CAR curves categorised by trading robots show certain movements in the right direction, but not more than 500 minutes after the event, when the effects of the events no longer have any influence.

The forecast analysis cannot be done for the Forex Factory analysts because one can not calculate the sentiment prior to the event without the event outcome.

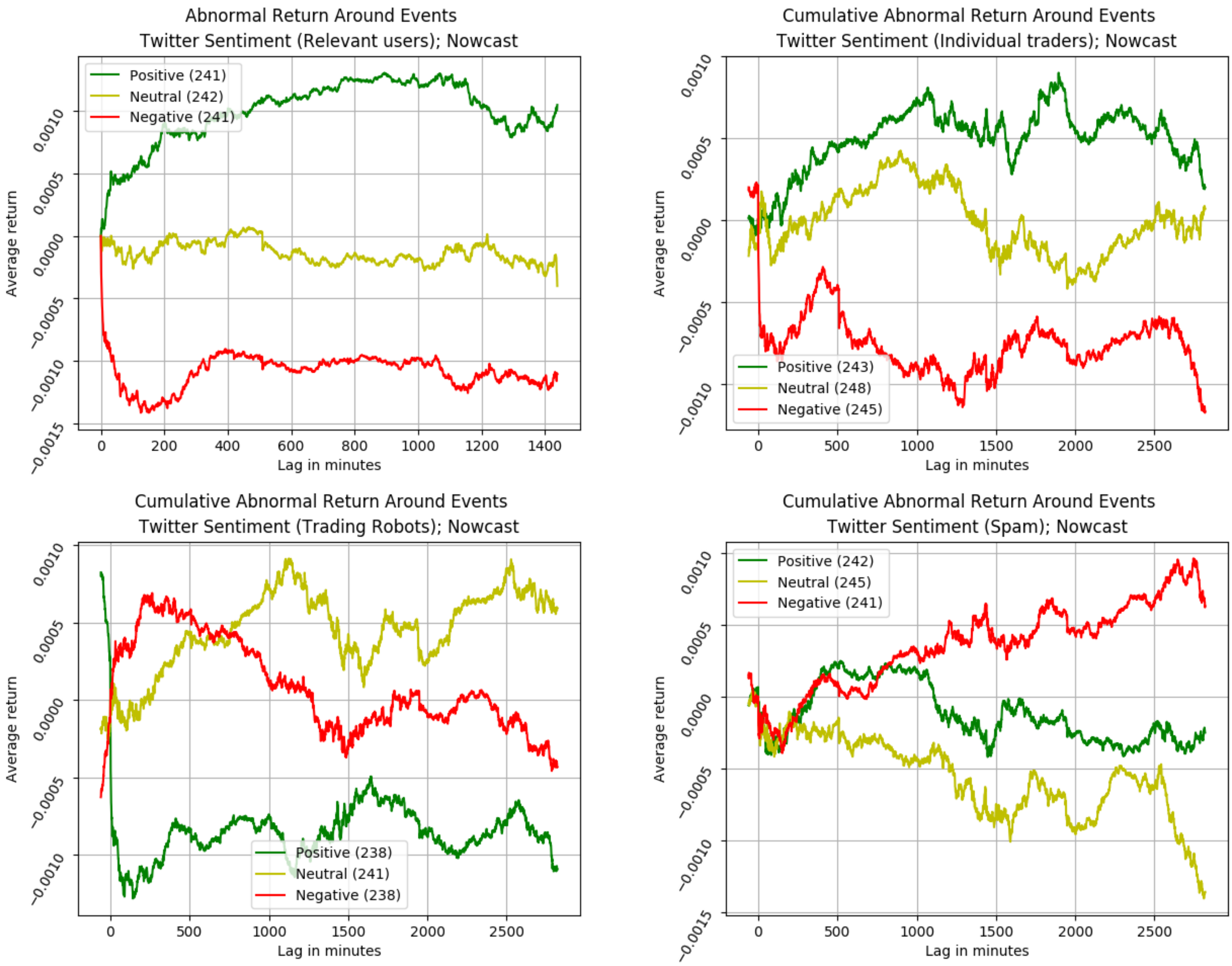

Figure 4.9: Cumulative abnormal return for all the user groups according to Twitter sentiment after the events.

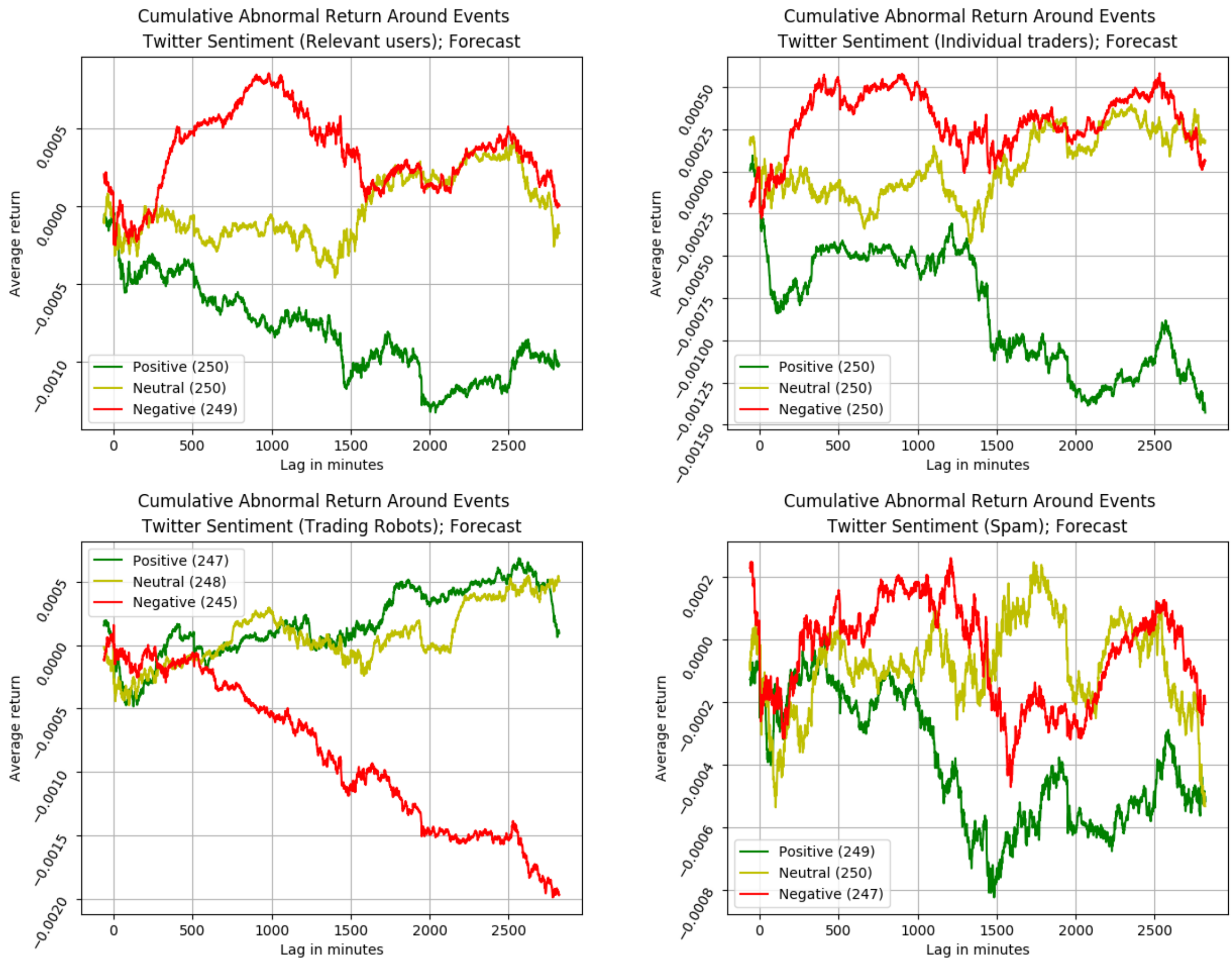

Figure 4.10: Cumulative abnormal return for all the user groups according to Twitter sentiment prior to the events.

Chapter 5

Conclusions

We analyse interactions between two complex systems, Forex and Twitter. Each system has its own characteristics. Forex is by far the largest financial market. There are several established methods for analysis of this system, such as technical analyses, mathematical modelling, statistical methods etc. Our focus is an application of event study methodology on Forex. Social media are another complex system, we focus on Twitter. We analyse the tweets with the content related to EUR and USD [27].

We used three years of data. The currency pair (EUR/USD) movements data were retrieved from histdata.com web page. The events data were retrieved from the Forex Factory web page. The tweets were provided by Sowa Labs company, which manually annotated a subset of tweets.

We annotated all collected tweets with the machine learning model trained on the manually annotated subset. We used a multi-class SVM algorithm. We validated the classifier using 10-fold cross validation and obtained surprisingly high classification accuracy of the model (81.4%). We found that the reason for this was a high share of robotic tweets, which have limited vocabulary and thus it is easy to determine their sentiment. Comparing to other published analyses, we used by far the largest set of labelled tweets for training the classifier.

We created a web application for visualisation and analysis of the data.

Front-end was created with state-of-the-art technologies. *HTML5*, *CSS3*, *Javascript*, *Angular* and *Highcharts* were used for the the design and dynamics of the front-end. Back-end was implemented in *Python*, which was used to connect the front-end and the database. Web application displays Forex movements, high impact events and their details, Twitter sentiment, as well as properties of the Twitter users and their tweets.

We analysed Twitter users and defined groups of users with distinct behaviour based on their activity, impact, vocabulary etc. According to these properties we named four major user groups: Forex trading robots, Relevant companies, Relevant individuals and Spam. We built a user classification model based on the properties of the groups. We consider this model to be an interesting achievement. In future work, the model can be improved by defining more user groups, using additional user properties and extending the model to other fields.

We analysed the tweets to confirm the correctness of the user model. We analysed the properties of the tweets for each user group, such as retweeting statistics, daily volume movement, volume movement around the time of events, etc. The analysis shows clear differences among the user groups. We noticed that groups of relevant companies and individuals are highly retweeted (thus relevant).

The events concerning EUR and USD are diverse so we focused on the events of high impact and frequency. The majority of analysed events are connected to USD rather than to EUR. We speculate that the reason is dispersion of announcements in Europe while the announcements in the USA are centralised. The results of the event analysis show different polarisations of the events with respect to the different user groups and analysts. In our analysis, we did not use speeches of presidents of major central banks (EU and USD) in spite of their high impact on currencies. The reason is that they cannot be simply numerically assessed. These types of events are left for further analyses.

The event study shows the impact of the events on financial markets and social media. By using the sentiment of different user groups in time close to the events, we calculated their polarity. According to the polarity of the events we computed an average cumulative abnormal returns (CAR). Detected differences in CAR curves between user groups confirm our user classification model.

5.1 Contributions

In the thesis we devised a classification model for classifying Twitter users. Model separates the users into distinct groups: users relevant to Forex prediction, irrelevant users and spam users. This classification enables us to follow or avoid certain groups of users. The model can be adapted to various other fields.

Our analysis shows the relation between Forex market movements and groups of users in connection with certain events.

We developed a web application that enables browsing through the currency movement, Twitter and events data.

Bibliography

[1] A. A. Baasher, M. W. Fakhr, Forex trend classification using machine learning techniques, in: Proceedings of the 11th WSEAS international conference on Applied computer science, 2011, pp. 41–47.

[2] J. Yao, C. L. Tan, A case study on using neural networks to perform technical forecasting of Forex, Neurocomputing 34 (1) (2000) 79–98.

[3] I. Kaastra, M. Boyd, Designing a neural network for forecasting financial and economic time series, Neurocomputing 10 (3) (1996) 215–236.

[4] J. Kamruzzaman, R. A. Sarker, I. Ahmad, SVM based models for predicting foreign currency exchange rates, in: Third IEEE International Conference on Data Mining, 2003. ICDM 2003., IEEE, 2003, pp. 557–560.

[5] S.-C. Huang, P.-J. Chuang, C.-F. Wu, H.-J. Lai, Chaos-based support vector regressions for exchange rate forecasting, Expert Systems with Applications 37 (12) (2010) 8590–8598.

[6] P. Domingos, Bayesian averaging of classifiers and the overfitting problem, in: International Conference on Machine Learning, 2000, pp. 223–230.

[7] R. Fatum, M. Hutchison, Effectiveness of official daily foreign exchange market intervention operations in japan, Journal of International Money and Finance 25 (2) (2006) 199–219.

[8] U. A. Müller, M. M. Dacorogna, R. B. Olsen, O. V. Pictet, M. Schwarz, C. Morgenegg, Statistical study of foreign exchange rates, empirical evidence of a price change scaling law, and intraday analysis, Journal of Banking & Finance 14 (6) (1990) 1189–1208.

[9] A. Vincent, M. Armstrong, Predicting break-points in trading strategies with Twitter, Available at SSRN: https://ssrn.com/abstract=1685150.

[10] P. Gabrovšek, D. Aleksovski, I. Mozetič, M. Grčar, Twitter sentiment around the earnings announcement events, arXiv preprint arXiv:1611.02090.

[11] A. Brvar, Elektronsko trgovanje na valutnem trgu s pomočjo twitterja, Ph.D. thesis, Univerza v Ljubljani (2015).

[12] A. Khadjeh Nassirtoussi, S. Aghabozorgi, T. Ying Wah, D. C. L. Ngo, Text mining of news-headlines for Forex market prediction, Expert Systems with Applications 42 (1) (2015) 306–324.

[13] M. Makrehchi, S. Shah, W. Liao, Stock prediction using event-based sentiment analysis, in: Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 2013 IEEE/WIC/ACM International Joint Conferences on, Vol. 1, IEEE, 2013, pp. 337–342.

[14] T. Rao, S. Srivastava, Modeling movements in oil, gold, forex and market indices using search volume index and Twitter sentiments, in: Proceedings of the 5th Annual ACM Web Science Conference, ACM, 2013, pp. 336–345.

[15] F. Jin, N. Self, P. Saraf, P. Butler, W. Wang, N. Ramakrishnan, Forex-foreteller: Currency trend modeling using news articles, in: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2013, pp. 1470–1473.

[16] S. S. Ozturk, K. Ciftci, et al., A sentiment analysis of Twitter content as a predictor of exchange rate movements, Review of Economic Analysis 6 (2) (2014) 132–140.

[17] G. T. Panger, Emotion in social media, Ph.D. thesis, University of California, Berkeley (2017).

[18] B. Born, M. Ehrmann, M. Fratzscher, Central bank communication on financial stability, The Economic Journal 124 (577) (2014) 701–734.

[19] L. A. Smales, N. Apergis, Does more complex language in fomc decisions impact financial markets?

[20] P. D. Azar, A. W. Lo, The wisdom of Twitter crowds: Predicting stock market reactions to FOMC meetings via Twitter feeds, The Journal of Portfolio Management 42 (5) (2016) 123–134.

[21] M. Grčar, D. Čerepnalkoski, I. Mozetič, The Hirsch index for Twitter: Influential proponents and opponents of Brexit, in: Book of abstracts, Institut Jožef Stefan, 2016.

[22] H.-S. Kim, W.-Y. Shin, D. Kim, J. Cho, Improved tweet bot detection using spatio-temporal information, Journal of the Korea Institute of Information and Communication Engineering 19 (12) (2015) 2885–2891.

[23] A. Duh, M. S. Rupnik, D. Korošak, Collective behaviour of social bots is encoded in their temporal Twitter activity, arXiv preprint arXiv:1706.00077.

[24] I. Mozetič, M. Grčar, J. Smailović, Multilingual Twitter sentiment classification: The role of human annotators, PloS ONE 11 (5) (2016) e0155036.

[25] M. Picault, T. Renault, Words are not all created equal: a new measure of ECB communication, Available at SSRN: https://ssrn.com/abstract=2980777.

[26] A. C. MacKinlay, Event studies in economics and finance, Journal of economic literature 35 (1) (1997) 13–39.

[27] V. Gholampour, E. van Wincoop, What can we learn from Euro-Dollar Tweets?, Tech. rep., National Bureau of Economic Research (2017).