UNIVERSITY OF CALIFORNIA,
IRVINE

Bayesian Nonparametric Models on Big Data

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Economics

by

Fulya Ozcan

Dissertation Committee:
Professor Dale J. Poirier, Chair
Professor Ivan G. Jeliazkov
Professor Eric T. Swanson

2017

DEDICATION

To my loving parents Ulku and Asim,
whose love and support made me who I am
To my sister Meltem,
who always pushes me to strive for higher
And to Merlin,
who brightens up my days with his existence.

Contents

List of Figures

Page

List of Tables

Page

List of Algorithms

ACKNOWLEDGMENTS

I grew up with my father's bed-time stories featuring many creatures from the phoenix to the anteater. I had the impression that anteaters were mythological and belonged in the stories. UC Irvine has taught me anteaters are not mythological but rather magical. I am indebted to UC Irvine for the countless opportunities and magical memories in these 5 amazing years.

This thesis would not have been possible without the support, advice and wisdom of countless brilliant individuals for whom I am full of gratitude. First and foremost, I would like to express my sincere gratitude to my advisor Professor Dale Poirier for his continuous support and encouragement through the course of my PhD journey. It is his open-mindedness and insightfulness that made this dissertation possible, and without him my PhD experience would not have been the same. He has been a real inspiration to me with his enthusiasm, energy and his constant pursuit of leading edge research and it is truly an honor to be his student. I would like to thank him for always challenging me and being a constant support as I pushed the limits of my abilities to strive for better. I feel lucky to have had the opportunity to work alongside him and his teachings will guide me in my future aspirations. I am also eternally grateful for Professor Ivan Jeliazkov for always being there for me and supporting me. He has not only been a brilliant mentor who inspired me to work harder but has also been a close friend who gives wonderful advice and whose wit and thought-provoking insights make each and every conversation with him memorable. It has been a blessing having him by my side during this journey, and I would like to thank him for everything. I am also very thankful to Professor Eric Swanson his for insightful advice and remarks on my papers. His comments have helped me greatly improve my work. I would like to also acknowledge Professor David Brownstone for making this dissertation possible by letting me use his node at the high performance computing cluster.

I feel lucky to be a part of such a supportive cohort. I cannot thank Dr. Michael Guggisberg enough for talking me into getting my Stats masters and making stats struggles fun. He has not only been a great influence, but has also added so much laughter and color to my life. I would also like to thank Tyler Boston and Tim Duffy for making my graduate school experience so enjoyable. I am very grateful to Tyler also for trusting me with his kitties while he's in Tanzania with the Peace Corps and making me the mom of the two coolest cats in the whole world! I would like to thank my hermanita Karina Hermawan for being the best travel mate, the best chef and being the creator of so many happy memories I know I will always cherish. Finally, I would also thank Kara Dimitruk for her vibrant, uplifting presence and for all our quality time together - nothing better than Doctor Who to put the grad school stress away.

None of my accomplishments in life, including this dissertation, would be possible if it were not for my parents. They are the source of my strength and creativity as well as the gentle yet firm push forward allowing me to persevere through all of my endeavors. I will always be astonished and humbled by how they simultaneously completed their PhDs, served as full time judges, and took care of toddler Fulya on top of it all, and I will always look up to them for their infinite kindness, diligence, humility and wisdom. They are the heroes that guide and inspire me to become not only a better academic but also a better, kinder human being and there is no better feeling than knowing that they are and will always be there for me. I am also thankful to them for blessing me with the most amazing sister one could have. She never fails to fascinate me with her intellect, compassion, creativity and wit and she's the living proof of how a little sister can outsmart the elder one. She's always an inspiration for me with everything she does, and I feel lucky to be her proud sister. I am also thankful to my husband for all the memories, for always believing in me and for helping me become who I am today. He has been an endless support that made my life much more fun and enjoyable, and I am beyond lucky to have him by my side.

I would like to thank Duygu Akdevelioglu for being my guardian angel, for always watching over me and always supporting me through thick and thin, for feeling my pain as her own when I was sick and sharing my joy when I'm happy. I am also very lucky to have Kumru Arslan as my honorary sister. I would like to thank her for always looking out for me, for pampering me while I was working on this dissertation and for cheering me up like no other. I am also grateful to Ecehan Uludag for all her support and kindness and to Dr. Caner Guclu for doing utmost for his friends. Finally, I cannot thank Dr. Cemil Can Coskun enough for always being there, his great sense of humor, and for every little thing he has done for me over these years. I am very lucky to have such an incredible, thoughtful and supportive friend group, and I will always cherish the memories of my Irvine days.

I am also very grateful to my managers and teammates at Amazon for their flexibility, encouragement and constant support while I completed the last chapter of this dissertation.

My research would not be possible without the funding and the resources provided by Department of Economics, School of Social Sciences, University of California, Irvine. My experience at UC Irvine has not only allowed me to accumulate a vast fund of knowledge and learn from such great professors, but has also shaped my career in an incredible way.

Zot zot zot!

CURRICULUM VITAE

Fulya Ozcan

EDUCATION

Doctor of Philosophy in Economics	**2017**
University of California	*Irvine, CA*
Master of Science in Statistics	**2016**
University of California	*Irvine, CA*
Master of Arts in Economics	**2014**
University of California	*Irvine, CA*
Master of Arts in Economics	**2012**
Bilkent University	*Ankara, Turkey*
Bachelor of Arts in Economics	**2010**
Bilkent University	*Ankara, Turkey*

EXPERIENCE

Research Scientist	**2017–**
Amazon	*Seattle, Washington*
Data Scientist	**2016**
Airbnb	*San Francisco, California*
Lecturer	**2015**
University of California, Irvine	*Irvine, California*
Econometrician	**2014**
Knightsbridge Asset Management	*Newport Beach, California*
Co-founder	**2013–2016**
BeatPool	*Irvine, California*
Teaching Assistant	**2012-2017**
University of California, Irvine	*Irvine, California*
Teaching Assistant	**2010-2012**
Bilkent University	*Ankara, Turkey*

HONORS & AWARDS

David Brownstone Best Paper in Econometrics Award **2016**
UC Irvine

Summer Research Fellowship **2015, 2016**
UC Irvine

Dean's Fellowship **2013**
UC Irvine

Bilkent University Fellowship **2010–2012**
Bilkent University, Turkey

Science Scholarship **2010–2012**
The Scientific and Technological Research Council of Turkey

Bilkent University Fellowship **2005–2010**
Bilkent University, Turkey

Science Scholarship **2005–2010**
The Scientific and Technological Research Council of Turkey

ABSTRACT OF THE DISSERTATION

Bayesian Nonparametric Models on Big Data

By

Fulya Ozcan

Doctor of Philosophy in Economics

University of California, Irvine, 2017

Professor Dale J. Poirier, Chair

This thesis focuses on the role investor type and sentiment play in financial markets, using data from social media. First paper investigates the effect of the interaction between asset maturity and liquidity restrictions in "on-the-run" phenomenon, using asset markets with search frictions. Under the presence of search frictions, investors would not prefer holding assets with very short time to maturity, since they will need to go back to the market and search for a counterpart to buy new assets every time their assets mature, incurring a search cost. However, they also would not want to hold assets with very long time to maturity due to liquidity considerations. An asset search model is set up to determine asset choices of investors with different liquidity preferences. Model considers two assets that differ in their maturities and two investor types who differ in their liquidity preferences. Main finding of this paper is that liquidity cost matters in the presence of search frictions as the model predicts a separating equilibrium where high type agents choose the long term asset and the low type agents choose the short term asset. When the two assets have the same time-to-maturities, the same separating equilibrium is obtained. Spread of the long term asset is found to be higher than that of the short term asset, which goes in line with the data and hence this paper shows that "on-the-run" phenomenon can be explained by higher search frictions in the off-the-run markets and investors with different liquidity preferences.

Second paper predicts intra-day foreign exchange rates by making use of trending topics from Twitter, using a sentiment based topic clustering algorithm. Twitter trending topics data provide a good source of high frequency information, which would improve the short-term or intra-day exchange rate predictions. This project uses an online dataset, where trending topics in the world are fetched from Twitter every ten minutes since July 2013. First, using a sentiment lexicon, the trending topics are assigned a sentiment (negative, positive, or uncertain), and then using a continuous Dirichlet process mixture model, the trending topics are clustered regardless of whether they are explicitly related to the currency under consideration. This unique approach enables to capture the general sentiment among users, which implicitly affects the currencies. Finally, the exchange rates are estimated using a linear model which includes the topic based sentiment series and the lagged values of the currencies, and a VAR model on the topic based sentiment time series. The main variables of interest are Euro/USD, GBP/USD, Swiss Franc/USD and Japanese Yen/USD exchange rates. The linear model with the sentiments from the topics and the lagged values of the currencies is found to perform better than the benchmark AR(1) model. Incorporating sentiments from tweets also resulted in a better prediction of currency values after unexpected events.

Third paper investigates the behavior of Reddit's news subreddit users and the relationship between their sentiment on exchange rates. Using graphical models and natural language processing, hidden online communities among Reddit users are discovered. The data used in this project are a mixture of text and categorical data from a news website. It includes the titles of the news pages, as well as a few user characteristics, in addition to users' comments. This dataset is an excellent resource to study user reaction to news since their comments are directly linked to the webpage contents. The model considered in this paper is a hierarhical mixture model which is a generative model that detects overlapping networks using the sentiment from the user generated content. The advantage of this model is that the communities (or groups) are assumed to follow a Chinese restaurant process, and therefore

it can automatically detect and cluster the communities. The hidden variables and the hyperparameters for this model can be obtained using Gibbs sampling.

Keywords: Natural Language Processing, Dirichlet Mixture Models, Exchange Rate Prediction, Hiearchical Mixture Models, Hidden Network Discovery, Overlapping Communities, Mixture Communities.

Chapter 1

Introduction

Over the past few years, developments in data collection and accumulation have enabled access to massive datasets in various disciplines, varying from finance to political science to healthcare. Definition of "big data" itself has also evolved, where the 3 V's of big data (volume, velocity and variety) are now replaced by the 7 V's of big data (volume, velocity, variety, variability, veracity, visualization and value). As data analysis techniques and tools also advanced, many different types of data became available at very fast rates, making it possible to get insight on even personal data. These high frequency and highly diverse data have led to many innovations in various disciplines. One of the most groundbreaking strand of literature that arose with the availability of these personal big data is the use of social media data in social sciences and business.

Social media data have been very popular in finance, particularly for modeling the movements in the stock market, using data from various means, varying from news headlines to Google Trends. Applications of techniques to economics, on the other hand, have been comparatively limited. The studies that focus on foreign exchange markets mostly use news headlines or news data to capture the market reaction to daily events, using text data. Incorporating this information into models enables medium and short-term exchange rate predictions; whereas by using only macroeconomic variables as features, only long-term prediction is available. Although the studies that incorporate news data into exchange rate prediction show improvements over models that only make use of macroeconomic variables, they miss out on events that occur with higher frequency.

This thesis shows that capturing those high frequency variables from social media data and using them as a proxy for general sentiment provides great value to short term exchange rate prediction. Social media has the capability to capture ongoing events and how people react to those events at every instance, providing real time feedback with very high frequency. This makes it possible to do predictions for a very short time frame, hence allowing acting as an early signal of where the currency will move given what is going on in the world and how people react.

Unlike collecting and accumulating these high frequency personal data, analyzing them to make use of them correctly is a challenging task. Earlier studies were mostly based on extracting the sentiment from the news headlines or tweets that are related to a company, and basing the prediction of the direction of its stock price on whether the sentiment is negative or positive. Although these provided improvement in prediction, they capture only a small percentage of what these data have to offer. In order to fully utilize the potential from these high frequency personalized data, this thesis resorts to hierarchical Bayesian models.

The second chapter presents a theoretical asset search model where investors experience liquidity shocks which affect their preferences. Assets differ in their maturities and as investors receive shocks, the proportion of the investor types in the population change, which in turn affects the prices of the assets. This chapter builds a theoretical foundation for asset valuation in which the state the investors are in affects the asset prices.

The following chapters present methodologies which move this theoretical approach to a real world environment. Instead of an ad-hoc approach to modeling the preference shocks that investors receive, or the states that the investors are in; second and third chapters treat these shocks or states as latent variables. By considering sentiment as observations resulting from these shocks, these chapters make inference on these latent states and use these latent states in improving short term prediction.

The third chapter uses data from Twitter's trending topics in the world, between 2013 and 2015 per 10 minute intervals to capture the general sentiment in the world. Then using a hierarchical Dirichlet process mixture model, sentiment clusters are created to capture the latent states, which improves the per 10 minute prediction of GBP/USD, EUR/USD, CHF/USD, and JPY/USD. There are some shortcomings of this approach due to limitations of this specific dataset, which are addressed in more detail in chapter 2 and improved in the following chapter.

Chapter 4 makes use of a more informative data set from Reddit's worldnews subreddit, which includes submissions on world news and comments on these submissions from 2013 to 2015. There are two main advantages of Reddit data set over the Twitter's trending topics data. Firstly it allows to cluster on news topics, which helps to differentiate the effects of different events. Secondly, it allows to detect networks of users depending on what they comment on (news clusters), how they comment (sentiment), and how their commenting behaviors change over time (evolving networks) and over topics. A four-layered hierarchical mixture model is used to combine information from each layer of the data set, resulting in an improvement in prediction compared to previous chapter.

Chapter 2

Time-to-Maturity and Liquidity Under Search Frictions

2.1 Introduction

This chapter is concerned with the effect of the interaction between asset maturity and liquidity restrictions in "on-the-run" phenomenon, using asset markets with search frictions. An asset search model is set up to determine asset choices of investors with different liquidity preferences.

Assets traded in the financial markets bear three types of risk: market, credit and liquidity. Market risk is the risk inherent in the market, beyond the control of the issuer and investor. Credit risk depends on the fundamentals of the asset itself, on the default probabilities of the asset. Liquidity risk, on the other hand, depends on the current market conditions, as well as investors' abilities to trade the assets. It is defined as the inability of investors of buying or selling the assets as quickly as desired, which is reflected in the bid and ask spreads. [1]

[1] There are many ways to define and calculate the liquidity risk. Garbade and Silber (1979) for instance,

Liquidity is considered to be one of the most important factors that the investors consider when purchasing the asset. In particular, in the markets where there is no immediate trade between counterparties, liquidity becomes an issue.

Investors who value liquidity more, i.e., liquidity constrained investors, would prefer holding assets that are more liquid, holding everything else constant. For instance, when there are liquidity crises people become value liquid assets more. During these times, the demand for liquid assets increase (Zhao (2013)), as well as the bid-ask spreads among assets that only differ in terms of their liquidity (Cooper (1999)); and investors care more about the liquidity of an asset rather than its quality, i.e., liquidity risk becomes more important than the credit risk (Beber et al. (2009)). Hence for a thorough investment choice analysis, liquidity considerations, especially time and time-to-maturity varying liquidity premium should be taken into account.

Presence of uncertainties in many aspects drives investors away from buying long maturity assets as they might need liquidity at any point in time. However, investors holding short term assets incur a search cost every time they try to replace their expired assets in the markets with trading frictions. This is especially the case for the over-the-counter (OTC) markets, which refer to the marketplace that has no physical location but two counterparts come together to trade the assets they hold. Prices in these markets are determined by the two counterparts. Market makers of the OTC markets are the dealers, instead of the original issuers, and investors trade with the dealers. These markets have trading frictions. These frictions arise from two main characteristics of the OTC markets. First, it takes time to find a counterpart dealer to trade the assets. Second, after a counterpart dealer is matched, usually a bargaining takes place between the two counterparts. This trade-off between hold-

define liquidity risk as the variance of the difference between the equilibrium price of an asset at the time a market participant decides to trade and the realized transaction price.

ing long and short term assets raises an important asset choice question: "How would the investors with different liquidity preferences decide on assets with different maturities under trading frictions?".

One of the main contributions of this chapter over the literature is that in this chapter there are two assets with the same dividend payments which only differ in their time to maturities, and which carry only the liquidity risk. This is to mimic the environment under the "on-the-run" phenomenon. Also, there are heterogeneous investors depending on their liquidity preferences. Hence this chapter looks at the role of the liquidity risk under search frictions when investors with different liquidity preferences decide on which asset to hold. The model is structured with search frictions with two assets which differ in their maturities, with the same time-to-maturities, where heterogeneous agents are subject to a liquidity preference shock to account for the "on-the-run" phenomenon.

This chapter is organized as follows. Related studies are discussed in Section-2. A general asset search model is presented in Section-3. The model for the "on-the-run" phenomenon is given in Section-4. Section-5 presents the results of this model. Section-6 concludes.

2.2 Related Literature

Valuation of an asset, along with many other factors, depends on its liquidity. Since illiquidity is costly to the investors, illiquidity of an asset is reflected in its price, in the form of illiquidity premium. Acharya and Pedersen (2005), through a liquidity-adjusted capital asset pricing model, find that return of an asset is expected to be higher as the asset gets more illiquid, implying an illiquidity premium. Amihud and Mendelson (1986) study the relationship between bid and ask spreads and the returns through a model with investors differing in their holding horizons. They find a “clientele effect” where investors with longer holding horizons hold stocks with higher spreads, to compensate for the illiquidity. The extent that liquidity of an asset affects its price depends on the type of the asset, i.e., whether it is a risky or a risk-free asset. Credit risk is reflected in the spread of an asset along with the liquidity risk. Although there are many studies that try to separate liquidity risk from the default risk, there is also evidence on a two-way interaction between the two. Liquidity is shown to be affecting the investment decisions indirectly as well as directly, through this two-way interaction with credit risk. Chen et al. (2014) document by a cross-sectional study that corporate bonds tend to be more liquid for bonds with higher credit ratings.

The interaction between liquidity and credit risk is studied through various theoretical models as well. He and Milbradt (2014) show, with a search theoretical model in OTC markets, where bargaining with dealers determines a bond’s endogenous liquidity, which depends on firm fundamentals and time to maturity; this in turn again affects credit risk. Huang and Huang (2012) show that credit risk is a function of maturity of the assets, and Geromichalos et al. (2016) explain positive term premia (investors holding long term bonds receive higher returns compared to investors holding short term bonds) with a search model. Through a general equilibrium model where investors are subject to endowment risk, Praz (2014) compares two correlated assets which differ in terms of the liquidity in the venues they are being traded (a liquid market and an illiquid OTC market). The endowment risk in the economy creates an endogenous liquidity risk, which increases the risk premium of both assets. Author finds the two markets interacting in two ways. The liquid market decreases the overall search frictions, where price discount on the illiquid asset falls, and the bargaining price in the OTC market increases; but liquid asset also captures some portion of the value of the illiquid asset due to risk-sharing. Rocheteau (2009) investigates this interaction on real assets and finds the illiquidity premium on real assets to be increasing in the riskiness and the abundance of the asset in a competitive and decentralized market setting (CM-DM).

Through another search model, Lagos (2010) studies the asset trade both in Walrasian markets where trading occurs instantly and in secondary markets where bargaining takes place between buyers and sellers. Author finds that assets are valued for their liquidity and also for their internal characteristics. In light of all these studies and more, in order to assess the pure effect of liquidity of an asset on its valuation, the two effects should be separated. Since treasury securities are almost risk-free, their yield spread can be considered to be reflecting the liquidity premium. For corporate bonds, which are riskier than treasury securities, the yield difference over the risk-free securities reflect both the default and the liquidity risk.[2] Therefore treasury securities provide a suitable environment to analyze the effect of liquidity of an asset on its valuation.

The relationship between liquidity and time to maturity, in addition to the maturity date of the asset, is crucial in understanding the investor behavior. Time-to-maturity is the time remaining until the maturity date stated on the security. Importance of this relationship in investment decisions is well documented in various empirical studies. Chakravarty and Sarkar (1999) for instance show that the liquidity premium is a function of time to maturity in municipal and corporate bond markets and Driessen (2005) estimates that liquidity premia accounts for about 20 % of the spread. However, in order to investigate the pure relationship between maturity and liquidity, the effect of credit risk on yield should be separated. This analysis can be done by looking at risk-free assets with different maturities. A good example would be comparing treasury bills, treasury notes and treasury bonds. Although all three assets are considered to be risk-free, they differ in terms of their maturities and hence in terms of their liquidity premia, which results in different yield spreads for them.[3] Treasury

[2] Although not as easy as in the case for risk-free securities, it is possible to separate liquidity risk from credit risk to analyze the liquidity risk premium of risky assets. Amihud et al. (2006), acknowledging the corporate bonds being less liquid than the treasury securities, discuss various studies that tries to estimate the liquidity premium of corporate bonds by separating it from the default premium. For instance, De Jong and Driessen (2012) estimate the liquidity premium on long term US corporate bonds to be 0.45%.

[3] These securities further differ from each other in terms of their tax treatments, which in the end affects the yield. Moreover, treasury bonds can be lent in repo markets, which earns the owner a lending fee, which

bills have maturities less than 1 year and pay no interest before maturity, whereas treasury notes and bonds have longer maturities and they make interest payments. Treasury notes have maturities ranging from 2 to 10 years, and they make interest payments semi-annually. Treasury bonds have the longest maturities among all, with at least 10 years, and they also pay interest twice a year.[4] Through a dynamic asset pricing model in a search setting, Weill (2008) investigates the role of liquidity in bargaining markets, controlling for risk premia; and shows different expected returns to be explained by differences in liquidity, caused by variation in the tradable shares of the assets.

Another way to compare the yield of these risk-free securities is to look at securities that have the same time-to-maturity. Then, ignoring the different tax treatments, these securities only differ in terms of being "on-the-run" and "off-the-run". Treasury securities that are most recently issued are referred to as on-the-run, whereas all previously issued treasury securities are called off-the-run. Since on-the-run securities are recently issued, they are more frequently traded and hence more liquid compared to the off-the-run securities. This difference in their liquidity is reflected in their yield, and hence in their price, where on-the-run securities are traded at higher prices (and give lower yield) and off-the-run securities are traded at lower prices (and give higher yield), implying an illiquidity premium on the off-the-run securities. When comparing treasury securities with different maturities but with same time-to-maturity, short maturity securities in consideration will be on-the-run whereas longer-maturity securities with same time-to-maturity will be off-the-run. This is because the short term securities are recently issued whereas longer term securities with the same time-to-maturity have been previously issued and are already being held in some investor's portfolio. Comparing treasury bills and notes with a time-to-maturity of less than six months (where treasury bills are on-the-run and notes are off-the-run), Amihud and

in turn again changes the yield.

[4] Treasury bonds are now issued with 30-years maturity only.

Mendelson (1991) find the average note yield to be 0.43% higher than bill yield, which shows the existence of an illiquidity premium for treasury notes over bills. Warga (1992) finds the yield difference between treasury bonds and notes to be 0.55%. As an extension of the general model, this chapter will investigate the on-the-run phenomenon through a search model.

In addition to securities with different life-cycles differing in their liquidities and hence valuations, liquidity of an asset changes also through its own life-cycle, which also affects its price. Sarig and Warga (1989) show bond liquidity to be inversely related to its age. To examine this phenomenon, Díaz and Escribano (2012) estimates the current liquidity and expected future liquidity of US Treasury bonds using a "liquidity life-cycle" function on GovPx dataset and find expected future liquidity affecting the bond prices more than does the current liquidity. Hence investors also take into account for how long assets will be locked in their portfolios. This expected liquidity considerations would also have a "clientele effect". Among different group of investors with different holding horizons (i.e., investors with different liquidity needs), investors with long holding horizons are expected to earn liquidity premia by investing in illiquid assets Lagos and Rocheteau (2009).

One of the biggest sources of illiquidity in asset markets are transaction costs. Hence, how the relationship between time-to-maturity and liquidity affects the asset choice also depends on how fast and how expensive it is to find a counterpart to purchase that asset, which depends on the microstructure of the market in which the asset is being traded. Amihud et al. (2006) list exogenous transaction costs, inventory risk, private information and search frictions as important components of the microstructure of the markets which affect the liquidity of the assets.Kamara (1994), uses the turnover ratios of different treasury securities (bills to notes) as a proxy for "expected length of transaction time in the note market relative to the bill market", and finds that as the turnover ratio increases transition time in note markets are expected to increase with respect to those in bill markets. Turnover of an asset in secondary markets depends on the number of dealers as well as the inventories that dealers hold. Garbade and Silber (1979) show that as the dealer participation increases, markets clear more efficiently, reducing the liquidity risk. Using their "liquidity risk" definition, Mendelson (1982) finds the optimal time between clearings to be decreasing with the price volatility; since as the number of dealers falls, every time markets clear, more orders will be executed, which results in less fluctuations in the execution price. Investigating the effect of inventory risk of the dealers and search costs in OTC markets, Jankowitsch et al. (2011) document the traded prices being different than the assets value in US corporate bond markets. This implies that there is a liquidity effect in OTC markets due to the fact that finding a counterpart in these markets requires time because of the search frictions. These studies hence show that frictions affect the bid and ask spreads through the liquidity risk.

This chapter falls into the strand of literature where liquidity is investigated through search theoretical models with trading frictions. The model presented here is close to the model by Duffie et al. (2005). In that paper, authors seek to answer how intermediation and asset prices are affected in OTC markets. There are two types of investors depending their on holding cost, which causes them to have different liquidity preferences. Authors find the bid-ask spread to be lower when it gets easier to find a counterpart in the OTC markets. Incorporating risk aversion into this framework, Duffie et al. (2007) show that as investors become more risk averse, the spread will increase. The two important characteristics of these search models in OTC markets are firstly heterogeneity of the agents in terms of their liquidity preferences, and secondly the market characteristics, ie., how fast it is to be matched with a dealer and the dealer's bargaining power. In most papers where investors differ in terms of their liquidity requirements or holding horizons, there is a "clientele effect" where investors with liquidity constraints hold short-term assets. Lagos and Rocheteau (2009) analyze how trading frictions affect the distribution of asset holdings and measures of liquidity through a search model where heterogeneous agents accomodate trading frictions by adjusting their asset positions. Vayanos and Wang (2007) examine why assets with similar cash flows differ in their liquidity and price and find that there is a clientele eqilibrium where all short horizon investors prefer the asset that has lower search times, but higher price compared to the other asset with identical pay-offs. Extending this to a spot and repo market setting, Vayanos and Weill (2008) show that short-sellers would prefer more liquid assets even though assets offer identical cash flows. Gârleanu (2009) finds that when portfolio choice is unrestricted, optimal positions depend on liquidity in a search setting. Geromichalos et al. (2016) use a CM-DM model where short term assets mature in time to take advantage of random consumption opportunities in markets with imperfect credit. They consider two types of buyers, ones that consume the good and other that do not consume during the DM market. They find positive liquidity premium arising due to agents being subject to stochastic consumption expenditures, markets being decentralized which creates a need for a liquid asset as a medium of exchange, and due to search frictions. Studies that concentrate on market characteristics investigate the role of bargaining power of the dealers as well as the market structure.Lagos and Rocheteau (2007) find that when dealers have a lower market power, trading costs will fall, which increases the trading volume. This will in the end lead to more dealers entering the market, which makes it easier to find a counterpart, further reducing the trading costs. Lagos et al. (2011) the implications of the two main frictions in the OTC markets, which are finding a counterpart dealer and bargaining with that dealer. Authors find that when the dealers' bargaining power is high, there will not be enough liquidity in the market, which calls for a government purchase of the assets in the markets to provide liquidity and improve welfare. Pagnotta (2013) analyzes how market structure affects the relationship between liquidity and asset prices and finds that in consolidated markets as the trading becomes faster, liquidity and prices increase whereas in fragmented markets this relationship between liquidity and prices might be opposite. These studies show that liquidity affects the equilibrium valuation of the assets through two channels: investor channel where liquidity constrained investors prefer short-term assets, and through market channel where trading frictions affect the bid-ask spread.

2.3 General Model

The model considered here is close to the one in Duffie et al. (2005). There are three kinds of agents considered in the model: investors, issuers and marketmakers (dealers). Marketmakers are not the issuers of the assets, but they act as middlemen. Marketmakers have an instant access to inter-dealer market and do not hold any asset inventories.[5] Issuers issue the assets in the primary market, and dealers trade these assets in the secondary market. Investors can buy the assets from either venue but can sell these assets only in the secondary markets. The assets are nondivisible and an investor can hold only 1 unit of an asset, or may not hold any (asset choice is in $\{0, 1\}$). All type of agents are infinitely lived, risk neutral and they discount future at rate r.

As in Duffie et al. (2005), there are two types of investors depending on their liquidity preferences: H and L. H-type investors do not face a holding cost for the asset, whereas L-type investors do face a holding cost (denoted by δ).[6] H-type investors receive a preference shock with Poisson rate γ_h and become L-type, and L-type investors receive a preference shock with Poisson rate γ_l and become H-type.

[5] The reason they do not hold any inventories in this model is to avoid the complications that the "inventory effect" would add to the interpretation of the yield spread. Kamara (1994), by comparing treasury bills and notes shows that when the dealers' inventories of the less liquid asset (treasury notes) increases, the yield spread differential between treasury bills and notes decreases. Amihud and Mendelson (1980) also document the dealer prices to be affected by the inventory effect. Since the sole interest of this chapter is on liquidity, any other factors that can possibly contribute to the yield spread differential should be omitted from the model.

[6] Duffie et al. (2005) list "... need for cash, high financing costs, hedging reasons to sell, relative tax disadvantage, lower personal use of the asset" as various reasons for the existence of the low type investors.

There are two types of assets that are subject to different maturity shocks: a short-term asset with maturity shock following a Poisson process with rate ω_s, and a long-term asset with maturity shock following a Poisson process with rate ω_l. Assets do not have a face value but they pay 1 unit of consumption as dividends. Once an asset is hit by a maturity shock, it no longer pays a dividend and hence maturity is equivalent to full depreciation. After an investor's asset is matured, he can go back to the asset markets to replace the matured asset. Assets are traded in 4 different venues as given in Figure-2.1.

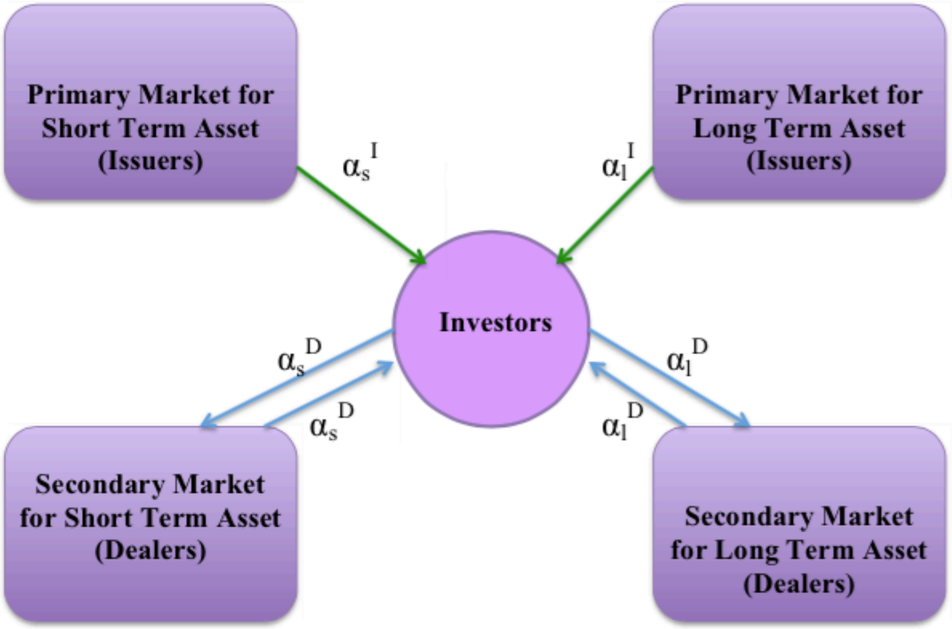

Primary Market for Short Term Asset (Issuers)
Primary Market for Long Term Asset (Issuers)
α_s^I
α_l^I
Investors
α_s^D
α_s^D
α_l^D
α_l^D
Secondary Market for Short Term Asset (Dealers)
Secondary Market for Long Term Asset (Dealers)

Figure 2.1: Asset Markets

There are two primary markets, one for the long term asset and one for the short term asset, where the issuers issue and sell their assets. There are also two secondary markets for each type of asset where dealers act as middlemen and sell and buy the assets. Prices of the assets are fixed at the primary markets whereas investors and dealers bargain over the transaction fee in the secondary markets. Long term assets are traded only in the primary and secondary markets for long term assets; and short term assets are traded only in the primary and secondary markets for short term assets. Investors who want to buy assets can search for an issuer in the primary markets or a dealer in the secondary markets whereas investors who want to sell assets can do so only in the secondary markets. Trade in all these four markets requires a match between the counterparts.

An investor seeking to purchase a short term asset can get matched with an issuer in the primary market following a Poisson process with rate α_s^I or with a dealer in the secondary market following a Poisson process with rate α_s^D, whichever happens the first. Similarly, a match between an investor seeking to purchase a long term asset and an issuer happens with a Poisson process with rate α_l^I or with a dealer in the secondary market with rate α_l^D. Although investors can purchase the type of asset they are looking for from either the primary or the secondary market, they can sell their assets back only to the dealers in the secondary market, not back to the issuers. Hence, a match between an investor trying to sell his long term asset and a dealer in the secondary market happens with a Poisson rate α_l^D, and a match between an investor trying to sell his short term asset and a dealer in the secondary market happens with a Poisson rate α_s^D. A match with a dealer occurs match with an issuer of the same type of asset, ie., $\alpha_s^D > \alpha_s^I$ and $\alpha_l^D > \alpha_l^I$. Hence the transaction fee in the secondary markets can be considered a convenience fee, since the foregone dividend payments increase as the investor keeps searching for an asset.

Since there are two type of investors and two type of assets, there are 4 equilibria where each type of investors are holding an asset:[7]

i. Pooling equilibrium at the short-term asset.

[7] It is also possible to have equilibria where one type or both types of agents hold no assets, which is not in the scope of this chapter.

ii. Separating equilibrium where high type chooses long-term asset and low type chooses short-term asset.

iii. Separating equilibrium where high type chooses short-term asset and low type chooses long-term asset.

iv. Pooling equilibrium at the long-term asset.

The remainder of the chapter focuses on the separating equilibrium in which high type investors choose to hold long term asset and sell short term asset, and low type investors choose to hold short term asset and sell long term asset. The conditions for this equilibria to hold are discussed in the following subsections.

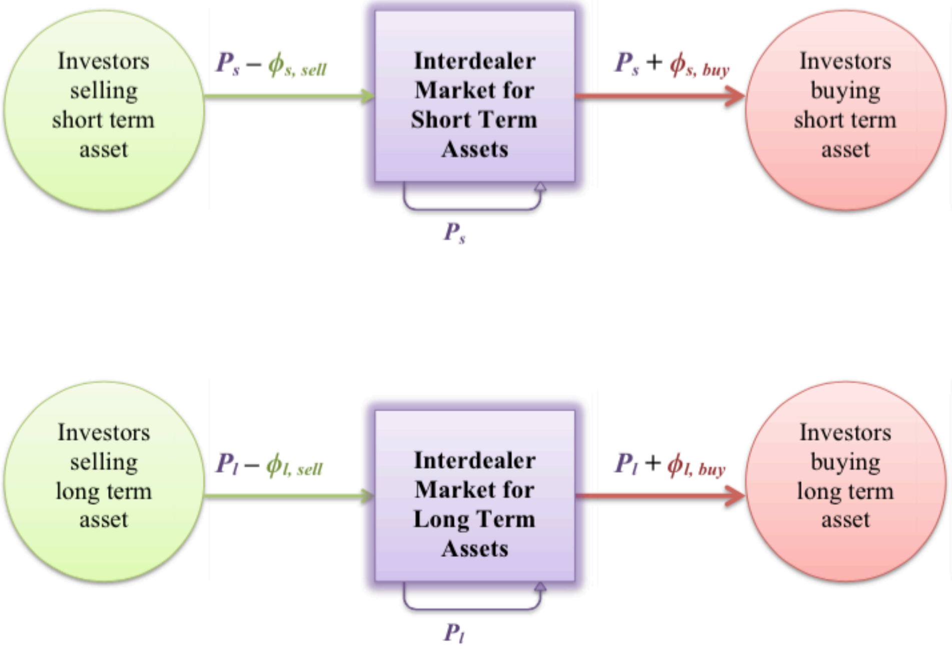

Investors selling short term asset
$P_s - \phi_{s,\ sell}$
Interdealer Market for Short Term Assets
$P_s + \phi_{s,\ buy}$
Investors buying short term asset
P_s
Investors selling long term asset
$P_l - \phi_{l,\ sell}$
Interdealer Market for Long Term Assets
$P_l + \phi_{l,\ buy}$
Investors buying long term asset
P_l

Figure 2.2: Interdealer Markets

2.3.1 Interdealer Markets

As mentioned earlier, dealers do not hold any inventories but they have an instant access to the interdealer market. Interdealer markets are depicted in Figure-2.2. When a dealer is matched with an investor who is seeking to purchase an asset, he buys the asset the investor is looking for at the interdealer price from the interdealer market for that type of asset. If instead the investor is trying to sell his asset to the dealer, the dealer buys the asset from the investor and sells it at the interdealer market at the interdealer price for that asset. An investor sells his asset of type j to the dealer at the interdealer price less the transaction fee $(P_j - \phi_{j,sell})$ where the transaction fee is determined as the result of the bargaining with the dealer and the investor. Similarly, an investor buy the asset of type j from the dealer at the interdealer price plus the transaction fee $(P_j + \phi_{j,buy})$ where the transaction fee is the outcome of the bargaining between dealer and the investor. Since investors on the selling side and the buying side are not the same, transaction fees are not necessarily equal. The bid and ask spread for asset type j is given by $\phi_{j,sell} + \phi_{j,buy}$.

Since dealers do not hold any inventories, the supply of the assets in the interdealer market comes from the investors selling their assets. Thus, the amount of each type of asset that dealers can sell is constrained by the amount of assets being sold by the investors. Similarly, the amount of each type of asset that dealers can buy is constrained by the amount of assets being demanded by the investors. This will be described in further detail in the following subsection.

2.3.2 Distribution of Investors

There are two types of investors depending on their liquidity preferences: high and low. Depending on whether an investor is holding an asset or not, and which type of asset he

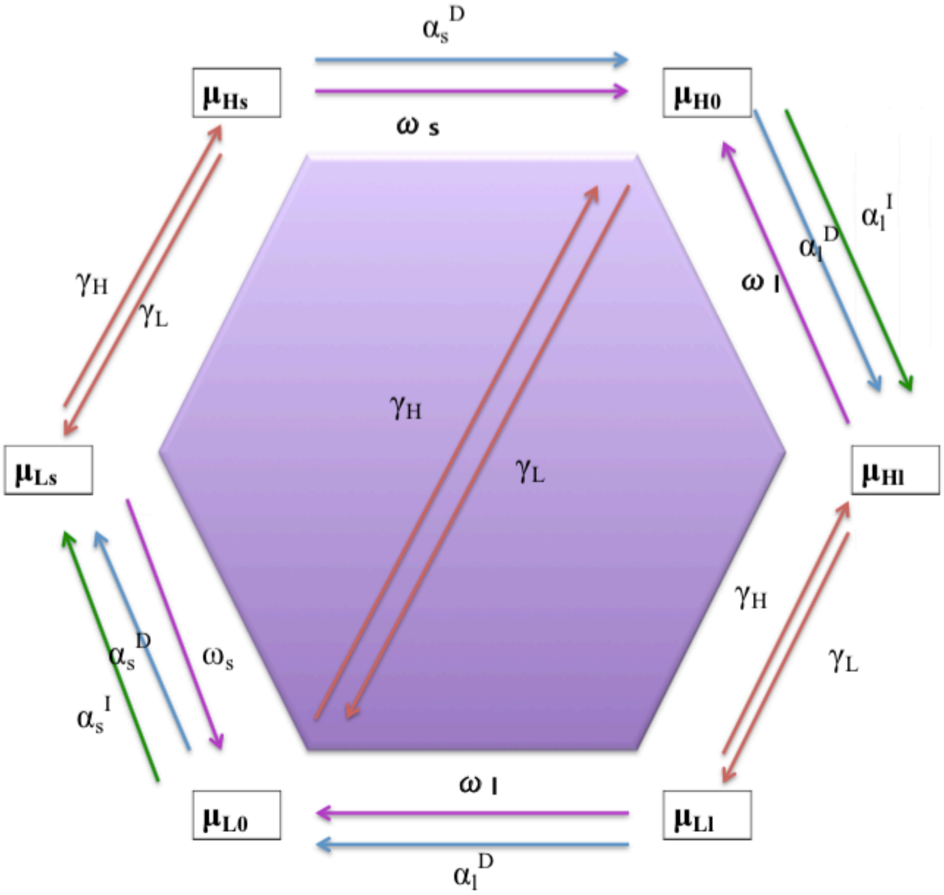


Figure 2.3: Cycle of Investment Behavior

is holding, there are six different states for an investor to be. Total number of investors is normalized to 1. Fraction of investors in each state is denoted by μ_{Jk} where subscript J indicates whether an investor is a high type or a low type ($J \in \{H, L\}$) and subscript k indicates whether an investor is not holding an asset or holding a short term asset or holding a long term asset ($k \in \{0, s, l\}$). Figure-2.3 shows how investors move between these states for the separating equilibrium where high type agents hold the long term asset and sell the short term asset, and the low type agents hold the short term asset and sell the long term asset.

If an investor is not holding any assets (μ_{H0}, μ_{L0}), there are 3 ways that he can leave this state. The investor can either get hit by a liquidity shock (with Poisson rate γ_L if the investor is high type or with Poisson rate γ_H if the investor is low type), or the investor can purchase an asset and leave the state. Purchase of an asset is possible by searching for and eventually getting matched with an investor (with Poisson rate α_l^I if the investor is seeking to purchase a long term asset or α_s^I if the investor is seeking to purchase a short term asset) or with a dealer (with Poisson rate α_l^D if the investor is seeking to purchase a long term asset or α_s^D if the investor is seeking to purchase a short term asset).

There are also 3 ways that an investor can arrive at this 0 state. An investor of the opposite type holding no asset can get hit by a preference shock and become the other type holding no asset. The asset of the investor of the same type holding an asset might mature (with Poisson rate ω_l if holding a long term asset or with rate ω_s if holding a short term asset) and as a result the investor enters the 0 state. Finally, an investor of the same type holding an asset might decide to sell his asset, and after getting matched with a dealer (with Poisson rate α_l^D if the investor is selling a long term asset or α_s^D if the investor is selling a short term asset), the investor enters the 0 state. The change in the fraction of the investors holding 0 assets are given by the following two equations:

$$\dot{\mu_{H0}} = -\gamma_L \mu_{H0} + \gamma_H \mu_{L0} + \omega_s \mu_{Hs} + \omega_l \mu_{Hl} + \alpha_s^D \min\{\mu_{Hs}, \mu_{L0}\}$$

$$-\alpha_l^D \min\{\mu_{H0}, \mu_{Ll}\} - \alpha_l^I \mu_{L0} \qquad (2.1)$$

$$\dot{\mu_{L0}} = -\gamma_H \mu_{L0} + \gamma_L \mu_{H0} + \omega_s \mu_{Ls} + \omega_l \mu_{Ll} - \alpha_s^D \min\{\mu_{Hs}, \mu_{L0}\}$$

$$+\alpha_l^D \min\{\mu_{H0}, \mu_{Ll}\} - \alpha_s^I \mu_{L0} \qquad (2.2)$$

Notice in these two equations that there is a minimum operator that constrains the amount of assets traded with the dealers, due to lack of inventories. In Equation-1, the number of high type investors with a short term asset becoming high type investors with no assets is constrained by the demand from the buyers' side, which are the low type investors with no assets. Dealers will buy the short term assets from the high type investors up to the amount that meets the demand from the low type investors with no assets, which is given by $\min\{\mu_{Hs}, \mu_{L0}\}$ and prices will adjust accordingly. Likewise, the number of high type investors with no assets becoming high type investors with long term assets is constrained by the supply from the sellers' side, which are the low type investors with long term assets. Dealers will sell the long term assets to the high type investors up to the amount that equals the supply from the low type investors selling their long term assets, which is given by $\min\{\mu_{H0}, \mu_{Ll}\}$ and prices will adjust accordingly.

The fractions of the investors holding the assets they prefer (short term assets for low type and long term assets for high type) are given as follows. Investors can enter the state Jk (where $Jk \in \{Hl, Ls\}$) in three ways. Investors of the opposite type holding the asset k can get hit by a preference shock an become the other type of investors holding the asset k. Moreover, investors can get matched with a counterpart to purchase their preferred asset. This can happen with Poisson rate α_k^I if the match is happened with an investor or with α_k^D if the match is happened with a dealer. Investors can leave their preferred state only in two ways: if they get hit by a preference shock and their type changes, or if their asset matures and they enter the 0 holding state. Since they are holding their preferred assets, there is no selling of the asset back to the dealer. The change in the fraction of the investors holding their preferred assets are given by the following two equations:

$$\dot{\mu_{Hl}} = -\gamma_L \mu_{Hl} - \omega_l \mu_{Hl} + \gamma_H \mu_{Ll} + \alpha_l^D \min\{\mu_{H0}, \mu_{Ll}\} + \alpha_l^I \mu_{H0} \tag{2.3}$$

$$\dot{\mu_{Ls}} = -\gamma_H \mu_{Ls} - \omega_s \mu_{Ls} + \gamma_L \mu_{Hs} + \alpha_s^D \min\{\mu_{Hs}, \mu_{L0}\} + \alpha_s^I \mu_{L0} \tag{2.4}$$

The amount of preferred assets that can be purchased from the dealers by the investors are again given by the minimum operators. Long term assets can be purchased by the high type investors holding no assets from the dealers up to the amount supplied by the low type investors holding long term assets, given by $\min\{\mu_{H0}, \mu_{Ll}\}$; and short term assets can be purchased by the low type investors holding no assets from the dealers up to the amount supplied by the high type investors holding short term assets, given by $\min\{\mu_{Hs}, \mu_{L0}\}$. Prices adjust accordingly.

The fractions of the investors holding the assets they do not prefer (long term assets for low type and short term assets for high type) are given as follows. Investors can enter the state Jk (where $Jk \in \{Hs, Ll\}$) in one way only, through the preference shock. Investors of the opposite type holding the asset k can get hit by a preference shock an become the other type of investors holding the asset k. This is the only way where an investor can end up holding an asset that he does not prefer. There are however three ways that investors can leave this unwanted state: if they get hit by a preference shock and their type changes, or if their asset matures and they enter the 0 holding state, or if they sell the unwanted asset to a dealer. The change in the fraction of the investors holding their unwanted assets are given by the following two equations:

$$\dot{\mu_{Hs}} = -\gamma_L \mu_{Hs} - \omega_s \mu_{Hs} + \gamma_H \mu_{Ls} - \alpha_s^D \min\{\mu_{Hs}, \mu_{L0}\} \tag{2.5}$$

$$\dot{\mu_{Ll}} = -\gamma_H \mu_{Ll} - \omega_l \mu_{Ll} + \gamma_L \mu_{Hl} - \alpha_l^D \min\{\mu_{H0}, \mu_{Ll}\} \tag{2.6}$$

The amount of unwanted assets that can be sold to the dealers by the investors are again given by the minimum operators. Long term assets can be sold by the low type investors to the dealers up to the amount demanded by the high type investors with no asset holdings, given by $\min\{\mu_{H0}, \mu_{Ll}\}$; and short term assets can be sold by the high type investors to the dealers up to the amount demanded by the low type investors with no asset holdings, given by $\min\{\mu_{Hs}, \mu_{L0}\}$. Prices adjust accordingly.

The steady state distribution of investors will be given by equating the left hand side of the equations 1-6 to zero.

2.3.3 Pricing of the Assets

There are in total 8 different prices: 2 in the primary markets and 6 in the secondary markets. The prices is fixed for each asset in the primary markets whereas there is bargaining in the secondary markets. Hence there is only one price in a primary market for each type of assets whereas there are three prices in the secondary market for each type of assets: ask price, bid price and the interdealer price.

Prices in the Primary Markets

In the primary markets, new assets are issued to replaced the matured ones. Investors can buy directly from issuers, but cannot sell back to the issuer. A match with a dealer occurs faster than a match with an issuer. Issuers provide the assets at the net present value, given by the Bellman equation:

$$rNPV = 1 + \omega(0 - NPV) + \gamma\alpha^D a(0 - NPV)$$

Net present value of each type of asset can then be expressed in terms of the interest rate, its maturity rate and the rate at which the investors change type and meet with a dealer to sell off their positions:

$$NPV = \frac{1}{r + \omega + \gamma\alpha^D} \tag{2.7}$$

Bargaining in the Secondary Markets

Dealers require a transaction fee on top of the interdealer prices. Transaction fees differ across long term and short term assets as well as across bid and ask prices. Fees will be determined as the outcome of the Nash bargaining, as a function of the trading surpluses. Trading surplus for an investor selling their asset is given by $V_0 - V_1 + P - \phi$ and trading surplus for an investor buying asset is given by $V_1 - V_0 - P - \phi$. Trading surplus for a dealer trading an asset is given by ϕ. Bargaining power of the dealer is given by η, interdealer price of an asset is given by P, and the value function of an investor is given by V (where V_1 is for an investor holding an asset and V_0 is for an investor with no asset). Then the transaction fees will be given by the following two equations.

Transaction fee when buying from a dealer:

$$\phi_A = \eta[V_1 - V_0 - P] \tag{2.8}$$

Transaction fee when selling to a dealer:

$$\phi_B = \eta[V_0 - V_1 + P] \tag{2.9}$$

where subscripts A and B denote if the transaction fee is for the ask price or the bid price respectively.

2.3.4 Value Functions

The value functions of the investors have subscripts $\{H, L\}$ to represent whether an agent is a high or low type; and $\{0\}$ to represent an agent not holding any assets, and $\{s, l\}$ to represent whether an agent is holding a short-term or a long-term asset.

The flow Bellman equations for the value functions are as follows. A high type investor holding a short term asset can become a low type investor holding a short term asset with a Poisson process with a rate γ_L and the change in his value function is $V_{Ls} - V_{Hs}$. With a Poisson process with a rate ω_s, investor's asset might mature, which turns him into a high type holding no asset. In this case, the change in his value function is $V_{H0} - V_{Hs}$. Finally, the investor might sell his asset back to a dealer, where the match occurs with a Poisson process with the rate α_s^D, and the investor becomes a high type holding no asset, receiving the trade surplus $(1-\eta)(V_{H0} - V_{Hs} + P_s)$. Since the investor is holding an asset, he will also get 1 unit of consumption as dividend payments:

$$rV_{Hs} = 1 + \gamma_L(V_{Ls} - V_{Hs}) + \omega_s(V_{H0} - V_{Hs}) + (1-\eta)\alpha_s^D(V_{H0} - V_{Hs} + P_s) \tag{2.10}$$

A high type investor holding a long term asset can become a low type investor holding a long term asset with a Poisson process with a rate γ_L and the change in his value function is $V_{Ll} - V_{Hl}$. With a Poisson process with a rate ω_l, investor's asset might mature, which turns him into a high type holding no asset. In this case, the change in his value function is $V_{H0} - V_{Hl}$. Since the long term asset is the preferred asset of the high type investors, he will not sell his asset back to a dealer. He will also get 1 unit of consumption for holding the asset:

$$rV_{Hl} = 1 + \gamma_L(V_{Ll} - V_{Hl}) + \omega_l(V_{H0} - V_{Hl}) \tag{2.11}$$

A low type investor holding a short term asset can become a high type investor holding a short term asset with a Poisson process with a rate γ_H and the change in his value function is $V_{Hs} - V_{Ls}$. With a Poisson process with a rate ω_s, investor's asset might mature, which turns him into a low type holding no asset. Hence, the change in his value function is $V_{L0} - V_{Ls}$. Since the short term asset is the preferred asset of the low type investors, he will not sell his asset back to a dealer. Low type investors pay δ while they are holding an asset, so the net dividend payment he gets is $1 - \delta$:

$$rV_{Ls} = 1 - \delta + \gamma_H(V_{Hs} - V_{Ls}) + \omega_s(V_{L0} - V_{Ls}) \tag{2.12}$$

A low type investor holding a long term asset can become a high type investor holding a long term asset with a Poisson process with a rate γ_H and the change in his value function is $V_{Hl} - V_{Ll}$. With a Poisson process with a rate ω_l, investor's asset might mature, which turns him into a low type holding no asset. Hence, the change in his value function is $V_{L0} - V_{Ll}$. Finally, the investor might sell his asset back to a dealer, where the match occurs with a Poisson process with the rate α_l^D, and the investor becomes a low type holding no asset, receiving the trade surplus $(1-\eta)(V_{L0} - V_{Ls} + P_l)$. He also receives the net dividend payments of $1-\delta$:

$$rV_{Ll} = 1 - \delta + \gamma_H(V_{Hl} - V_{Ll}) + \omega_l(V_{L0} - V_{Ll}) + (1-\eta)\alpha_s^D(V_{L0} - V_{Ll} + P_l) \tag{2.13}$$

When investors are not holding any assets, with a Poisson process with rate α_s^D they get matched with a dealer with a short-term asset, with a Poisson process with rate α_l^D they get matched with a dealer with a long-term asset, and purchase an asset depending on the result of the bargaining. It is also possible to purchase the asset directly from the issuer at the fixed price with no bargaining; where a match with an issuer of the short term asset happens with Poisson rate α_s^I, and a match with an issuer of the long term asset happens with Poisson rate α_l^I. Whether the investors purchase their preferred asset from the dealer or the investor depends on which one arrives first. Investors are not subject to the liquidity shock when they are not holding the asset. Again, with Poisson process with rate γ, they get hit by the liquidity preference shock. Bellman equation for a high type not holding an asset (V_{H0}):

$$rV_{H0} = \gamma_L(V_{L0} - V_{H0}) + \alpha_l^I(V_{Hl} - V_{H0} - NPV_l) + (1-\eta)\alpha_l^D(V_{Hl} - V_{H0} - P_l) \tag{2.14}$$

Bellman equation for a low type not holding an asset(V_{L0}):

$$rV_{L0} = \gamma_H(V_{H0} - V_{L0}) + \alpha_s^I(V_{Ls} - V_{L0} - NPV_s) + (1-\eta)\alpha_s^D(V_{Ls} - V_{L0} - P_s) \tag{2.15}$$

2.3.5 No Arbitrage Condition Between the Primary and the Secondary Markets

When purchasing an asset, investors are indifferent between the primary and the secondary markets. The choice depends only on whichever counterpart arrives first. If a dealer arrives first, the investor would buy the asset from the dealer at the bargaining price instead of waiting for an issuer. Since the arrival of an issuer or a dealer of the same type of asset are two independent Poisson processes, the probability that the dealer and the issuer will arrive first can be calculated respectively as follows:

$$Probability\{dealer\} = \frac{\alpha^D}{\alpha^D + \alpha^I}$$

$$Probability\{issuer\} = \frac{\alpha^I}{\alpha^D + \alpha^I}$$

For the high type investor to be indifferent between buying from an issuer and dealer, the following should hold:

$$(1-\eta)\frac{\alpha_l^D}{\alpha_l^D + \alpha_l^I}(V_{Hl} - V_{H0} - P_l) = \frac{\alpha_l^I}{\alpha_l^D + \alpha_l^I}(V_{Hl} - V_{H0} - NPV_l)$$

Similarly, for the low type investor,

$$(1-\eta)\frac{\alpha_s^D}{\alpha_s^D + \alpha_s^I}(V_{Ls} - V_{L0} - P_s) = \frac{\alpha_s^I}{\alpha_s^D + \alpha_s^I}(V_{Ls} - V_{L0} - NPV_s)$$

These two equations define the interdealer prices in terms of the net present values of the assets as well as the trade surpluses:[8]

$$P_l = \frac{[(1-\eta)\alpha_l^D - \alpha_l^I](V_{Hl} - V_{H0}) + \alpha_l^I NPV_l}{(1-\eta)\alpha_l^D} \tag{2.16}$$

$$P_s = \frac{[(1-\eta)\alpha_s^D - \alpha_s^I](V_{Ls} - V_{L0}) + \alpha_s^I NPV_s}{(1-\eta)\alpha_s^D} \tag{2.17}$$

2.3.6 Separating Equilibrium Conditions

For the separation equilibrium where high type investors choose the long term asset and the low type investors choose the short term asset to happen, the following conditions should hold:[9]

i. High type agents should benefit from buying long term asset at least as much as they would from buying short term assets:
$\alpha_l^I(V_{Hl} - V_{H0} - NPV_l) > \alpha_s^I(V_{Hs} - V_{H0} - NPV_s)$

ii. The benefit from trade should be nonnegative:
$(V_{Hl} - V_{H0} - NPV_l) \geq 0$

iii. High type agents should benefit from selling short term assets :
$(1-\eta)[V_{H0} - V_{Hs} + P_s] > 0$

iv. Low type agents should benefit from buying short term asset at least as much as they would from buying long term assets:
$\alpha_s^I(V_{Ls} - V_{L0} - NPV_s) > \alpha_l^I(V_{Ll} - V_{L0} - NPV_l)$

[8] Net present values of the assets are given in Equation-7.

[9] To save from notation, the comparisons are done with respect to buying from an issuer. As shown in the previous subsection, investors are indifferent from buying from a dealer or from an investor hence the same conditions hold if instead investors buy from a dealer.

v. The benefit from trade should be nonnegative:

   $(V_{Ls} - V_{L0} - NPV_s) \geq 0$

vi. High type agents should benefit from selling short term assets :

   $(1-\eta)[V_{L0} - V_{Ll} + P_l] > 0$

Next subsection solves the model depending on the population distributions.

2.3.7 Results

Solving for the value functions and plugging in for the transaction fees and the net present values of the assets, spreads can be solved as follows:

$Spread_s =$

$$\eta \frac{(r+\omega_s+\gamma_H\alpha_s^D)(\gamma_H-\alpha_s^I-\alpha_s^D(1-\eta))(r+\gamma_L+\omega_s+(1-\eta)\alpha_s^D)}{(r+\omega_s+\gamma_H\alpha_s^D)(r+\gamma_L+\omega_l)(\gamma_H\gamma_L-(\omega_s+(1-\eta)\alpha s^D)(\alpha_s^I+\alpha_s^D(1-\eta)))}$$

$$-\frac{\eta(r+\gamma_L+\omega_l)(\alpha_s^I+(r+\omega_s+\gamma_H\alpha_s^D)\alpha_s^D(1-\eta)P_s)(\gamma_L-\omega_s-(1-\eta)\alpha_s^D)}{(r+\omega_s+\gamma_H\alpha_s^D)(r+\gamma_L+\omega_l)(\gamma_H\gamma_L-(\omega_s+(1-\eta)\alpha s^D)(\alpha_s^I+\alpha_s^D(1-\eta)))}$$

$$+\eta\frac{\omega_s-\gamma_H}{\gamma_H}\frac{(r+\omega_l+\gamma_L\alpha_l^D)(\alpha_l^I+\alpha_l^D(1-\eta))(1-\delta+(1-\eta)\alpha_l^D P_l)}{(r+\omega_l+\gamma_L\alpha_l^D)(\omega_l(\alpha_l^I+\alpha_l^D(1-\eta))-\gamma_H\gamma_L)}$$

$$+\eta\frac{\omega_s-\gamma_H}{\gamma_H}\frac{\gamma_H(\alpha_l^I+(r+\omega_l+\gamma_L\alpha_l^D)\alpha_l^D(1-\eta)P_l)}{(r+\omega_l+\gamma_L\alpha_l^D)(\omega_l(\alpha_l^I+\alpha_l^D(1-\eta))-\gamma_H\gamma_L)}+\eta\frac{\delta-1}{\gamma_H} \tag{2.18}$$

$Spread_l =$

$$\eta\frac{(\gamma_H-\omega_l)(\alpha_l^I+(r+\omega_l+\gamma_L\alpha_l^D)\alpha_l^D(1-\eta)P_l)}{(r+\omega_l+\gamma_L\alpha_l^D)(\omega_l(\alpha_l^I+\alpha_l^D(1-\eta))-\gamma_H\gamma_L)}$$

$$+\eta\frac{(r+\omega_l+\gamma_L\alpha_l^D)(\alpha_l^I+\alpha_l^D(1-\eta)-\gamma_L)(1-\delta+(1-\eta)\alpha_l^D P_l)}{(r+\omega_l+\gamma_L\alpha_l^D)(\omega_l(\alpha_l^I+\alpha_l^D(1-\eta))-\gamma_H\gamma_L)}$$

$$+\eta\frac{\gamma_L-\omega_l}{r+\gamma_L+\omega_l}\frac{\gamma_L(r+\gamma_L+\omega_l)(\alpha_s^I+(r+\omega_s+\gamma_H\alpha_s^D)\alpha_s^D(1-\eta)P_s)}{\gamma_L(r+\omega_s+\gamma_H\alpha_s^D)(\gamma_H\gamma_L-(\omega_s+(1-\eta)\alpha_s^D)(\alpha_s^I+\alpha_s^D(1-\eta)))}$$

$$+\eta\frac{\gamma_L-\omega_l}{r+\gamma_L+\omega_l}\frac{(r+\omega_s+\gamma_H\alpha_s^D)(\alpha_s^I+\alpha_s^D(1-\eta))(r+\gamma_L+\omega_s+(1-\eta)\alpha_s^D)}{\gamma_L(r+\omega_s+\gamma_H\alpha_s^D)(\gamma_H\gamma_L-(\omega_s+(1-\eta)\alpha_s^D)(\alpha_s^I+\alpha_s^D(1-\eta)))}$$

$$-\frac{\eta}{\gamma_L} \tag{2.19}$$

Equations 18 and 19 are in terms of the interdealer prices, which depend on the trade surpluses as shown in Equations 16 and 17. Trade surpluses

- Number of investors normalized to 1:

$$\mu_{Hl}+\mu_{Hs}+\mu_{Ll}+\mu_{Ls}+\mu_{H0}+\mu_{L0}=1 \tag{2.20}$$

- Letting $s_l \equiv \mu_{Hl}+\mu_{Ll}$ and $s_s \equiv \mu_{Hs}+\mu_{Ls}$:

$$s_l+s_s+\mu_{H0}+\mu_{L0}=1 \tag{2.21}$$

- Then in steady state, the following hold:

$$\mu_{H0} = \frac{\omega_l s_l}{\alpha_l^I},\ \mu_{L0} = \frac{\omega_s s_s}{\alpha_s^I},\ \mu_H = \frac{\gamma_H}{\gamma_H+\gamma_L},\ \mu_L = \frac{\gamma_H}{\gamma_H+\gamma_L}$$

$$s_l\left(1+\frac{\omega_l}{\alpha_l^I}\right) + s_s\left(1+\frac{\omega_s}{\alpha_s^I}\right) = 1$$

2.4 Investigating On-the-run Phenomenon

This section of the chapter presents the model to account for the on-the-run phenomenon. On-the-run phenomenon arises from the "liquidity" and "specialness" of the short term treasury securities over the long term treasury securities. Short term treasury securities are more liquid compared to the long term ones with the same time to maturity remaining. This is because the long term securities have been issued previously and are locked in some investor's portfolio, and hence it is more difficult to find a counterpart to buy or sell a long term treasury security that was issued previously. Moreover, the interest rate on the on-the-run securities is lower than that on the off-the-run securities, which makes it more advantageous to borrow money in the repo market with on-the-run securities as collateral. This is the "specialness" of the on-the-run securities. On-the-run phenomenon is investigated with a search theoretical model by Vayanos and Weill (2008). By allowing short selling and repo markets, authors show that because of the price and transaction cost difference between the on-the-run and off-the-run securities, short-sellers prefer on-the-run securities, which makes these securities more liquid. This part of the chapter tries to account for the on-the-run phenomenon by focusing only on the liquidity characteristics of the two risk-free securities with the same time to maturity. Unlike in Vayanos and Weill (2008), short-selling is not allowed; hence investors can pick only one asset, and hold only 1 units of it.

To explain "on-the-run" phenomenon, the model has two assets with identical dividend payments (1 unit of consumption) and they have no face values. They differ in their maturities, i.e, one is short-term and the other one is long term. However, they have the same time to maturity. Hence the short-term asset mimics on-the-run securities and long term asset mimics the off-the-run securities. More specifically these two assets can be considered as a treasury note that has just been issued and a treasury bond that has been issued previously but has the same time-to-maturity remaining as the treasury note. Since they have the same time to maturity, the arrival rate of the maturity shock is ω for both assets. However, getting matched with a counterpart dealer is easier for investors seeking to buy or sell on the run securities. Hence, two assets differ in their rates of matches with the dealer. Since the interest here is to explain the choice between on the run and off the run assets, model only considers dealers for simplicity.[10] A match with a dealer for a short term asset occurs with a Poisson rate α_s and a match with a dealer for a long term asset occurs with a Poisson rate α_l, where $\alpha_s > \alpha_l$. Again, there are two types of investors, high types and low types. Low types who hold an asset pay a holding cost δ during the holding period. An investor can hold 1 unit of the either asset. When an investor is holding an asset, three things can happen. He can be hit by a liquidity shock which arrives with a Poisson rate γ and changes his type. With a Poisson rate ω his asset matures, and with the rate α_s or α_l depending on which asset he is holding, he can be matched with a counterpart dealer to sell his asset if there is benefit from this trade.

Investors are subject to the liquidity shock only when holding an asset. When investors are not holding any assets, they decide on which asset to buy depending on the difference between the value they will get from holding the asset and the ask price of that asset.

[10] Off-the-run securities can be provided only by the dealers, issuers would only matter for the on-the-run securities. However, considering only dealers for the on-the-run securities is sufficient to explain the phenomenon.

The change in the fraction of the investors holding 0 assets are given by the following two equations, similar to the general model:

$$\dot{\mu_{H0}} = -\gamma_L \mu_{H0} + \gamma_H \mu_{L0} + \omega(\mu_{Hs} + \mu_{Hl}) + \alpha_s \min\{\mu_{Hs}, \mu_{L0}\} - \alpha_l \min\{\mu_{H0}, \mu_{Ll}\} \tag{2.22}$$

$$\dot{\mu_{L0}} = -\gamma_H \mu_{L0} + \gamma_L \mu_{H0} + \omega(\mu_{Ls} + \mu_{Ll}) - \alpha_s \min\{\mu_{Hs}, \mu_{L0}\} + \alpha_l \min\{\mu_{H0}, \mu_{Ll}\} \tag{2.23}$$

The change in the fraction of the investors holding their preferred assets are given by the following two equations:

$$\dot{\mu_{Hl}} = -\gamma_L \mu_{Hl} - \omega \mu_{Hl} + \gamma_H \mu_{Ll} + \alpha_l \min\{\mu_{H0}, \mu_{Ll}\} \tag{2.24}$$

$$\dot{\mu_{Ls}} = -\gamma_H \mu_{Ls} - \omega \mu_{Ls} + \gamma_L \mu_{Hs} + \alpha_s \min\{\mu_{Hs}, \mu_{L0}\} \tag{2.25}$$

The change in the fraction of the investors holding their unwanted assets are given by the following two equations:

$$\dot{\mu_{Hs}} = -\gamma_L \mu_{Hs} - \omega \mu_{Hs} + \gamma_H \mu_{Ls} - \alpha_s \min\{\mu_{Hs}, \mu_{L0}\} \tag{2.26}$$

$$\dot{\mu_{Ll}} = -\gamma_H \mu_{Ll} - \omega \mu_{Ll} + \gamma_L \mu_{Hl} - \alpha_l \min\{\mu_{H0}, \mu_{Ll}\} \tag{2.27}$$

Using the bargaining prices from the previous section, the Bellman equations can be written as follows:

$$rV_{Hs} = 1 + \gamma_L(V_{Ls} - V_{Hs}) + \omega(V_{H0} - V_{Hs}) + (1-\eta)\alpha_s(V_{H0} - V_{Hs} + P_s) \tag{2.28}$$

$$rV_{Hl} = 1 + \gamma_L(V_{Ll} - V_{Hl}) + \omega(V_{H0} - V_{Hl}) \tag{2.29}$$

$$rV_{Ls} = 1 - \delta + \gamma_H(V_{Hs} - V_{Ls}) + \omega(V_{L0} - V_{Ls}) \tag{2.30}$$

$$rV_{Ll} = 1 - \delta + \gamma_H(V_{Hl} - V_{Ll}) + \omega(V_{L0} - V_{Ll}) + (1-\eta)\alpha_s(V_{L0} - V_{Ll} + P_l) \tag{2.31}$$

Bellman equation for a high type not holding an asset (V_{H0}):

$$rV_{H0} = \gamma_L(V_{L0} - V_{H0}) + (1-\eta)\alpha_l(V_{Hl} - V_{H0} - P_l) \tag{2.32}$$

Bellman equation for a low type not holding an asset(V_{L0}):

$$rV_{L0} = \gamma_H(V_{H0} - V_{L0}) + (1-\eta)\alpha_s(V_{Ls} - V_{L0} - P_s) \tag{2.33}$$

When the conditions for the separating equilibrium where the high types choose "off-the-run" assets and low types choose "on-the-run" assets, the spreads will be given by:

$$Spread_s =$$

$$\eta\frac{(r+\omega+\gamma_H\alpha_s)(\gamma_H-\alpha_s(1-\eta))(r+\gamma_L+\omega+(1-\eta)\alpha_s)}{(r+\omega+\gamma_H\alpha_s)(r+\gamma_L+\omega_l)(\gamma_H\gamma_L-(\omega+(1-\eta)\alpha s)(\alpha_s^D(1-\eta)))}$$

$$-\frac{\eta(r+\gamma_L+\omega)((r+\omega+\gamma_H\alpha_s)\alpha_s(1-\eta)P_s)(\gamma_L-\omega-(1-\eta)\alpha_s)}{(r+\omega+\gamma_H\alpha_s)(r+\gamma_L+\omega)(\gamma_H\gamma_L-(\omega+(1-\eta)\alpha s)(\alpha_s(1-\eta)))}$$

$$+\eta\frac{\omega-\gamma_H}{\gamma_H}\frac{(r+\omega+\gamma_L\alpha_l)(\alpha_l(1-\eta))(1-\delta+(1-\eta)\alpha_l P_l)}{(r+\omega_l+\gamma_L\alpha_l)(\omega(\alpha_l+\alpha_l(1-\eta))-\gamma_H\gamma_L)}$$

$$+\eta\frac{\omega-\gamma_H}{\gamma_H}\frac{\gamma_H((r+\omega_l+\gamma_L\alpha_l)\alpha_l(1-\eta)P_l)}{(r+\omega+\gamma_L\alpha_l)(\omega(\alpha_l+\alpha_l(1-\eta))-\gamma_H\gamma_L)}+\eta\frac{\delta-1}{\gamma_H} \tag{2.34}$$

$$Spread_l = \eta\frac{(\gamma_H-\omega)((r+\omega+\gamma_L\alpha_l)\alpha_l(1-\eta)P_l)}{(r+\omega+\gamma_L\alpha_l)(\omega(\alpha_l(1-\eta))-\gamma_H\gamma_L)}$$

$$+\eta\frac{(r+\omega+\gamma_L\alpha_l)(\alpha_l(1-\eta)-\gamma_L)(1-\delta+(1-\eta)\alpha_l P_l)}{(r+\omega+\gamma_L\alpha_l)(\omega(\alpha_l(1-\eta))-\gamma_H\gamma_L)}$$

$$+\eta\frac{\gamma_L-\omega}{r+\gamma_L+\omega}\frac{\gamma_L(r+\gamma_L+\omega)((r+\omega+\gamma_L\alpha_s)\alpha_s(1-\eta)P_s)}{\gamma_L(r+\omega+\gamma_H\alpha_s)(\gamma_H\gamma_L-(\omega+(1-\eta)\alpha_s)(\alpha_s(1-\eta)))}$$

$$+\eta\frac{\gamma_L-\omega}{r+\gamma_L+\omega}\frac{(r+\omega+\gamma_H\alpha_s)(\alpha_s(1-\eta))(r+\gamma_L+\omega+(1-\eta)\alpha_s)}{\gamma_L(r+\omega+\gamma_H\alpha_s)(\gamma_H\gamma_L-(\omega+(1-\eta)\alpha_s)(\alpha_s(1-\eta)))}$$

$$-\frac{\eta}{\gamma_L} \tag{2.35}$$

2.5 Numerical Applications to "On-the-Run" Phenomenon

This section presents the numerical results of the general model since closed form solutions are not easy to interpret. The values for interest rate r and the bargaining power of the dealer η are taken as 0.05 and 0.8 following Duffie et al. (2005). The low type agents are assumed to face a liquidity cost of $\delta = 0.01$.

For the other parameters, various numerical values were explored, where the results hold for a great range of values. The following sections investigate how changes in various parameters affect the bid and ask spreads of the "on-the-run" and "off-the-run" securities. Bid and ask spread is found to be always higher for the off-the-run asset compared to the on-the-run asset, as seen in the data.

2.5.1 Changes in Bargaining Power

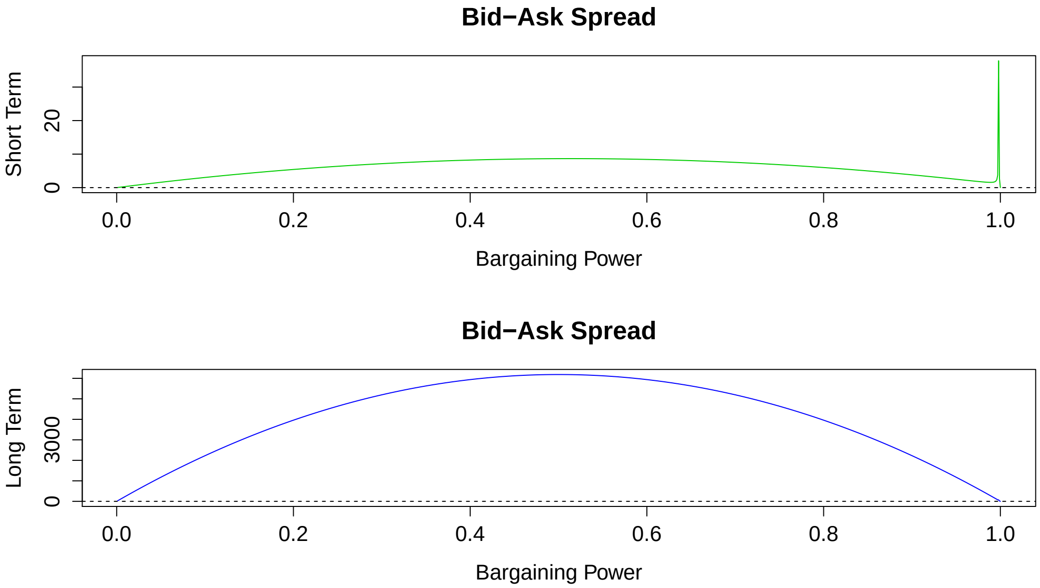


Figure 2.4: Spreads as η changes

Figure-2.4 shows how spreads change in response to the bargaining power. Spreads get higher when the bargaining power is around 0.5, and it falls as one counterpart stars to get more power.

2.5.2 Changes in Illiquidity Cost

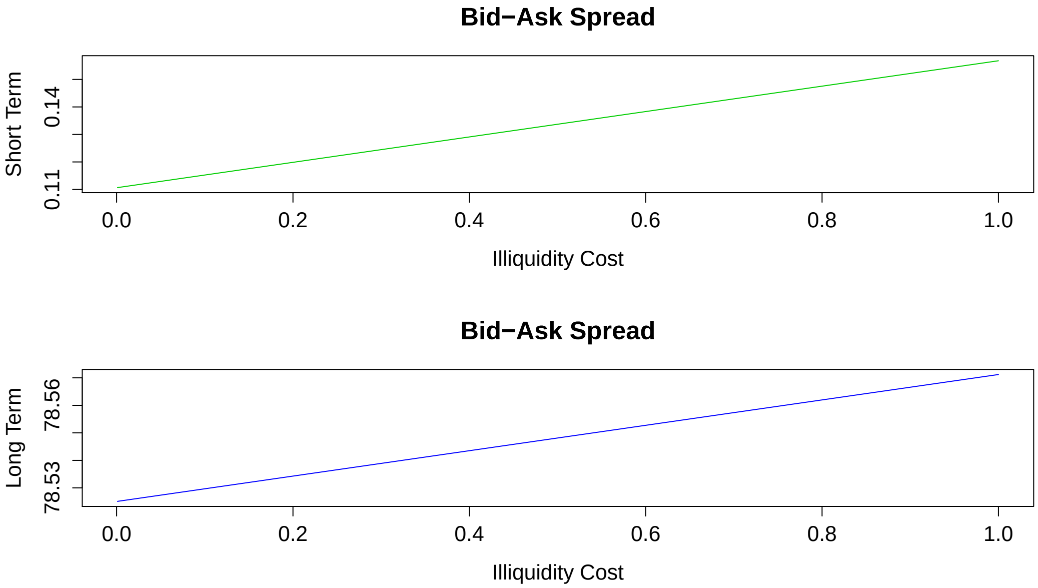


Figure 2.5: Spreads as δ increases

Figure-2.5 shows that spreads increase linearly in response to the illiquidity cost, since the cost is an additive cost.

2.5.3 Changes in Rate of the Preference Shock

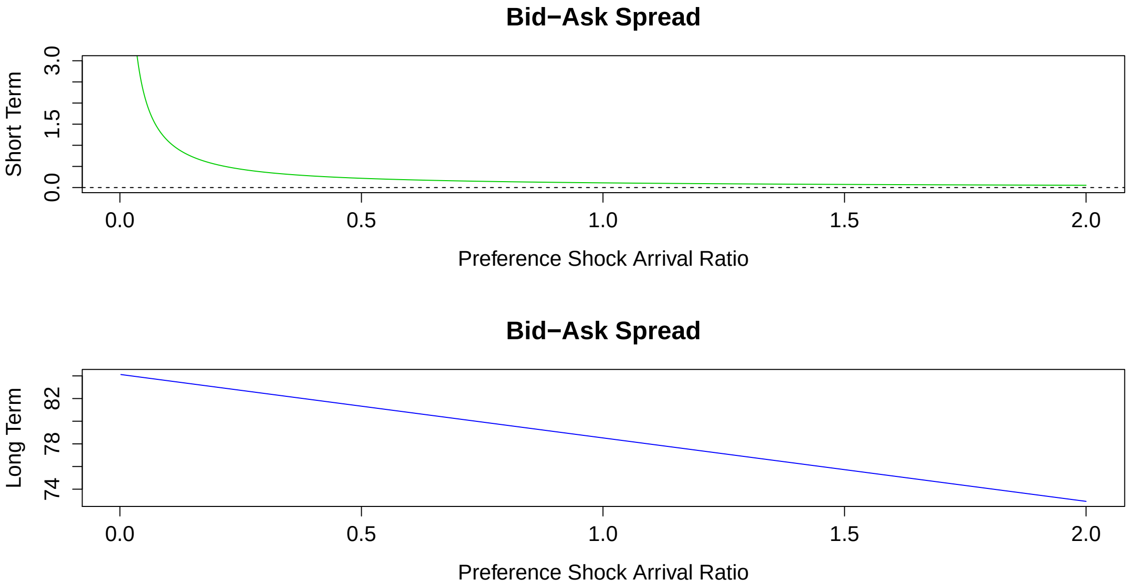


Figure 2.6: Spreads as γ ratio increases

Figure-2.6 presents how spreads change in response to the γ_H/γ_L ratio. Spread of short term approaches to zero as the ratio increases. This is because there are less investors who are low type and who hold short term asset as γ_L increases against γ_H.

2.5.4 Changes in Trading Friction Ratios

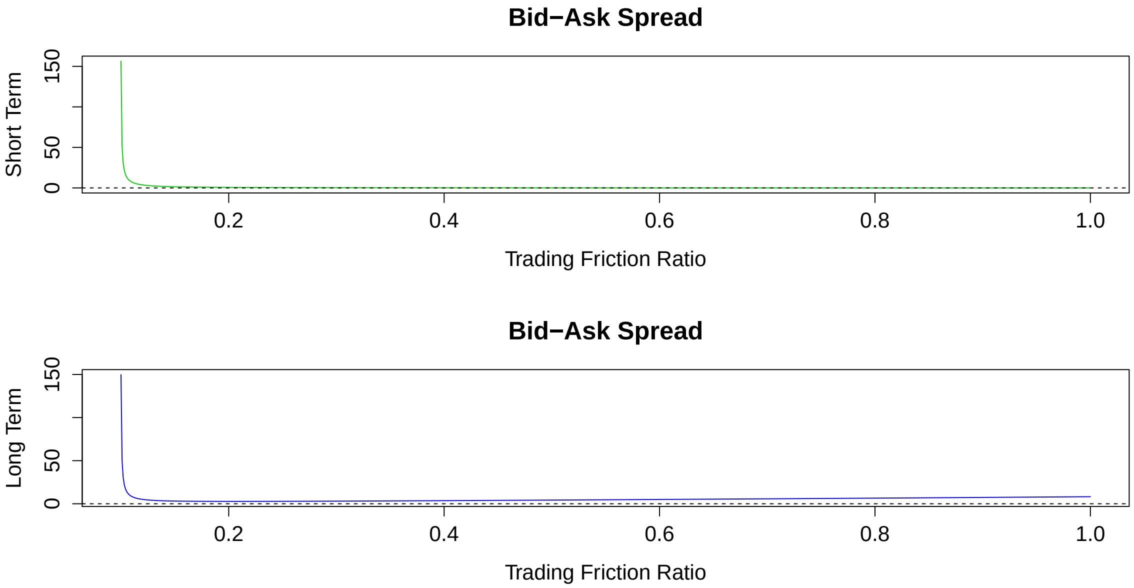


Figure 2.7: Spreads as α ratio increases

Figure-2.7 shows how spreads change as the ratio of α_l/α_s increases, i.e., when it gets easier to find a dealer of a long term asset. As the ratio increases, spreads of both assets fall. When the ratio is in close vicinity of 1, spreads approach to zero.

2.6 Conclusion

This chapter studied choice of heterogeneous investors with different liquidity preferences over assets that differ in their maturities under the presence of search costs in order to explain the "on-the-run" phenomenon. Under the general model which had two assets with identical dividend payments with different maturities, the model predicts a separating equilibrium where high type agents choose the long term asset and the low type agents choose the short term asset. This shows that liquidity cost matters in the presence of search frictions. The same equilibrium was obtained from the model that was constructed to explain the on-the-run phenomenon. Moreover, the model shows that the spread of the long term asset is always higher than that of the short term asset when they have the same time-to-maturities, which goes in line with the data. Henceforth, “on-the-run” phenomenon can be explained by higher search frictions in the off-the-run markets and investors with different liquidity preferences.

Chapter 3

Exchange Rate Forecasting from Twitter's Trending Topics

This chapter forecasts intra-day foreign exchange rates by making use of trending topics from Twitter, using a sentiment based topic clustering algorithm. Twitter trending topics data provide a good source of high frequency information, which would improve the short-term or intra-day exchange rate forecasts. This project uses an online dataset, where trending topics in the world are fetched from Twitter every ten minutes since July 2013. First, using a sentiment lexicon, the trending topics are assigned a sentiment (negative, positive, or uncertain), and then using a continuous Dirichlet process mixture model, the trending topics are clustered regardless of whether they are explicitly related to the currency under consideration. This unique approach captures the general sentiment among users, which implicitly affects the currencies. Finally, the exchange rates are estimated using a linear model which includes the topic based sentiment series and the lagged values of the currencies, and a VAR model on the topic based sentiment time series. The main variables of interest are Euro/USD, GBP/USD, Swiss Franc/USD and Japanese Yen/USD exchange rates. The linear model with the sentiments from the topics and the lagged values of the currencies is found to perform better than the benchmark AR(1) model. Incorporating sentiments from tweets also resulted in a better forecast of currency values after unexpected events.

3.1 Background

Making use of the data available from social media helps incorporate the beliefs, expectations, information levels and behaviors of agents into the models. This has been vastly explored in finance, and lately in economics. Using social media data with machine learning algorithms is found to increase the predictive power of models in these disciplines. This chapter utilizes text data from social media in exchange rate prediction models. In particular, the aim is to forecast foreign exchange rates making use of the trending topics from Twitter, using natural language processing.

Previous studies that aim at forecasting exchange rates by extracting sentiment from text data have used news or news headlines as their features. This chapter takes one step ahead and uses Twitter trending topics data instead of news. Trending topics come from 400 million potential users, rather than a much smaller number of reporters, therefore representing a much greater variety of sentiment. Another unique approach taken in this chapter is that the trending topics are taken irrespective of whether the tweets explicitly mention the currencies under consideration, whereas previous studies analyze the tweets that contain the names or symbols of the stocks or the currencies that they are interested in. This approach enables to capture any implicit effect of the trending topics on the currency values.

Trending topics come from 400 million potential users, which represent a much greater variety of sentiment. The data comes from an online dataset, where trending topics in the world are fetched from Twitter every ten minutes since July 2013.[1] Twitter trending topics are words, phrases and hashtags that suddenly increase in popularity, rather than

[1] The data are being collected from the website: `http://tt-history.appspot.com/`.

being tweeted at a high volume. Trending topics are a good reflection of the concerns, interests, ideas and feelings of the users. Twitter trending topics data hence provides a good source of high frequency information, which would improve especially the short-term or intra-day exchange rate forecasts. The advantage of using the Twitter trending topics data set is that since it embodies news as well as people's reactions to the news, it also provides a reflection of sentiment or mood in the economy. The data sets used in the economics literature include only the news or the mood, but not both. This project will help fill in this gap in the literature; by using natural language processing on Twitter trending topics data, which incorporates both news and sentiment, to forecast the movements in foreign exchange markets.

The main variables of interest are Euro/ US Dollar, GBP/ US Dollar, Swiss Franc/ US Dollar and Japanese Yen/ US Dollar exchange rates, and the short term movements of these rates are forecasted using the sentiments clusters through a Dirichlet Process Mixture Model. This chapter is organized as follows. Related studies are discussed in Section-2. Section-3 introduces the data set and the preprocessing methods. Section-4 presents the model. Results are given in Section-5. Section-6 concludes.

3.2 Related Literature

This chapter contributes to two different strands of literature in two disciplines: exchange rate forecasting and machine learning. This is the first paper that makes use of data from social media to predict exchange rates with an unsupervised topic clustering model. Studies available in this literature mostly make use of news headlines as their data, and use neural networks or ensembles as their models. The novelty of this paper in exchange rate forecast literature is the use of Twitter data with a Dirichlet process mixture algorithm to cluster the sentiment-based topics to use them in a time series model to forecast the short term exchange rates. This chapter is also the first paper that makes use of unsupervised learning in topic extraction, in exchange rate forecasting using natural language processing.[2]

Long term exchange rates are affected by macroeconomic fundamentals such as trade balance, purchasing power parity, inflation, whereas short term fluctuations in exchange rates cannot be easily forecasted since there are no high frequency macroeconomic variables that can be used in forecast. Short term variations in exchange rates might be caused by some unobservable fundamentals that act like a noise such as the risk premium of a currency. However Wang et al. (2008) show that they do not have a role in forecasting long term variations. One improvement over the models with fundamentals was when Meese and Rogoff (1983) found that a random walk model was as useful as forecasting the exchange rates as fundamentals, which suggests that for an accurate forecast, relying only on fundamentals is not sufficient.

In addition to these monetary models, there are two more theoretical models that have been used to forecast the exchange rates: overshooting and portfolio balance models. The overshooting model by Dornbusch (1976) states that exchange rates overshoot their long run values temporarily when the fundamentals change. Portfolio balance model treats currencies like financial assets, therefore determining the exchange rates by the demand for and the supply of foreign and domestic currencies (Branson et al. (1977)). However, these models were also shown to be inaccurate and empirically insufficient (Backus (1984)). The failures of theoretical models call for more technical analyses in order to forecast short term movements in the currency values.

There are also many empirical studies that make use of econometric models such as GARCH, VAR, and spectral analysis for exchange rate forecasting, however they are still not sufficient for medium and short term forecasts. One reason why these empirical models are insufficient

[2] Unsupervised learning refers to the applications of machine learning techniques on data without having pre-specified outputs. This will be explained in more detail in the following sections.

is because they do not include the sentiment into their analyses. Exchange rates are shown to be affected by market sentiment instead of the fundamentals (Hopper et al. (1997)), and hence for an accurate forecasts, market sentiment should also be used. This chapter addresses this gap in the literature, by using natural language processing for sentiment and Dirichlet Process Mixture Model for topic extraction to forecast short term movements in the exchange rates.

There are many studies that make use of text data for short term exchange rate forecasting. This paper is unique in its use of data set and methodology. All papers that forecast exchange rate from text data employ news data or news headlines, and although their methods vary from neural networks to ensembles to latent Dirichlet allocation, Dirichlet process mixture models are not as commonly used.

There are two main strands of machine learning literature that aim to improve upon the current exchange rate prediction. Using neural networks in exchange rate prediction has been a popular method since the last two decades, and the advances in deep learning computing techniques sparked interest in neural networks for the last five years. However, the increasing popularity of microblogs and online news started a new era in information technology; which makes use of text data and the sentiment from the text data.[3] This paper falls into the second strand which uses text mining techniques in time series prediction.

3.2.1 Neural Network Models in Exchange Rate Prediction

Artificial neural networks are machine learning models that are inspired by the information processing of nervous system of animals.[4] A basic neural network model consists of three

[3] There are of course many other machine learning techniques used in exchange rate prediction, such as classification and ensembles; however the only main two are discussed here. For instance, Talebi et al. (2014) use Bayesian voting to find ensemble trends in forex market and show improvement over models that use only ensembles.

[4] For a detailed definition and history of neural networks, refer to Rojas (2013).

sets of interconnected nodes: input, hidden output. The data enters in the input node and is outputted usually as a classifier from the output node.

Kaastra and Boyd (1996) and Refenes et al. (1993) are two of the first studies which apply neural network models to financial and economic time series prediction. Kuan and Liu (1995) , Hann and Steurer (1996), Tenti (1996), Muhammad and King (1997), Gencay (1999) and Yao and Tan (2000), are among many earlier studies that show that using neural networks improve the prediction performance, on exchange rate data up to weekly frequency. More recent studies not only use higher frequency data but also more advanced techniques such as particle swarm optimization, fuzzy networks (both stochastic and non-stochastic), hybridization techniques with neural networks, and gene expressions Sermpinis et al. (2013), Sermpinis et al. (012a), Bahrepour et al. (2011), Ni and Yin (2009), Wang et al. (2008) and Sermpinis et al. (012b)).[5]

There are many advantages of using these computationally advanced techniques due to factors such as nonlinearity of the time series, latency, noisy time series (Palit and Popovic (2006)), and also due to their high prediction accuracy. For example, Evans et al. (2013) obtain 72.5% intra-day prediction accuracy where they predict if the currency value will go up or down, and 23.3% annualized net return by employing artificial neural networks and genetic algorithms, and they also show that the forex series are not randomly distributed. Using a recurrent Cartesian Genetic Programming evolved artificial network, Rehman et al. (2014) were able to reach up to 98.872% prediction accuracy. Although neural networks result in such high accuracy, one major drawback is that these techniques are not intuitive and often are not easy to express with models.

[5] These techniques are even more commonly applied in finance literature. For example, for applications in stock market forecasting, see Atsalakis and Valavanis (009a)), Atsalakis and Valavanis (009b), Kim and Shin (2007) and Hadavandi et al. (2010).

3.2.2 Text Mining in Exchange Rate Prediction

Text data based time series prediction models are not as computationally complex, and they often provide more intuitive results compared to the neural networks based studies. These models became popular with the availability of big digital data and the tools to analyze these big data, such as text mining algorithms. Collecting high volumes of digital data and extracting information for description and prediction purposes is called knowledge discovery in databases (Fayyad et al. (1996)). Text mining is a part of a knowledge discovery in databases, which is concerned with extracting and enumerating patterns in the text data to be used in further analyses. Text data resources vary from online news websites to microblogs or social media accounts such as Twitter and Facebook to customer reviews on websites such as Amazon and Yelp.[6]

These sources provide unstructured qualitative data which require a careful pre-processing; which is especially the case for the social websites. The reason is that there are low barriers to entry to these websites, which makes it very easy for users to publish any information of their choice (Mitra and Mitra (2011)). This leads to misinformation due to factors that are often encountered in social media and online reviews such as use of sarcasm, trolling to entertain the readers as well as the writer, or giving fake opinions about businesses or people in return for money or discounts.[7][8][9] To avoid using wrong information the pre-processing should be done in a way to detect the real meaning of the sentences.

In order to obtain the real meaning from a text, natural language processing (NLP) tech-

[6] There is a very interesting study by Kwak et al. (010a) which analyzes the topological characteristics of Twitter and finds it to be closer to the online news media rather than social networks.

[7] For a list of sarcastic Amazon product reviews, for example, see `http://www.businessinsider.com/stupidest-amazon-product-reviews-2013-1?`

[8] Trolling is posting comments or opinions irrelevant to the context, and is especially encountered very commonly on YouTube. This piece of news talks about why some famous YouTubers opted out of having comments sections `http://www.theguardian.com/media-network/media-network-blog/2014/sep/18/psychology-internet-trolls-pewdiepie-youtube-mary-beard`.

[9] Luca and Zervas (2016) has an empirical study on Yelp review fraud which shows that the fraud reviews are very common and their extend depend significantly on business characteristics.

niques are commonly used. NLP is the first step in a study which uses text data. It enables computer human interaction to understand and manipulate text or speech data (Chowdhury (2003)). With these techniques, researchers are able to extract the real meaning, sentiment and sarcasm from a text; as well as check the language and grammar and suggest corrections.[10] [11]

Once the correct meaning is extracted from text documents, data are ready to be applied to machine learning techniques. Models that use news data in exchange rate prediction are more common than the ones with social media data. This is mostly because news data are more informative and require less preprocessing since the language and the grammar are often correct. Peramunetilleke and Wong (2002) use news headlines to predict intraday movements in exchange rates and their model outperforms the random walk model. In another attempt to predict short to middle term movements, Zhang et al. (2005) use a framework where they use the news instead of only the headlines and classify the news as good and bad news.[12] These studies use historical data whereas there are studies that aim to do real time prediction from news data as well. The end results of some of these studies are or will be available as a program to users. Forex-Foreteller by Jin et al. (2013) and the Curious Negotiator by Zhang et al. (2007) are two examples of automated exchange rate movement predictors from news data.[13]

The methodology in these studies is to predict movements in exchange rates depending on the negativity or positivity of news. There is, however, more information that can be extracted from news or any text data. These data are usually subjective data, and extracting

[10] Resorting to NLP with semi-supervised learning, Davidov et al. (2010) are able to recognize sarcasm in Twitter tweets and Amazon reviews with F-scores of 0.83 and 0.78 respectively.

[11] Eisenstein (2013) discusses and compares performances of various techniques to detect bad language on social media and Baldwin et al. (2013) looks at the linguistic noise in social media text.

[12] For the applications of news mining on stock markets, see Feuerriegel and Prendinger (2016), Leinweber and Sisk (2011), and Minev et al. (2012).

[13] Automated news based prediction is very common in finance also. For stock price prediction, see Mittermayer and Knolmayer (2006) for an application called NewsCATS and Hagenau et al. (2012).

this subjectivity, namely "sentiment", is called sentiment analysis or opinion mining.[14] The sentiment in a document can be extracted using NLP and computational linguistics to be used in further analyses.

Sentiment analysis has became a popular forecasting tool in finance with the rise of the social media. The seminal work by Bollen et al. (2011) show that by using sentiment, classified as calm, alert, sure, vital, kind and happy; the prediction accuracy in the movements of the Dow Jones Industrial index increases up to 86.7%. Zhang et al. (2011) uses more sentiment classes and extends the framework to predict Dow Jones, NASDAQ, S&P 500 and VIX. Google also has a patent by Bollen and Mao (2013), granted in 2013 which applies a self-organizing fuzzy neural network model on the public mood to forecast the Dow Jones Industrial Index. Sentiment analysis is being commonly used in other fields as well. Tumasjan et al. (2010) is an important work in political science where they apply sentiment analysis from Twitter to predict election results. Public health also benefits from sentiment analysis. For example, Paul and Dredze (2011) has an interesting work where they predict flu trends from Twitter sentiment.

Although sentiment analysis is a very common tool for these fields, there are only a few studies that make use of it in exchange rate prediction. Using sentiment with news data is even less common compared to the ones with social media data. Nassirtoussi et al. (2015) is one of the few studies that incorporate sentiments into forex prediction from news data. They use a multi-layer algorithm to solve the co-referencing in text, then to obtain sentiment and finally to update the model with the latest information. With this algorithm they can reach to an accuracy level of 83.33%.

The studies that use social media data in forex prediction almost always work within a subset of tweets that the authors define to be relevant to the currency whose rate is to be forecasted. Papaioannou et al. (2013) forecast hourly EUR/USD exchange rate using a neural network

[14] Liu (2012) provides a very thorough reference for sentiment analysis and its application in various fields.

algorithm with Twitter sentiment, and their model outperforms the random walk model. Their methodology is to take tweets that contain "EUR/USD" symbol, and to extract their sentiment, which is different than the framework in this paper. Here, the sentiment from all trending topics are extracted regardless of whether they contain the symbol of the currency in question, and are clustered in an unsupervised manner. In a similar study, Janetzko (2014) forecasts the EUR/USD exchange rate, which again outperforms the random walk model. Similar to Papaioannou et al. (2013), he also preselects the tweets, using a list of concepts and names related to Eurozone and sentiments. In a more supervised study, Ozturk et al. (2014) predict the movements in USD/TRY by extracting a buy, sell, neutral sentiment from tweets and then employing a logistic regression. This paper is novel in its approach as it picks tweets without assuming a predefined relationship, and clusters them using an unsupervised Dirichlet process mixture model. This way, selected tweets do not impose anything on the clusters, and hence the analysis is not limited to the researcher's definitions on what should be related to the currency rates under consideration.

3.2.3 Topic Models

Once sentiment is extracted from data, the usual next step is to cluster the documents depending on their sentiment. Some studies prefer to do classification instead of clustering, where classification is a supervised algorithm that classifies documents depending on whether they fit into pre-defined classes or not. Clustering is unsupervised; the algorithm lets the data decide how documents should be clustered together, without any pre-specified class labels. The researcher may decide to limit the number of clusters to a fixed number, or that could be decided by the data as well. Moreover, Dirichlet process is language agnostic.

Topic models are the generative models that define documents as mixtures of different topics. Topics of a document is usually decided by the inverse frequency of the given words. The reason for this is that the most frequent words in a document are usually unimportant words whereas the words about the topic of the document are mentioned only a few times. Since the aim is to extract the topic in a document, the ordering of the words in the document does not matter. The most common practice to find the frequency of words is using the Bag of Words model. In this model, occurrence count of each word is recorded in a big sparse matrix, where each column corresponds to every single word that exists in all documents (corpus), and the rows correspond to documents.[15] Hence, each entry shows how many times a word specified in a given column occurred in the document specified in the row. The resulting matrix, called a document-term matrix, is populated by zeros since only some of the words of a document are present in the corpus. Document-term matrix is then used as the input of the topic clustering algorithm.

Unsupervised topic clustering models are defined by a set of hidden (latent) topics. These topics represent the underlying semantic structure of the corpus (Ritter et al. (2010)). One of the most popular method is latent Dirichlet allocation (LDA) by Blei et al. (2003). In LDA, documents are mixtures over latent topics, which also have distributions over words. Ritter et al. (2011), Zhao et al. (2011), Shirota et al. (2014), Ritter et al. (2010), and Zhao and Jiang (2011) are some of the important studies that apply LDA to documents in various disciplines from finance to business. Jin et al. (2013) is the only study that uses LDA for exchange rate prediction from news data.

Although LDA is commonly used, it is not suitable for the analysis in this paper due to two main reasons. First, the trending topics data consist of very short text, often even less than the 140 character limit imposed by Twitter. LDA is a more appropriate model for longer documents. Secondly, for LDA, the researcher needs to assume and specify a number of topics for the corpus. However, in this study the aim is to have a completely unsupervised

[15] The collection of all documents under study are called a corpus. Hence the columns of the bag of words matrix are given by every word that exist in the corpus. In this study, each trending topic is considered a document, and the corpus is the all trending topics available in the data set.

algorithm, which decides on the number of topics in the corpus itself. Hence, this study resorts to Dirichlet process mixture models (DPMM) which are more suitable for short text data and the researcher does not necessarily have to assume a predetermined number of topics for the corpus.[16]

Neal (2000) is an excellent resource for DPMM algorithms for different cases of priors and sampling methods. Yu et al. (2010) and Yin and Wang (2014) apply DPMM to short news and tweet data. This paper follows the methodology of Si et al. (2013) closely. In that study, tweets that contain the symbols of certain stocks (given by the $ sign) are crawled from Twitter, whose sentiments are then used in a DPMM model to create a sentiment based time series. The stock prices are then predicted using a VAR model on these sentiment based time series. This study differs from that in some important aspects. That study only takes the tweets that are relevant to the stocks whose prices are to be predicted, whereas here all trending topics are used regardless of whether they are related to the currencies in consideration or not. Hence, the sentiment extracted from the data is a more general sentiment. Also, in that study full tweets are used, which are around 140 characters. The data on this paper comes from trending topics, which are much shorter than that since they often consist of a word or two. This makes the analysis more difficult since the tweets are clustered based on very short text (documents).

Next section describes the data set in more detail and explains the preprocessing and challenges with that step.

[16] Also, since there are no pre-specified set of topics, DPMM can be applied to any human language without necessarily knowing the meaning of the words.

3.3 Data Set and Pre-Processing

3.3.1 Data Description

Twitter data used in this study were obtained from a website which keeps a track of Twitter trending topics and for how long they were trending.[17] The data set includes the trending topics in the world from July 2013 to June 2015. The trending topics are fetched and timestamped every 10 minutes. There are 1,048,575 trending topics in the data set; 523,383 of which are the trending topics from the world. In the world subset, which is used in this paper, there are 102,552 unique timestamps and 139,365 unique trending topics. The first few rows of the trending topics in the world data set look as follows as given in Table-3.1

	woeid	time	trend	timestamp
0	1	10	#AnakCiumanGaraGaraSinetronn	1411716752
1	1	10	#LiamYouMakeUsHappy	1429925450
2	1	10	#ÖzgürBasınSusturulamaz	1418579426
3	1	10	Britney Is My Life	1394316703

Table 3.1: First 5 rows of the World trending topics data. Data is not sorted to give the reader an idea of the availability of the various characters.

The first column in the dataset indicates row number. The variable in the second column is "woeid", which is the geographical id assigned to the tweets by Twitter, where $woeid = 1$ indicates that the tweets were trending in the world. Third column is for the time when the tweets were fetched from Twitter, and they are all equal to 10 since the trends are fetched every 10 minutes. The trending topics are given in the fourth column, and their timestamps are in the fifth column.

As also seen in the first 5 rows, the data are not clean. The first problem with the data is the foreign characters. The second problem is the hashtagged entries. Since they start with a hashtag, the functions used on strings during the pre-processing part did not work on those

[17] The website `http://tt-history.appspot.com/`. is built by Mustafa Ilhan. The GitHub project can be found at `https://github.com/mustilica/tt-history`.

entries and hence their hashtags needed to be removed. Finally, most of the hashtagged entries are sentences without spaces, and to make sense of these entries, they needed to be separated into words. Next subsection explains how these problems are tackled.

The exchange rate data were gathered from the website called Forex Rate.[18] Historical data were extracted using per minute intervals, for Euro/Dollar, British Pound/Dollar, Dollar/Swiss Franc and Dollar/Yen exchange rates, from January 2013 to November 2015.Figure-3.1 shows how the exchange rates used in this paper change over time:

[18] `http://www.forexrate.co.uk`.

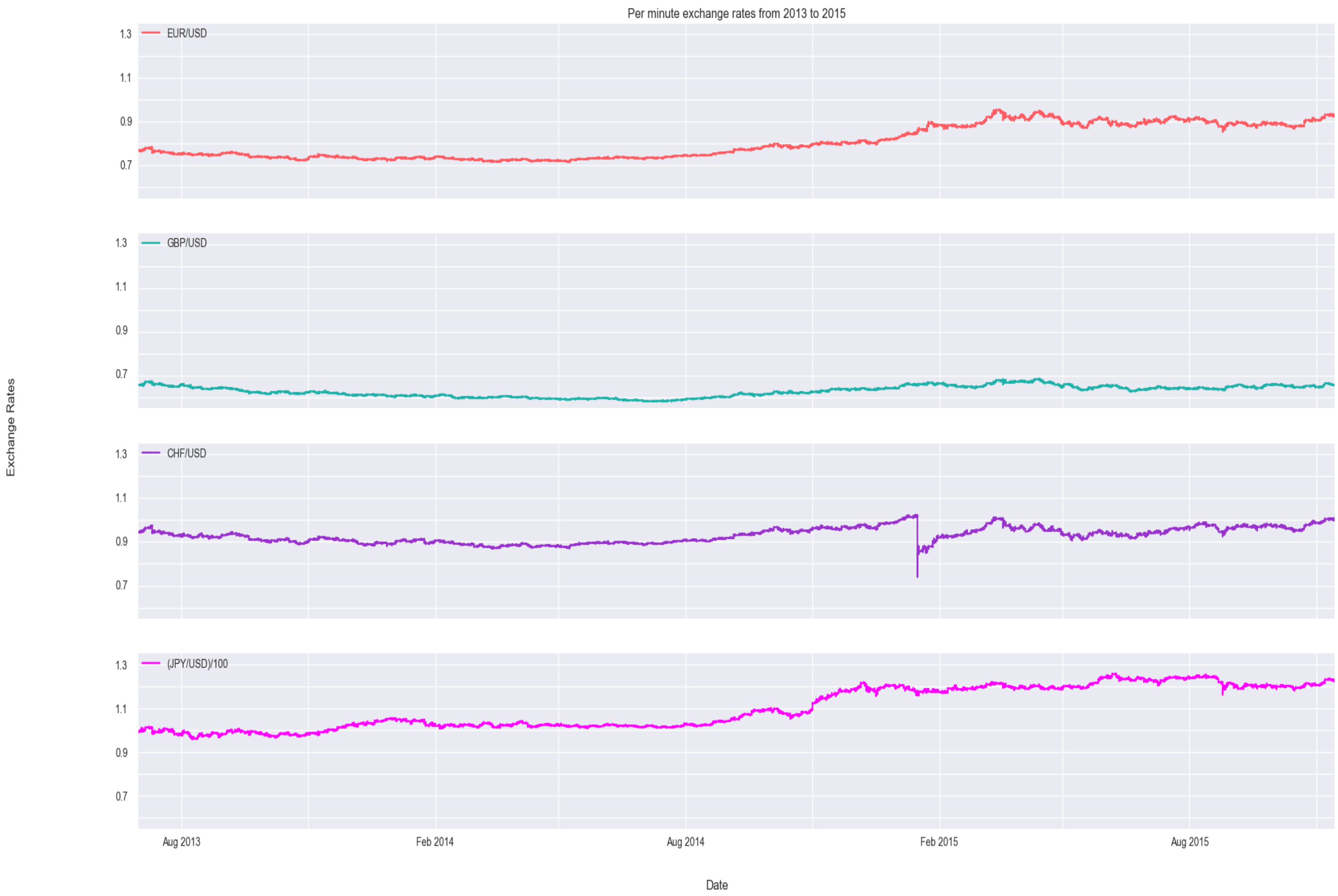


Figure 3.1: Exchange rates from 2013 to 2015.

3.3.2 Pre-Processing

Natural Language Processing (NLP) is the sub-division of artificial intelligence that deals with human-computer interaction. Applications of NLP techniques in research as well as in industry vary from interpreting text data, speech recognition, part-of-speech tagging to spam recognition. NLP applies to people's everyday life from auto text correction to voice commands.

The first step in handling the dataset was removing the hashtags from the entries that begin with a hashtag. Hashtagged entries become trending topics when users use those hashtagged words within their 140 character tweets. Once the hashtags are removed, a binary variable called "hashtag" is created in order not to lose information from that entry being hashtagged or not. The new variables are presented in Table-3.2.

trend	timestamp	nohash_trend	hashtag
#AnakCiumanGaraGaraSinetronn	1411716752	AnakCiumanGaraGaraSinetronn	1
#LiamYouMakeUsHappy	1429925450	LiamYouMakeUsHappy	1
#ÖzgürBasınSusturulamaz	1418579426	ÖzgürBasınSusturulamaz	1
Britney Is My Life	1394316703	Britney Is My Life	0

Table 3.2: Data set with hashtags removed: Contains two new variables, "nohashtrend" and the binary "hashtag".

After removing hashtags, all entries were turned into lower-case entries since the functions used in the next steps consider all text with capital letters as meaningful words and disregard them. The next step was to separate the entires with no spaces into words. To do that, every letter combination in the entry is compared against the items in a use specified dictionary.[19]

Table-3.3 shows the data set which includes the lower-cased and separated entires:

Also, to reduce noise, only English entries are selected. This was done using the "langid" module

[19] The dictionary used here is the one called "words by frequency" which includes the most frequently and commonly used English words, including different verb forms and slang words. The dictionary is available at: `https://raw.githubusercontent.com/alseambusher/columbus/master/words-by-frequency.txt` . The function used to separate the entries with spaces is "infer_spaces()' by Mark Hal (from Mankind Software) and can be found at `https://github.com/mankindsoftware/tweetsec/blob/master/lib/wordFinder.py` The second function "put_spaces" was used to apply "infer_spaces" to only entries with no space.

trend	timestamp	lower-cased spaced trend	hash-tag
#AnakCiumanGaraGaraSinetronn	1411716752	an akc i um angara gara sine tron n	1
#LiamYouMakeUsHappy	1429925450	liam you make us happy	1
#ÖzgürBasınSusturulamaz	1418579426	öz g ür b a s ı n s u s t u r u l a m a z	1
Britney Is My Life	1394316703	britney is my life	0

Table 3.3: Data set with all lower-cased entries and separated entries. -Non-English letters print as if they are not lower case but this is due to encoding of the Latex document.

in Python, in which the"langid.classify()" function returns the expected language of a given string with its probability of being in that language. After removing the non-English entries, there are 196,567 rows left in the dataset.[20]

Stop-Word Removal

Stop-words are the words which do not add any meaning to the sentiment of the sentences, but rather help to make grammatical sense, such as "the", "a", "this" etc. Removing these words results in a more refined data for sentiment analysis purposes. Stop-words were identified by the "stopwords" corpus in the nltk module, and were removed by searching each entry for these words. Table-3.4 shows the data set which includes the entries with no stop-words, which are sorted in an ascending order of the timestamps.

trend	timestamp	trend with stop-word removed	hashtag
Monumental	1373234195	monumental	0
Andy Murray	1373236603	andy, murray	0
PRI	1373250449	pri	0
PREP	1373258875	prep	0
Lanata	1373259477	lanata	0

Table 3.4: Data set with stop words removed. Each entry contains the list of the words that remain in the trending topic after the stop words are removed.

Tokenization is the method of making all words in a sentence an independent entry such as words, phrases or symbols. Before tokenization, each entry in the data set were one string including multiple words separated by spaces. After tokenization, each entry becomes a list of strings (where strings are the words), which are separated by commas. For tokenization, "word_tokenize" was used from the nltk package. Table-3.5 shows the data set with tokenized

[20] The langid module was developed by Lui and Baldwin (2012), and can be obtained at `https://github.com/saffsd/langid.py`.

entries.

trend	timestamp	tokenized trend	hashtag
Monumental	1373234195	[‘monumental’]	0
Andy Muray	1373236603	[‘andy’, ’murray’]	0
PRI	1373250449	[‘pri’]	0
PREP	1373258875	[‘prep’]	0
Lanata	1373259477	[‘lanata’]	0

Table 3.5: Tokenized data. -Data are sorted in an ascending order of the timestamps.

Stemmer

Stemming is a method used for collapsing the words into distinct forms. This was needed to obtain the words with the same roots as well as combining derivationally related words as one single word. There are three main stemmers available in nltk module (Porter, Lancaster and Snowball). Porter stemmer, by Martin Porter is the most commonly used stemming algorithm, probably due to being one of the oldest stemming algorithms. Although it is a popular algorithm, since it is computationally intensive, it was not used in this paper. Lancaster stemmer is considered to be an aggressive stemmer, and often returns unintuitive words. It is often used due to being the fastest stemming algorithm. Snowball, again by Martin Porter, is a slightly faster version of the Porter Stemmer, and less aggressive than Lancaster stemmer. In this paper Snowball stemmer was used due to being fast enough and not aggressive. This stemmer returns unicoded string values, which were encoded to ASCII after stemming.

Table-3.6 shows the data set which includes the lower-cased and separated entires.

trend	timestamp	tokenized trend	hashtag
Monumental	1373234195	[‘monument’]	0
Andy Muray	1373236603	[‘andi’, ’murray’]	0
PRI	1373250449	[‘pri’]	0
PREP	1373258875	[‘prep’]	0
Lanata	1373259477	[‘lanata’]	0

Table 3.6: Tokenized data. -Data are sorted in an ascending order of the timestamps.

Sentiment Lexions

In order to extract the sentiment from the stemmed trending topics, the dictionary developed by Loughran McDonald was used.[21] This extensive dictionary contains more than 80,000 words, and was updated in 2014. The words included in this lexicon contain at least two letters, so one letter words are excluded since they are not regarded as critical to the content. Table-A.1 in the appendix prints the first 70 rows of this lexicon.

The lexicon includes the word count, proportion of the word compared to the other words in the scanned documents, the average proportion and the standard deviation, number of documents the word occurred, whether the word has a negative, positive, uncertain, litigious, constraining, superfluous or interesting sentiment. The variable modal indicates how strong the modal is, taking 3 possible values where "1" corresponds to the "strong modal" (such as "always"), "2" corresponds to moderate modal" (such as "usually"), and "3" corresponds to "weak modal" (such as "almost"). Irregular verb shows whether the verb is irregular, syllables indicate how many syllables the word has. Harvard IV is the sentiment from derived from "Harvard Psycho-sociological Dictionary, which is constructed from applications in psychology and sociology. Finally, source shows which dictionary the authors used, and 2of12inf is the dictionary that includes words without abbreviations, acronyms, or names; and the words are not in the stemmed form in order to capture more content.

In order to make use of this dictionary, it was preprocessed to suit the Twitter dataset. Since this dictionary includes the inflections and different word form for the words originating from the same root, the entries in the dictionary were stemmed using the same algorithm which was used to stem the words in the trending topics dataset. After the words are stemmed, there were multiple entires with the same stem coming from different word forms. These

[21] The sentiment dictionary is available at: `http://www3.nd.edu/~mcdonald/Word_Lists_files/LoughranMcDonald_MasterDictionary_2014.xlsx`. Documentation about this dictionary can be found at: `https://www3.nd.edu/~mcdonald/Word_Lists_files/Documentation/Documentation_LoughranMcDonald_MasterDictionary.pdf`.

multiple entries were reduced into one single entry. The final sentiment dataset is given in Table-A.3 in the appendix.

3.4 Model

Topics and words from the data set are exchangeable (De Finetti (1989)) and have a representation as an infinite mixture distribution. When exchangeability is satisfied, the most commonly used topic clustering method is Latent Dirichlet Allocation (LDA), following Blei et al. (2003). According to LDA, *words* are the basic unit of discrete data who are of *documents*, whose collection make up the text *corpus*. LDA is a generative probabilistic model where documents are random mixtures over latent topics, and each topic has a distribution over words. Although it is common practice to use LDA in this context, for 2 main reasons it is not suitable for the analysis in this paper.

Firstly, LDA requires large documents and hence a large corpus. In this study, each trending topic is considered to be a document, which are limited to 140 characters by Twitter, and in fact are almost always shorter than 140 characters. Secondly, LDA requires a pre-specified number of topics to cluster the documents into. However, the topics in the trending topics dataset are too broad to pre-specify a number. Hence, in order to avoid limiting the analysis to a pre-specified number of topics and to be able to treat all tweet entries as one document, a Dirichlet Process Mixture model was used instead.

Dirichlet Process Mixture Model

The model that would be applicable the short documents (tweets) and high number of topics is a Dirichlet Process Mixture Model. This model accommodates very short documents and hence it is widely used in studies that make use of Twitter data, and also does not limit the analysis to a pre-specified number of topics.

A Dirichlet process is a probability distribution over probability distributions. Dirichlet process mixture models are hierarchical models that consist of K Dirichlet components. K can be a finite number; in that case there are a finite number of components (in the scope of this paper, "topics"), or K can be taken to infinity to represent infinite number of topics (clusters).

Following the Algorithm 3 from Neal (2000), topics are clustered as follows:

- Each trending topic (document) x, exchangeably and independently comes from a mixture of distributions F(θ), over the parameter θ:

  $$x_i|\theta_i \sim F(\theta_i)$$

- Each θ_i is the parameter of the mixing distribution G:

  $$\theta_i|G \sim G$$

- where G is generated from a Dirichlet process prior with a concentration parameter α and a base measure ζ:

  $$G \sim DP(\zeta, \alpha)$$

- Then, x_i can be represented as:

  $$x_i \sim \lim_{k \to \infty} \sum_{k=1}^{K} \pi_k p(x_i|y_i = k)$$

- where y_i is the cluster label assigned to document x_i and π_k is the weight of the corresponding cluster ($\pi_k \geq 0$ and $\sum_k \pi_k = 1$), K is the number of topics, and $p(x_i|y_i = k)$ is the topic models $\{\phi_k\}_{k=1}^K$. Since the trending topics entries are very small (even less than the 140 character limit by Twitter), there is only one topic y_i in each trend x_i.

The number of topics, K is not a pre-determined number, but rather is estimated from the given time-stamp's trending topics. This is a similar framework to that of Si et al. (2013), where the authors estimate topics for each day whereas here instead of each day, in order to estimate the topics for each time-stamp; in order to capture the intra-day changes. Similar to this framework, it is also acknowledged in this paper that the neighboring timestamps might have the same or related topics and hence they evolve with time. For this reason, the model has a dynamic generative process. The observed trending topics $\{x_{n,i}\}_{i=1}^{|T_n|}$ are grouped for each timestamp T_n, $(n : 1, .., N)$, and are generated by the latent topics $\{\theta_{n,i}\}_{i=1}^{|T_n|}$. The topics are generated for each timestamp T_n, so a DPM model is built on every time stamp. The topics learned in the previous timestamp are used as priors for the topics in the current timestamp, which is the continuous nature of this model.

For the very first timestamp, there is no previous topic that was learned that could work as prior. Although the model is a continuous DPM, for the first entry, a standard DPM was needed to be used where only two parameters that start the generating process are α_0 and ζ_0. For all other timestamps at any time n, the priors are α_n, ζ_n and G_{n-1}. Then, the probability that x_i belongs to cluster k is given in two components as follows:

$$P(y_i = k|y_{-i}, x_i, \alpha, \zeta) \propto P(y_i = k|y_{-i}, \alpha)P(x_i|x_{-i}, y_i = k, y_{-i}, \zeta) \tag{3.1}$$

The first component can take four different values depending on whether k is a new topic or not, and whether it takes a symmetric base prior or a topic as a prior:

$$P(y_i = k|y_i, \alpha) = \begin{cases} \dfrac{\alpha}{l-1+\alpha} & \text{if } k \text{ is a new topic and takes a symmetric base prior} \\ \dfrac{\alpha\pi_{n-1,k}}{l-1+\alpha} & \text{if } k \text{ is a new topic and takes a topic as prior} \\ \dfrac{l_k^{-i}}{l-1+\alpha} & \text{if } k \text{ is an existing topic} \end{cases} \tag{3.2}$$

The second component can be rewritten as:

$$P(x_i|x_{-i}, y_i = k, y_{-i}, \zeta) = \int P(x_i|\theta_k)\left[\Pi_{j\neq i, y_j=k} P(x_j|\theta_k)\right] G(\theta_k|\zeta)d\theta_k \tag{3.3}$$

Figure-3.2 shows the process of the model graphically:

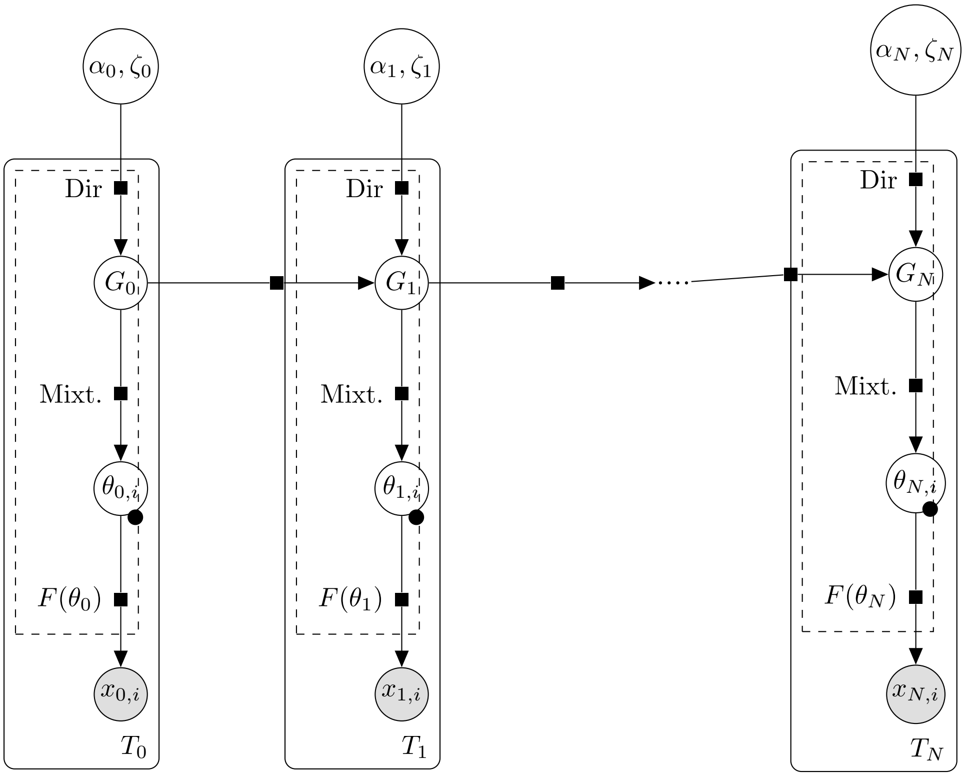


Figure 3.2: Continuous Dirichlet Process Mixture Model

Model Inference

Following Bishop (2006) and Si et al. (2013), a collapsed Gibbs sampling was used with hyper-parameters following from Si et al. (2013), Sun et al. (2010) and Teh et al. (2006). Accordingly, the concentration parameter α is taken to be constant at 1, and each base measure follows a Dirichlet distribution, $\zeta_n \sim Dir(\beta)$, where $\beta = 0.5$.

When sampling for the current topic y_i, there are two things that can happen. The label of the topic can be a new topic k^* or can be an existing topic k. In both cases, there are two different scenarios. If it is a new topic, then k^* can either take a symmetric base prior $Dir(\beta)$, or it can take one of the topics from $\{\phi_{n-1,k}\}_{k=1}^{K_{n-1}}$ learned from tweets T_{n-1}, where K_{n-1} is the number of topics learned at timestamp $n-1$. If k^* is new and takes a symmetric base prior, then the posterior becomes:

$$p(y_i = k^*|y_{-i}, x_i, \zeta) \sim \frac{\alpha}{l-1+\alpha} \frac{\Gamma(\beta|W|)}{\Gamma(\beta|W|+l_i)} \frac{\Pi_{w=1}^{|W|}\Gamma(\beta + l_{i,w})}{\Pi_{w=1}^{|W|}\Gamma(\beta)} \tag{3.4}$$

where $|W|$ is the size of the vocabulary, l_i is the length of tweet x_i, and $l_{i,w}$ is the "term frequency" of word w in x_i. If k^* is new but takes a topic from $\{\phi_{n-1,k}\}_{k=1}^{K_{n-1}}$ from the learned topics from T_{n-1}, then the posterior becomes:[22]

$$p(y_i = k^*|y_{-i}, x_i, \zeta) \sim \frac{\alpha\pi_{n-1,k}}{l-1+\alpha} \frac{\Gamma(\beta|W|)}{\Gamma(\beta|W|+l_i)} \frac{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w) + l_{i,w})}{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w))} \tag{3.5}$$

where $\phi_{n-1,k}(w)$ is the probability of word w in the topic k of the previous day.

If k is an existing topic, then its prior is already known. In this case, k would either take a symmetric base prior $\phi_{n,k}(w) = (\beta + l_{k,w})/(\beta|W| + l_{k,(.)})$, where $l_{k,w}$ is the frequency of word w in and $l_{k,(.)}$ is the marginalized sum over all words; or it can take $\phi_{n-1,k}$ as its prior, which

[22] Term frequency of a word shows how frequently that word occurs in the given document.

then will be used to calculate $\phi_{n,k}(w) = (\beta|W|\phi_{n-1,k}(w) + l_{k,w})/(\beta|W| + l_{k,(.)})$. Hence, if k is an existing topic which takes a symmetric base prior, then the posterior becomes:

$$p(y_i = k|y_{-i}, x_i, \zeta) \sim \frac{l_k^{-i}}{l-1+\alpha} \frac{\Gamma(\beta|W| + l_{k,(.)}^{-i})}{\Gamma(\beta|W| + l_i + l_{k,(.)}^{-i})} \frac{\Pi_{w=1}^{|W|}\Gamma(\beta + l_{i,w} + l_{k,w}^{-i})}{\Pi_{s=1}^{|W|}\Gamma(\beta + l_{k,w}^{-i})} \tag{3.6}$$

where l_k^{-i} is the number of tweets assigned to topic k except for the current tweet x_i, $l_{k,w}^{-i}$ is the term frequency of the word w in topic k, excluding the current tweet x_i, and $l_{k,(.)}^{-i}$ is the marginalized sum over all words in topic k, again excluding x_i. If, on the other hand, k is an existing topic with the prior $\phi_{n-1,k}$, then the posterior becomes:

$$p(y_i = k|y_{-i}, x_i, \zeta) \sim \frac{l_k^{-i}}{l-1+\alpha} \frac{\Gamma(\beta|W| + l_{k,(.)}^{-i})}{\Gamma(\beta|W| + l_i + l_{k,(.)}^{-i})} \frac{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w) + l_{i,w} + l_{k,w}^{-i})}{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w) + l_{k,w}^{-i})} \tag{3.7}$$

Topic weights for each day are calculated as:

$$\pi_k = \frac{l_k}{\sum_{k'} l_{k'}} \tag{3.8}$$

The sampling process for the current topic y_i given all other topics y_{-i} is summarized as follows:

Incorporating Topic Clusters into Sentiments

In the NLP section, each day's sentiment in the data set was matched using Loughran McDonald's sentiment lexicon. In order to simplify the model, only 3 sentiments were extracted from the tweets: negativity, positivity and uncertainty. This sentiments are collected under one variable named "sentiment", $S \in \{-1, 0, 1\}$, where the class labels $c = \{-1, 0, 1\}$ correspond to negative, neutral and positive respectively. The entries whose all three sentiment

Algorithm 1 Conditional distribution of y_i given all other assignments y_{-i}

```
if $k^*$ is a new topic then there are two candidate priors:
    if $k^*$ takes a symmetric base prior $Dir(\beta)$ then
        $p(y_i = k^*|y_{-i}, x_i, \zeta) \sim \frac{\alpha}{l - 1 + \alpha} \frac{\Gamma(\beta|W|)}{\Gamma(\beta|W| + l_i)} \frac{\Pi_{w=1}^{|W|}\Gamma(\beta + l_{i,w})}{\Pi_{w=1}^{|W|}\Gamma(\beta)}$
    if $k^*$ takes a topic from $\{\phi_{n-1,k}\}_{k=1}^{K_{n-1}}$ from the learned topics from $T_{n-1}$ then
        $p(y_i = k^*|y_{-i}, x_i, \zeta) \sim \frac{\alpha\pi_{n-1,k}}{l - 1 + \alpha} \frac{\Gamma(\beta|W|)}{\Gamma(\beta|W| + l_i)} \frac{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w) + l_{i,w})}{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w))}$
if $k$ is an existing topic then the prior is known:
    if $k$ takes a symmetric base prior $Dir(\beta)$ then
        $p(y_i = k|y_{-i}, x_i, \zeta) \sim \frac{l_k^{-i}}{l - 1 + \alpha} \frac{\Gamma(\beta|W| + l_{k,(.)}^{-i})}{\Gamma(\beta|W| + l_i + l_{k,(.)}^{-i})} \frac{\Pi_{w=1}^{|W|}\Gamma(\beta + l_{i,w} + l_{k,w}^{-i})}{\Pi_{s=1}^{|W|}\Gamma(\beta + l_{k,w}^{-i})}$
    if $k$ takes a topic from $\{\phi_{n-1,k}\}_{k=1}^{K_{n-1}}$ from the learned topics from $T_{n-1}$ then
        $p(y_i = k|y_{-i}, x_i, \zeta) \sim \frac{l_k^{-i}}{l - 1 + \alpha} \frac{\Gamma(\beta|W| + l_{k,(.)}^{-i})}{\Gamma(\beta|W| + l_i + l_{k,(.)}^{-i})} \frac{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w) + l_{i,w} + l_{k,w}^{-i})}{\Pi_{w=1}^{|W|}\Gamma(|W|\beta\phi_{n-1,k}(w) + l_{k,w}^{-i})}$
```

were zero are assumed to have neutral sentiment.

In order to incorporate topic clusters into sentiment dataset, the usual practice would be to create a topic based sentiment score. Sentiment score for topic k at timestamp n is given as the weighted sum of sentiment (opinion) classes, $c(o)$, weighted by the word probability for topic k, $\phi'_{n,k}(o)$:

$$S(n,k) = \sum_{o=1}^{|O|} \phi'_{n,k}(o)c(o) \tag{3.9}$$

However, since since each trending topic is considered to be a document, not to lose any information, the weight of words are assumed to be 1. The final dataset is converted to a sparse $0-1$ matrix, using CountVectorizer from scikit-learn package in Python. The size of this matrix is 34 GB, consisting of 196,567 rows and 20,979 columns. The columns are given by the unique words in the data set; hence there are 20,979 unique words in the corpus. However, most of these words (which make up the columns) were not meaningful or not full words. Instead of working with all words, word count and sentiment (positive, negative, uncertain) at each timestamp is used to cluster the topics. The algorithm divided the tweets into 9 unique topics.

3.4.1 Sentiment Time Series Prediction

Two different methods are used to forecast exchange rates. First, an AR model is fitted to the exchange rate series $\{z_t\}$ from July 2013 to June 2015 in order to obtain baseline values to compare to, without the topics from the tweets:

$$z_t = \alpha_0 + \alpha_1 z_{t-k} + \epsilon_{z,t} \tag{3.10}$$

where $\{\alpha\}$ are the model coefficients, $\{k\}$'s are the lags and $\{\epsilon\}$ are white noises.

Table-3.7 gives the number of observations from each currency during this period. Both AIC and BIC were considered while choosing the optimal lag, and they both resulted in 1-period lag. This implies that the best lagged predictor of a currency value is its values 10 minutes ago.

	no. obs.	AIC	BIC
EUR/USD	89389	-1174842.737	-1174814.535
GBP/USD	89419	-1256379.221	-1256351.017
CHF/USD	89419	-998922.197	-998893.994
JPY/USD	89419	-306488.525	-306460.322

Table 3.7: The minimum AIC and BIC from AR models.- Both criteria yield a 1- period lag as optimum.

In order to test against the forecast results against the results from this model, a linear model which includes 1-period lagged value of the currency itself and the 9 topics as 8 dummies are fitted as follows:

$$z_t = \alpha_0 + \alpha_1 z_{t-k} + \alpha_2 D_1 + ... + \alpha_9 D_8 + \epsilon_{z,t} \tag{3.11}$$

where $\{\alpha\}$ are the model coefficients, $\{k\}$'s are the lags, $\{D\}$'s are the dummies for the topics and $\{\epsilon\}$ are white noises.

Secondly, following Si et. al (2013), after clustering the tweets, a VAR model is built for the tweet series $\{x_t\}$ and each exchange rate series $\{z_t\}$:

$$x_t = \nu_{11} x_{t-1} + \nu_{12} z_{t-1} + \epsilon_{x,t}$$

$$z_t = \nu_{21} z_{t-1} + \nu_{22} x_{t-1} + \epsilon_{z,t} \tag{3.12}$$

where $\{\nu\}$ are the model coefficients and $\{\epsilon\}$ are white noises.

From the timestamps with multiple trending topics, the ones with the most word count is taken to increase the information that can be obtained from the tweets data. This reduced dataset includes 89428 entries with 10 minute intervals from July 2013 to June 2015. Before fitting the VAR, the series are divided into training and test subset in order to test the predictive power of the models. 70 % of the observations are taken as training set (from July 2013 to November 2014), and VAR is trained on this set. The remaining 30% (from December 2014 to June 2015) is used to test the forecast. Table-3.8 shows the results from AIC and BIC. While according to BIC the optimal number of lags is 12 for each currency, according to AIC the optimal lags are 29, 29, 28 and 19 for Euros, British Pound, Swiss Franc and Japanese Yen respectively.

	Lags from AIC	Lags from BIC	AIC	BIC
EUR/USD	29	12	-15.38	-15.37
GBP/USD	29	12	-15.81	-15.80
CHF/USD	28	12	-14.72	-14.71
JPY/USD	19	12	-5.056	-5.048

Table 3.8: The minimum AIC and BIC from VAR models.- Each criteria yields different number of lags as optimum.

Next section compares the prediction errors from these 4 models.

3.5 Results

The prediction errors from the AR(1) models during the test period is given in Figure-3.5. Prediction errors are not too far from zero most of the time, although they are either below or above zero for long intervals instead of fluctuating around zero.

One important exception to this is for Swiss Franc on 1/15/2015, where the big spike in the errors is observed (which is also observed in the currency value in Figure-1). This is when the Swiss Central Bank announced after an unscheduled meeting that it would discontinue keeping the exchange rate at a minimum of 1.20 CHF per Euro, and that it is lowering the interest rate to -0.75%, and changing the target range for the three month Libor further from the $(-0.75\% - 0.25\%)$ interval to to $(-1.25\%, -0.25\%)$ interval.[23]

In order to improve the prediction over AR(1), the topics are fitted as dummy variables. This improved the sum of squared prediction errors, as well as the errors across time. Figure-3.4 and Table-3.9 document these error values.

Prediction errors are smaller compared to those from the AR(1) across time, and they oscillate around zero except for a few observations, and except for CHF on 1/15/2015. There seems to be almost no improvement in terms of prediction error on 1/15/2015 for CHF with the model with topics compared to that from AR(1), although after that day the error falls back to zero; whereas with AR(1), errors stays above zero for two weeks. Figure-3.5 compares the errors from both models.

One of the main reasons why the model with topics could not capture this movement is that the Swiss Central Bank meeting was an unannounced emergency meeting and hence the movement in the currency value was probably simultaneous to if not preceding the reaction

[23] The press release from the Swiss Central Bank can be found at: `https://www.snb.ch/en/mmr/reference/pre_20150115/source/pre_20150115.en.pdf`.

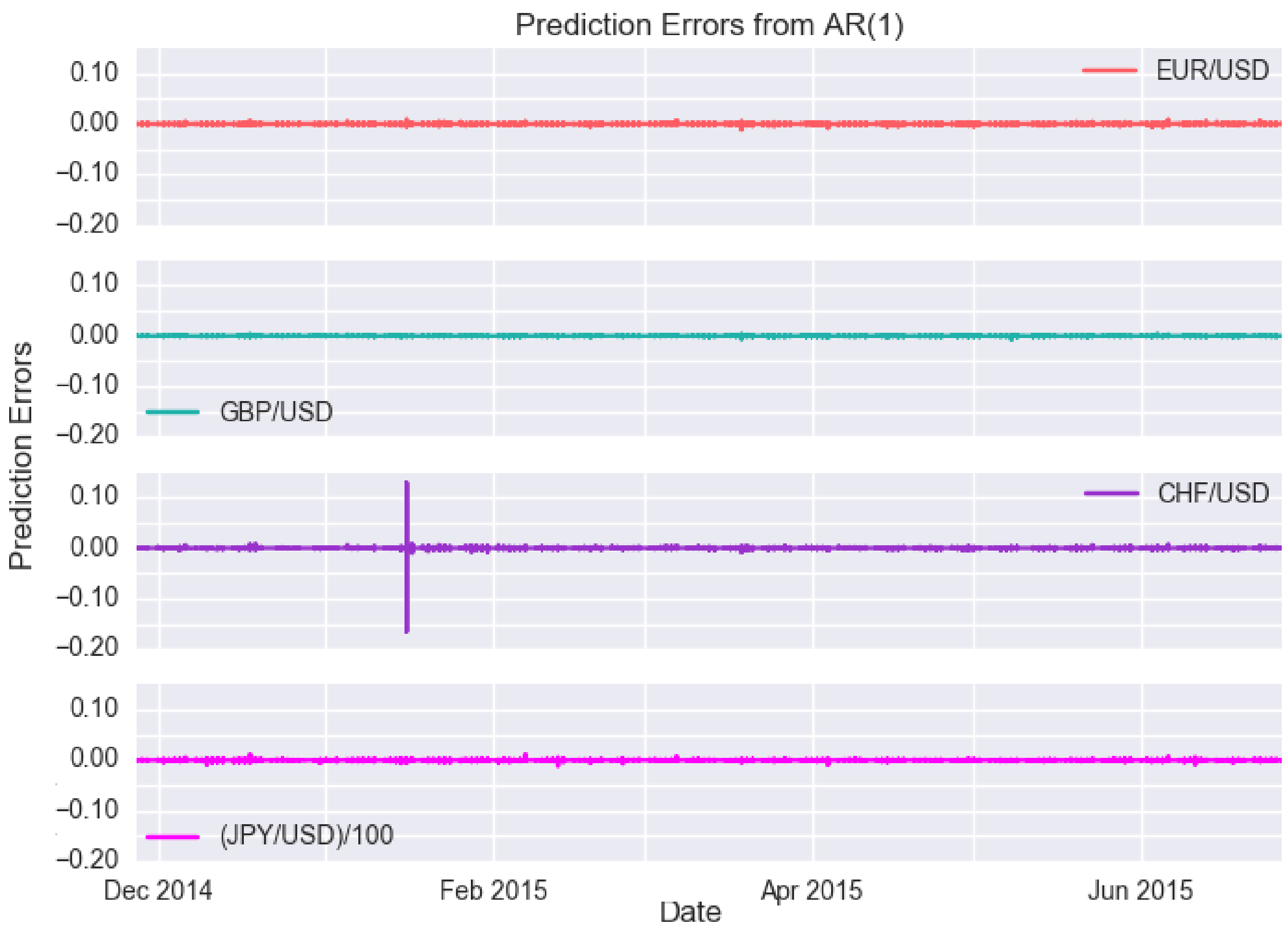


Figure 3.3: Prediction Errors from AR(1) model during test period.- Errors are close to zero but they tend to stay above or below zero instead of oscillating around zero.

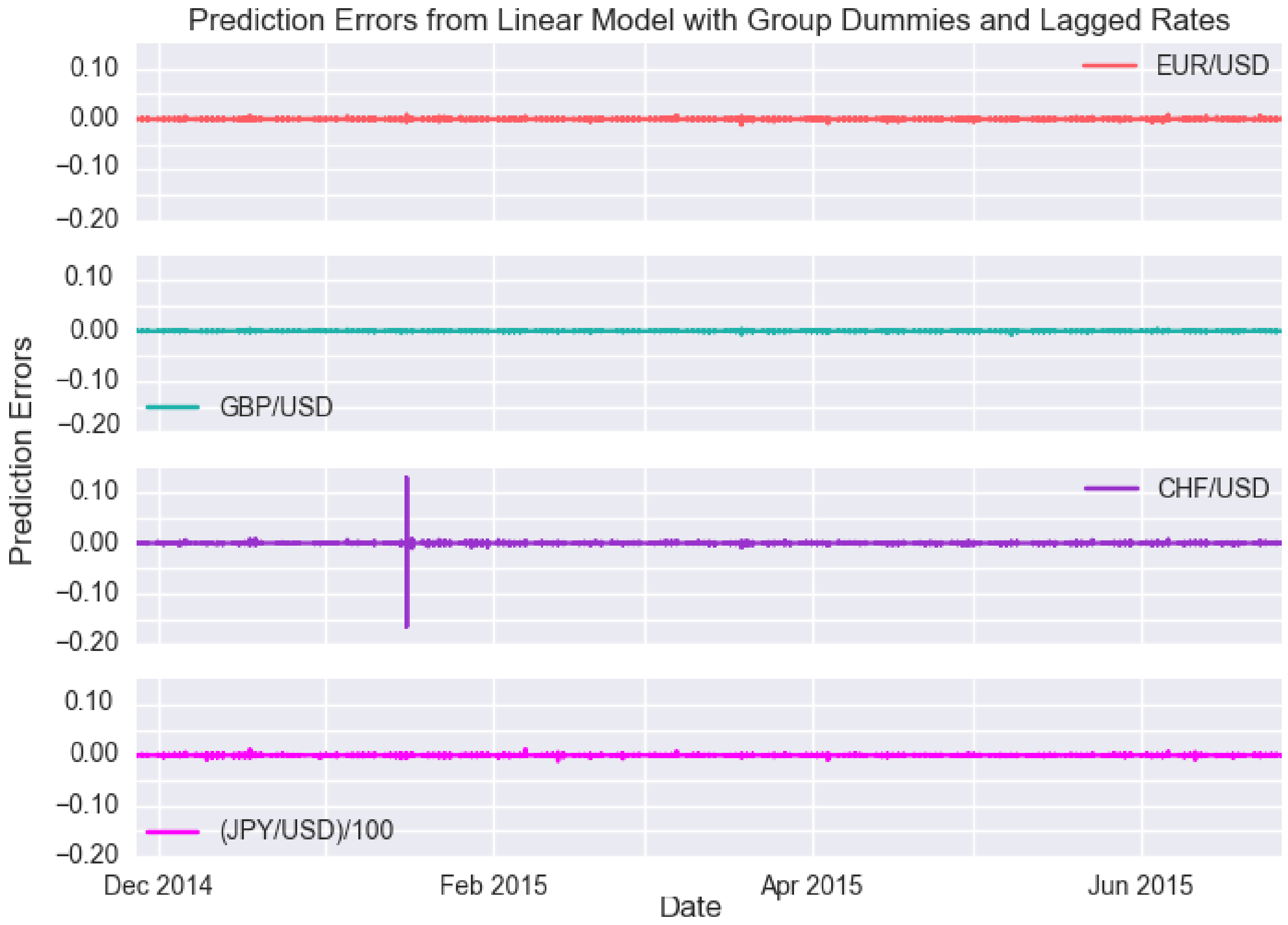


Figure 3.4: Prediction Errors from linear regression during test period with 1-period lagged variables and the topics as dummy variables.- Errors oscillate around zero, except for CHF on 1/15/2015, and a few other observations.

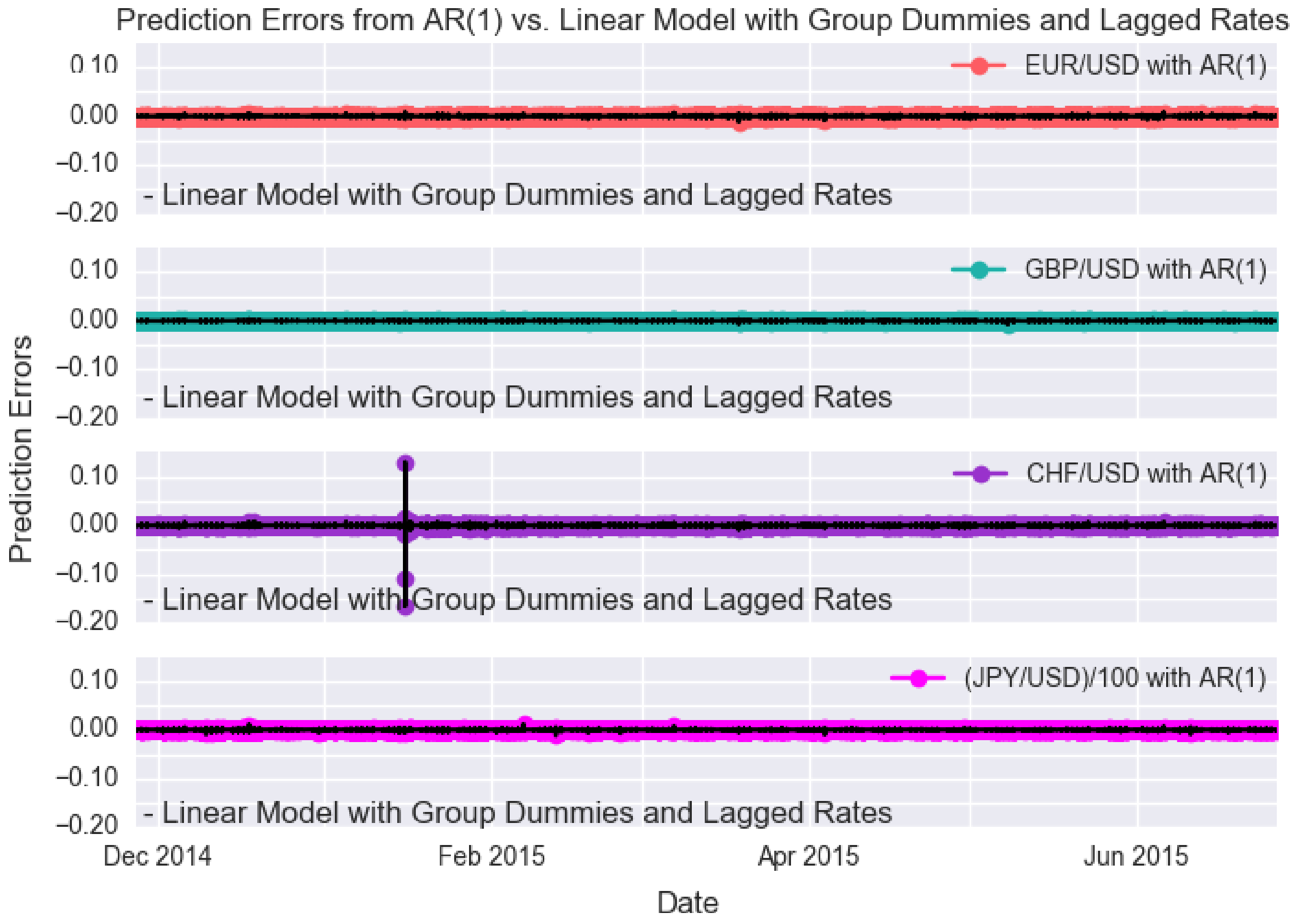


Figure 3.5: Errors from both models.- As can also be seen from Table-3.9, prediction from the model with the factors is much smaller that those from AR(1).

	AR	Linear Regression
EUR/USD	0.0109	0.0069
GBP/USD	0.0022	0.0020
CHF/USD	0.0674	0.0673
JPY/USD	76.4965	70.7816

Table 3.9: Sum of squared errors during the test period.- Sum of squared prediction errors are smaller when topic dummies are added to the model than in AR(1).

in the social media.[24] Although the model with the topics could not predict this surprise event, it was still successful to predict the value of the currency right after the event, which AR(1) failed to predict for two weeks.

In addition to the linear model with topics with dummies, VAR models (depending on AIC and BIC) are fitted in order to capture the effects of topics on exchange rates as well as the effects of rates on topics. Figures 3.6 and 3.7, and Table-3.10 show the prediction errors from these models.

	VAR (AIC)	VAR (BIC)
EUR/USD	93.9337	92.5235
GBP/USD	45.9247	45.9361
CHF/USD	53.1797	53.1658
JPY/USD	58258314.8725	55696777.3534

Table 3.10: Sum of squared errors from VAR's during the test period.- Sum of squared prediction errors are slightly smaller when lag=12 is picked following BIC.

[24] This event was trending in Twitter with the hashtag "#Francogeddon". This hashtag does not appear in the data set used here, probably because it was not trending in the world.

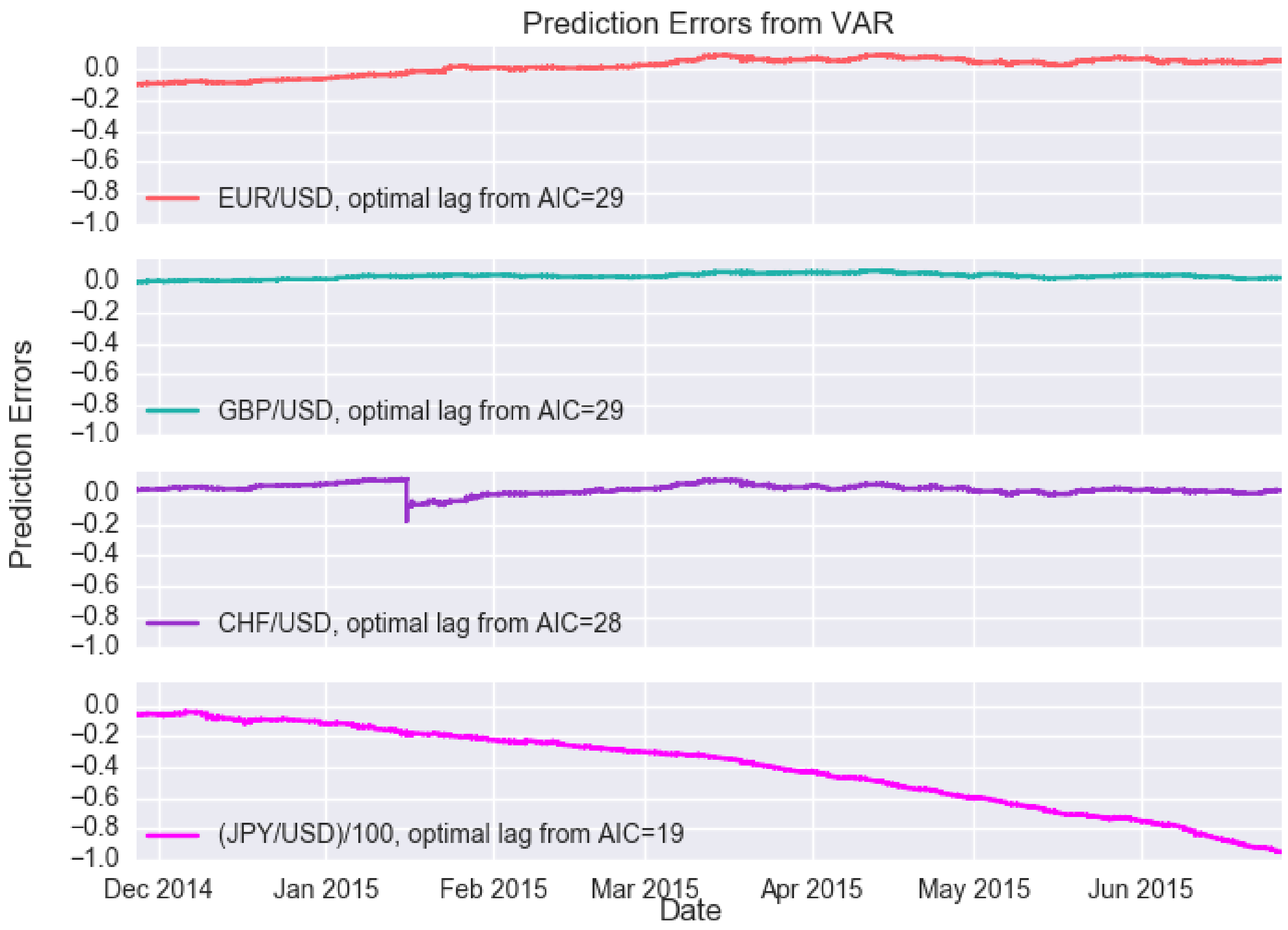


Figure 3.6: Prediction Errors from VAR following AIC criterion.- VAR does not improve the prediction over AR(1).

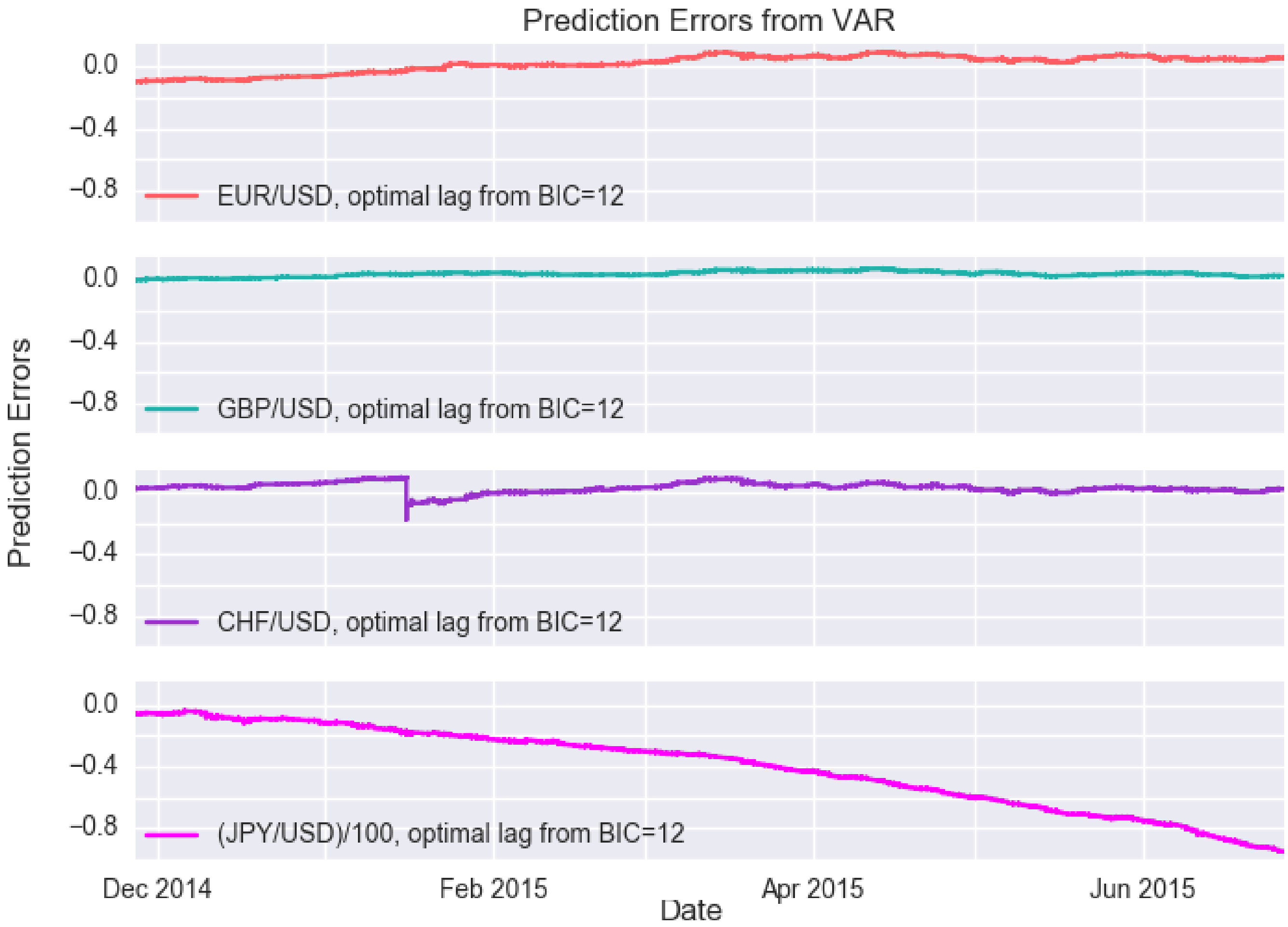


Figure 3.7: Prediction Errors from VAR following BIC criterion.- VAR does not improve the prediction over AR(1).

Compared to AR(1), or the linear model with dummies and lagged currency values, on average VAR models do not result in better prediction. However VAR models result in a better prediction for the Swiss Franc after the announcement shock. This shows the effect of sentiment from the topics on the currency value. However, the overall discrepancy between the results from AR and VAR on average could be attributed to the challenges from the quality of the Twitter dataset.

3.6 Conclusion

This chapter investigated whether incorporating sentiment extracted from Twitter's trending topics would improve the intra-day exchange rate predictions. What makes this paper unique is that unlike previous similar studies which only consider tweets that contain the symbol or the name of the currency or stock, it looks at all trending topics irrespective of whether they contain the name or the symbol of the currency. This allows to capture the general sentiment among the users, which implicitly affects the exchange rates.

One of the important results from this study is that although this approach cannot predict surprise events, since events need to be heard by the users before a user sentiment can be obtained, it performs better than the benchmark AR(1) model when forecasting what will happen after such unexpected events. Moreover, when the topics are added as dummies to the lagged values of the currencies, the prediction errors are found to be lower than the ones from using only the lagged values AR(1). This shows the implicit effect of the sentiment among users on the currency values.

VAR models performed better after the unexpected shock to Swiss Franc, but they did not result in a improved forecasts compared to the benchmark AR(1) on average. The reason why VAR's could not over-perform AR(1) on average could be due to the choice of the data set. The fact that the trending topics can only include an implicit sentiment about the currencies worked well for the linear model; however for VAR to work it might require to use tweets that explicitly mention the currencies, or the countries in which those currencies are used. In fact, only using the tweets that explicitly mention the currencies is the general approach taken in most of all previous studies. However since that approach ignores the sentiments that might implicitly appear in the tweets, it was not preferred in this chapter.

A further step that can be taken after this chapter would be to expand the dataset to include the whole tweets instead of only the trending topics as using only trending topics had many limitations. First of all, the trending topics are too short in length, almost always shorter than the 140 character limit imposed by Twitter. This makes it difficult to extract the extract sentiment from those short expressions. Moreover, most clustering algorithms such as LDA do not work for such short text. Secondly, the trending topics could be in any language, which needed to be detected by the algorithm. Although the language detection algorithm is a powerful one, due to typos, extra or missing characters or numbers it might not have worked as well as desired. Moreover, some of the expressions with missing or extra characters, although the algorithm corrected for the ones that it could detect, were still present in the data set after preprocessing, which also made sentiment extraction and clustering more difficult.

Overall, despite the linguistic limitations from the Twitter's trending topics, this chapter successfully showed that by incorporating sentiment from trending topics intra-day exchange rate forecasts can perform better than the forecasts from AR(1). The next step would be to improve the forecasts from VAR. Using longer tweets or taking the whole tweets which contain a trending topic would be a good starting point for this.

Chapter 4

Sentiment-Based Overlapping Community Discovery

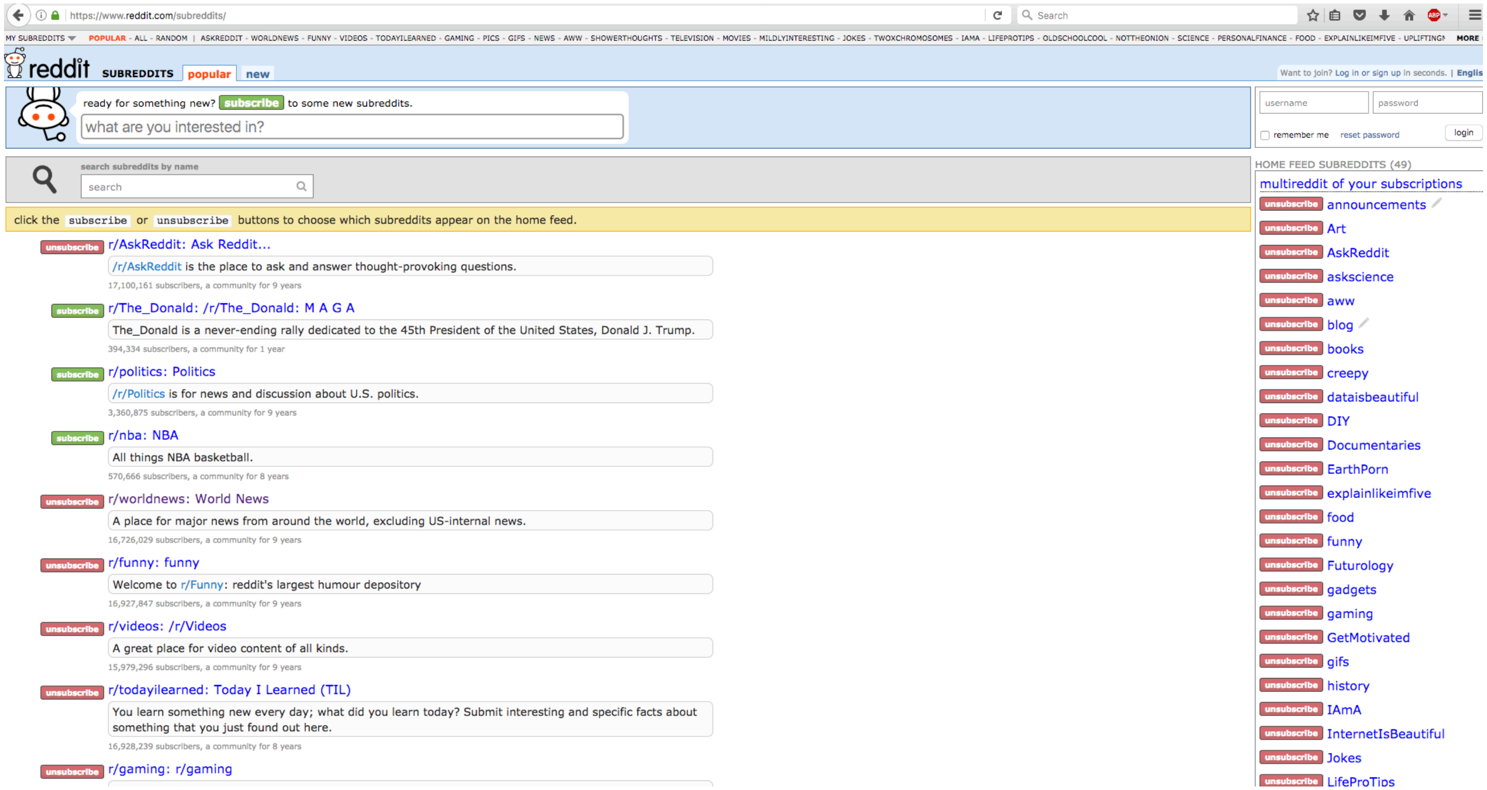
This chapter investigates the behavior of Reddit's news subreddit users and the relationship between their sentiment on exchange rates. Using graphical models and natural language processing, hidden online communities among Reddit users are discovered. The data used in this project are a mixture of text and categorical data from Reddit's news subreddit. It includes the titles of the news pages, as well as a few user characteristics, in addition to users' comments. This dataset is an excellent resource to study user reaction to news since their comments are directly linked to the webpage contents. The model considered in this chapter is a hierarchical mixture model which is a generative model that detects overlapping networks using the sentiment from the user generated content. The advantage of this model is that the communities (or groups) are assumed to follow a Chinese restaurant process, and therefore it can automatically detect and cluster the communities. The hidden variables and the hyperparameters for this model are obtained using Gibbs sampling.

4.1 Background

Detecting user networks is important in understanding the structure and the functions of the communities. This is especially useful for online communities, where discovering clusters of users with high accuracy result in better ad-targeting or improved user behavior prediction. There are many studies that attempt to identify the underlying network structures in communities, most of which use social network datasets, where the links between the users can be easily extracted through various metrics. This study will be the first one that will cluster the users into communities with no pre-defined links between them, by integrating a topic clustering and opinion extraction algorithm into community detection. This paper utilizes a mixture of text and categorical data from Reddit to discover hidden communities among users and extract opinions from these communities to analyze their effect on exchange rate prediction. In particular, the aim is to identify the clusters of users in the news subreddit, using graphical models and natural language processing. [1]

Reddit is an online community consisting of contents that are submitted, voted and discussed by registered users. Areas of interests are divided into subreddits, which are specific forums for relevant shared contents. After a user submits a content such as a news article, a photo, or a video to the relevant subreddit, submissions are discussed as comments and are voted up or down.

[1] `https://about.reddit.com/`.


https://www.reddit.com/subreddits/
Search
MY SUBREDDITS
POPULAR - ALL - RANDOM | ASKREDDIT - WORLDNEWS - FUNNY - VIDEOS - TODAYILEARNED - GAMING - PICS - GIFS - NEWS - AWW - SHOWERTHOUGHTS - TELEVISION - MOVIES - MILDLYINTERESTING - JOKES - TWOXCHROMOSOMES - IAMA - LIFEPROTIPS - OLDSCHOOLCOOL - NOTTHEONION - SCIENCE - PERSONALFINANCE - FOOD - EXPLAINLIKEIMFIVE - UPLIFTING
MORE
reddit
SUBREDDITS
popular
new
Want to join? Log in or sign up in seconds. | Englis
username
password
□ remember me
reset password
login
search subreddits by name
search

ready for something new? subscribe to some new subreddits.

what are you interested in?

click the `subscribe` or `unsubscribe` buttons to choose which subreddits appear on the home feed.

unsubscribe r/AskReddit: Ask Reddit...

/r/AskReddit is the place to ask and answer thought-provoking questions.

17,100,161 subscribers, a community for 9 years

subscribe r/The_Donald: /r/The_Donald: M A G A

The_Donald is a never-ending rally dedicated to the 45th President of the United States, Donald J. Trump.

394,334 subscribers, a community for 1 year

subscribe r/politics: Politics

/r/Politics is for news and discussion about U.S. politics.

3,360,875 subscribers, a community for 9 years

subscribe r/nba: NBA

All things NBA basketball.

570,666 subscribers, a community for 8 years

unsubscribe r/worldnews: World News

A place for major news from around the world, excluding US-internal news.

16,726,029 subscribers, a community for 9 years

unsubscribe r/funny: funny

Welcome to r/Funny: reddit's largest humour depository

16,927,847 subscribers, a community for 9 years

unsubscribe r/videos: /r/Videos

A great place for video content of all kinds.

15,979,296 subscribers, a community for 9 years

unsubscribe r/todayilearned: Today I Learned (TIL)

You learn something new every day; what did you learn today? Submit interesting and specific facts about something that you just found out here.

16,928,239 subscribers, a community for 8 years

unsubscribe r/gaming: r/gaming

HOME FEED SUBREDDITS (49)

multireddit of your subscriptions

- unsubscribe announcements
- unsubscribe Art
- unsubscribe AskReddit
- unsubscribe askscience
- unsubscribe aww
- unsubscribe blog
- unsubscribe books
- unsubscribe creepy
- unsubscribe dataisbeautiful
- unsubscribe DIY
- unsubscribe Documentaries
- unsubscribe EarthPorn
- unsubscribe explainlikeimfive
- unsubscribe food
- unsubscribe funny
- unsubscribe Futurology
- unsubscribe gadgets
- unsubscribe gaming
- unsubscribe GetMotivated
- unsubscribe gifs
- unsubscribe history
- unsubscribe IAmA
- unsubscribe InternetIsBeautiful
- unsubscribe Jokes
- unsubscribe LifeProTips

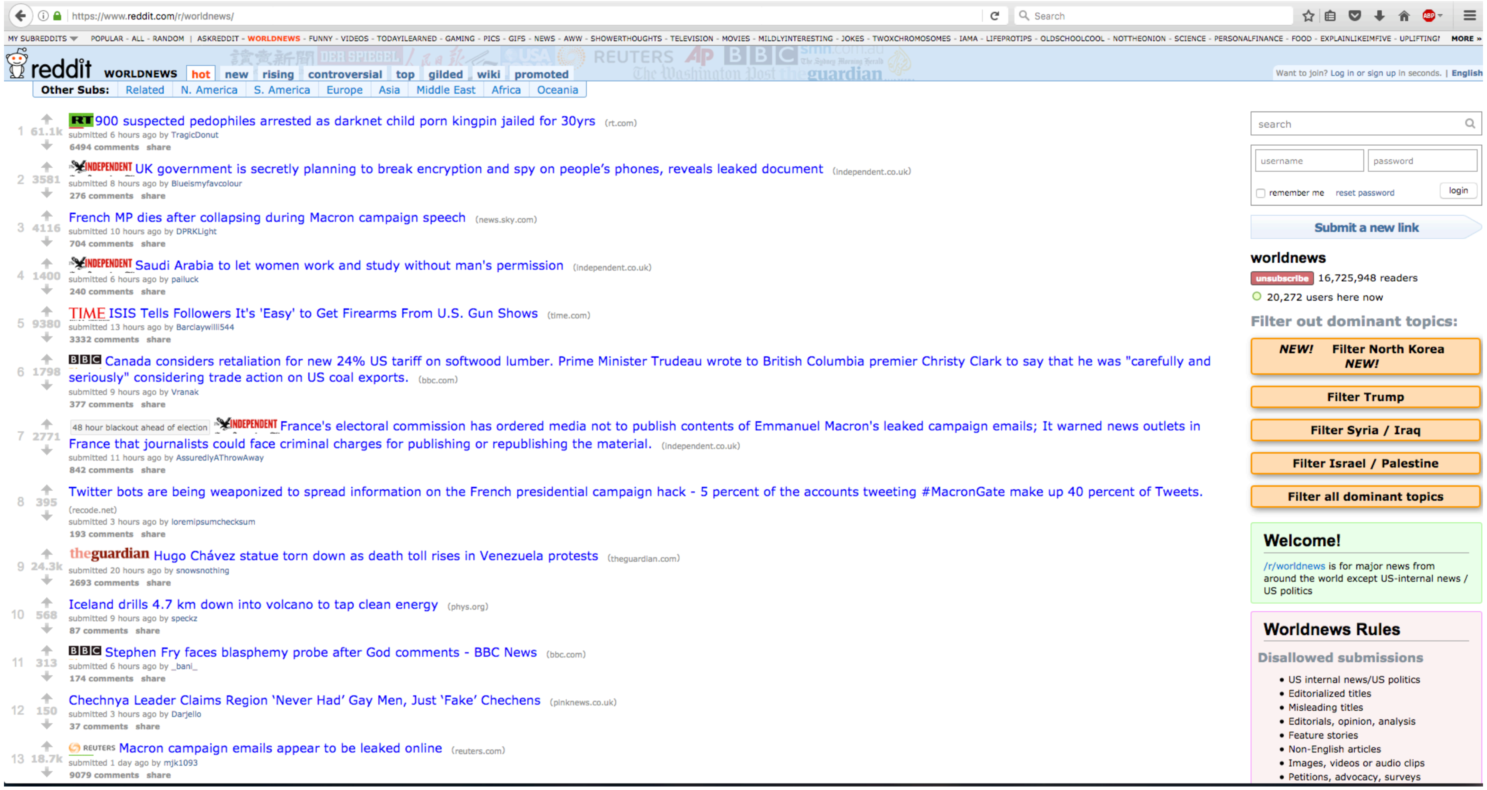
Figure 4.1: Some of the subreddits on Reddit, accessed on 5/6/17.

Within each subreddit, submissions are displayed depending on number of votes they receive; and similarly, comments under each submission are ranked and displayed according the number of votes. There are 250 million registered Reddit users, and 50,000 active Reddit communities.[2] Reddit's wide user base with diverse interests provide an excellence resource for public sentiment. Subreddits that have similar content make up communities; for example subreddits "/r/dogs" and "/r/puppies" can be considered as being under the same community. Unlike subreddits, Reddit communities are not strictly defined as they often overlap with each other. Figure-4.1 shows some of the subreddits available on Reddit.

This chapter is interested in detecting the communities under the "world news subreddit" where users post a URL to a news article as their submission, which get comments from other users.[3] Both the comments and the original submissions receive up or down votes. The reason why this subreddit is chosen is because it contains the headline from the articles on world news, and how people react to that piece of news through comments. This is very important in understanding which communities follow what kind of news (topic-wise) and how their opinions affect short term exchange rates.

Although there are many studies which look into community characteristics of social networks by using comments as links, Xia and Bu (2012) is one of the few studies that analyzes the comment content in social communities. Weninger (2014) is the first study that explores post and comment content in Reddit community. Using a hierarchical LDA model, they find strong evidence comment threads on Reddit exhibit a topical hierarchy. In a following study Weninger (2014) performs another hierarchical LDA analysis on comments and posts in Reddit. In a longitudinal study, Singer et al. (2014) show that the variety of topics has increased in Reddit over the years and it transformed into a "self-referential community from a webpage for content-sharing. These studies concentrate on the content aspect of Reddit, and ignore the community aspect of users who create that content as well as their opinions.

[2] Updated on 4/20/2017. Source: `http://expandedramblings.com/index.php/reddit-stats/`.
[3] `https://www.reddit.com/r/worldnews/`.


https://www.reddit.com/r/worldnews/
Search
MY SUBREDDITS
POPULAR - ALL - RANDOM | ASKREDDIT - WORLDNEWS - FUNNY - VIDEOS - TODAYILEARNED - GAMING - PICS - GIFS - NEWS - AWW - SHOWERTHOUGHTS - TELEVISION - MOVIES - MILDLYINTERESTING - JOKES - TWOXCHROMOSOMES - IAMA - LIFEPROTIPS - OLDSCHOOLCOOL - NOTTHEONION - SCIENCE - PERSONALFINANCE - FOOD - EXPLAINLIKEIMFIVE - UPLIFTING! MORE »
reddit
WORLDNEWS
hot
new
rising
controversial
top
gilded
wiki
promoted
Want to join? Log in or sign up in seconds. | English
Other Subs:
Related
N. America
S. America
Europe
Asia
Middle East
Africa
Oceania
search
username
password
□ remember me
reset password
login
Submit a new link
unsubscribe
NEW! Filter North Korea NEW!
Filter Trump
Filter Syria / Iraq
Filter Israel / Palestine
Filter all dominant topics

1 61.1k RT 900 suspected pedophiles arrested as darknet child porn kingpin jailed for 30yrs (rt.com)
submitted 6 hours ago by TragicDonut
6494 comments share

2 3581 INDEPENDENT UK government is secretly planning to break encryption and spy on people's phones, reveals leaked document (independent.co.uk)
submitted 8 hours ago by BlueIsmyfavcolour
276 comments share

3 4116 French MP dies after collapsing during Macron campaign speech (news.sky.com)
submitted 10 hours ago by DPRKLight
704 comments share

4 1400 INDEPENDENT Saudi Arabia to let women work and study without man's permission (independent.co.uk)
submitted 6 hours ago by pailuck
240 comments share

5 9380 TIME ISIS Tells Followers It's 'Easy' to Get Firearms From U.S. Gun Shows (time.com)
submitted 13 hours ago by Barclaywilli544
3332 comments share

6 1798 BBC Canada considers retaliation for new 24% US tariff on softwood lumber. Prime Minister Trudeau wrote to British Columbia premier Christy Clark to say that he was "carefully and seriously" considering trade action on US coal exports. (bbc.com)
submitted 9 hours ago by Vranak
377 comments share

7 2771 48 hour blackout ahead of election INDEPENDENT France's electoral commission has ordered media not to publish contents of Emmanuel Macron's leaked campaign emails; It warned news outlets in France that journalists could face criminal charges for publishing or republishing the material. (independent.co.uk)
submitted 11 hours ago by AssuredlyAThrowAway
842 comments share

8 395 Twitter bots are being weaponized to spread information on the French presidential campaign hack - 5 percent of the accounts tweeting #MacronGate make up 40 percent of Tweets. (recode.net)
submitted 3 hours ago by loremipsumchecksum
193 comments share

9 24.3k theguardian Hugo Chávez statue torn down as death toll rises in Venezuela protests (theguardian.com)
submitted 20 hours ago by snowsnothing
2693 comments share

10 568 Iceland drills 4.7 km down into volcano to tap clean energy (phys.org)
submitted 9 hours ago by speckz
87 comments share

11 313 BBC Stephen Fry faces blasphemy probe after God comments - BBC News (bbc.com)
submitted 6 hours ago by _bani_
174 comments share

12 150 Chechnya Leader Claims Region 'Never Had' Gay Men, Just 'Fake' Chechens (pinknews.co.uk)
submitted 3 hours ago by Darjello
37 comments share

13 18.7k REUTERS Macron campaign emails appear to be leaked online (reuters.com)
submitted 1 day ago by mjk1093
9079 comments share

worldnews

16,725,948 readers

20,272 users here now

Filter out dominant topics:

Welcome!

/r/worldnews is for major news from around the world except US-internal news / US politics

Worldnews Rules

Disallowed submissions

- US internal news/US politics
- Editorialized titles
- Misleading titles
- Editorials, opinion, analysis
- Feature stories
- Non-English articles
- Images, videos or audio clips
- Petitions, advocacy, surveys

Figure 4.2: World news subreddit on Reddit, accessed on 5/6/17.

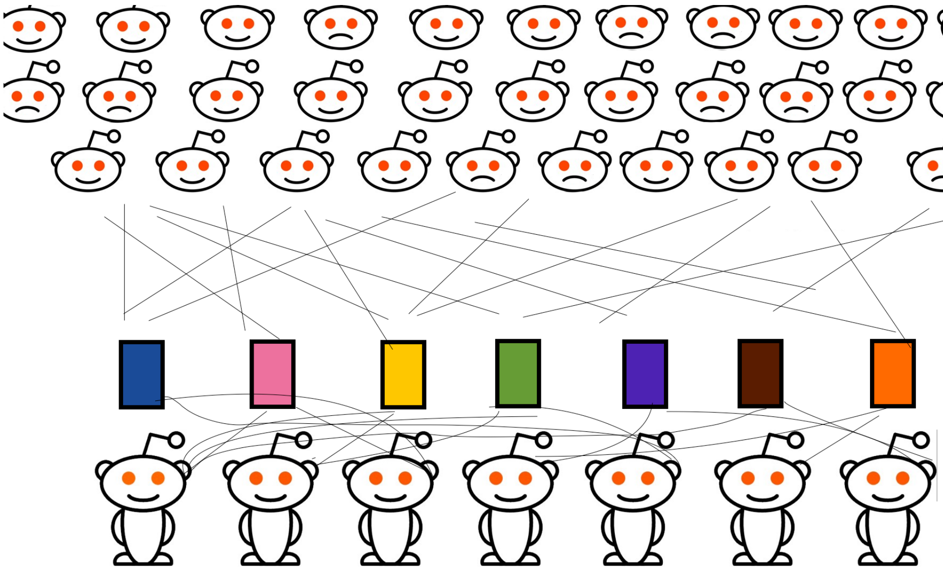


Figure 4.3: Post and comment structure in Reddit.

There are also some other papers that analyze the community structure in microblogs. Laniado et al. (2011) is among the papers that analyze discussion threads in Wikipedia and show that users with many replies interact with inexperienced users, and users who receive comments by many users are more likely to interact with other users. Community detection has been very important for disciplines like biology, sociology and marketing, but it has not received much attention in economics. This paper is also the first one that applies community detection to an economic problem by using the community sentiments in exchange rate prediction. This chapter is organized as follows: Section-II discusses the related literature, Section-III explains the data and methodology, Section-IV presents the topic-based community detection model, Section-V presents the exchange rate prediction model. Results are given in Section-VI and Section-VII concludes.

4.2 Related Literature

This chapter is related to various sub-disciplines such as community detection and topic clustering. Although there are also many studies that apply topic modeling to community detection (topic-based community detection literature), this paper takes it one step further and incorporates opinions into community detection.

4.2.1 Community Detection Literature

This chapter focuses on detecting hidden communities through the content of their submissions and comments, and their sentiment. The earlier line of work in the community detection literature focused on discovering communities. Girvan and Newman (2002) have their famous community detection algorithm which uses centrality indices to find community boundaries, by removing edges from the original graph. Although this is a very famous model, it has its limitations since all communities need to be equal size and all nodes should have the same expected degree (Lancichinetti and Fortunato (2009)). Newman and Girvan (2004) algorithm iteratively removes edges from the network to divide into communities. These methods work well for small communities the magnitude of current networks call for further techniques.

Fortunato (2010) presents an excellent summary of community detection methods. There are the traditional methods such as graph partitioning, hierarchical clustering, partitional clustering and spectral clustering; divisive algorithms such as Girvan and Newman (2002), optimization methods such as greedy techniques, simulated annealing, external optimization and spectral optimization; dynamic algorithms such as spin models, random walk and syncronization; and overlapping community detection methods such as clique percolation and eigenvalue matrix factorization.

Li (2016) combines node attributes and structure of social network without assuming that they have correlation. Newman and Leicht (2007) introduced a model where they take a probabilistic mixture modeling approach and expectation maximization to detect the network structures. Gopalan et al. (2012) develop stochastic optimization algorithms by sampling from usually unconnected nodes. Ruan et al. (2013) combine content and link information in graph structures where they cluster the communities using community discovery algorithms such as Metis and Markov clustering. Tayal et al. (2012) proposes a hierarchical double Dirichlet process mixture model and show that they especially work well with non-stationary time series. Xu et al. (2012) develop a Bayesian probabilistic model for attributed graphs which provides a principled and natural framework for capturing both structural and attribute aspects of a graph. Cheng et al. (2016) propose a model where they combine density and distance, apply a density ordered tree partition problem.

In some networks, communities overlap with each other, whereas in others they are not related. With crisp (non-fuzzy) community assignments, nodes are assigned to only one community whereas fuzzy assignments make it possible for nodes to belong more than one community through a belonging factor (Gregory (2011)). Xie et al. (2013) show that in real networks, nodes belong to 2 or 3 overlapping communities. Anandkumar et al. (2014) develop a tensor approach to detect overlapping hidden communities defined with mixed membership Dirichlet model where they allow fractional membership in multiple communities. They use a scoring algorithm to find the closeness between users based on the semantic similarity in their comments. Wu et al. (2012) start their overlapping community detection algorithm by discovering communities as non-overlapping, and then finding the ones that are related. Yang et al. (2013) develop an overlapping community detection method for networks with node attribute information, based on a generative model for networks with node attributes. Wang et al. (2009) develop an algorithm which detects overlapping communities and especially unstable nodes. Gopalan and Blei (2013) detect overlapping communities by subsampling subgraphs from full graphs and then analyzing them under the current estimate of the communities. They employ a mixed-membership stochastic blockmodel where each node can belong to multiple communities. Yang et al. (2013) develop an algorithm which detects hierarchical communities probabilistically, using quadratic optimization on a random walk based heuristic. Yang and Leskovec (2012) detect overlapping, non-overlapping and hierarchically nested communities.

Another important task in network discovery it to determine the number of underlying communities, which becomes more challenging when the communities are overlapping. Chen et al. (2016)) develop an algorithm to determine the number of nodes as well as the links between them automatically. Chen et al. (2017) propose an algorithm for Bayesian mixture networks through which they can detect overlapping communities in mixing networks. Zhu and Jiang (2016) introduce a community detection algorithm based on random walks and hierarchical Dirichlet process, which allows them to automatically detect the number of communities as well as the communities.

4.2.2 Topic Modeling Literature

First stage in community detection in this chapter is to identify the content of the submissions and the comments, which is done by modeling the topics within submissions, and within comments under submissions.

One of the most commonly used topic clustering algorithm is Latent Dirichlet Allocation (LDA) of Blei et al. (2003). LDA is a generative probabilistic model where documents are random mixtures over latent topics and each topic has a distribution over words. The topic distribution in LDA has a Dirichlet prior. Dirichlet process is a distribution over distributions and is often referred to as the "Chinese restaurant process". Aldous (1985) defines the Chinese restaurant process with a scenario where a fixed number of patrons enter a restaurant with infinitely many tables sequentially. Each customer's seat assignment is proportional to the number of people occupying each table, and more crowded tables have a higher probability of being selected. Chinese restaurant process is being used extensively in topic modeling.

There are many extensions of the Chinese restaurant process that model Griffiths and Ghahramani (2005) extend the Chinese restaurant process to an Indian buffet process where infinitely many dishes are served. A fixed number of customers enter a restaurant and each patron takes a serving from each dish, where their stopping rule is defined by a Poisson process. Thibaux and Jordan (2007) show that Indian buffet process has beta process as the underlying mixing distribution and implement an application on text classification. Jameel et al. (2015) introduce the concept of "buddy customers" to Teh and Jordan (2010)'s Chinese restaurant "franchise model". The Chinese restaurant franchise model allows multiple restaurants to serve the same dishes, so that the factors are shared not only within the groups but also between groups. Introduction of "buddy customers" makes it possible to keep the order of the words and to automatically infer the number of latent topics, which helps discover the n-grams in topics. Blei et al. (2010)'s nested Chinese restaurant process allows "infinitely-deep and infinitely-branching" trees of probability distributions.

Similarly, Blei's LDA is also being extended to model with many different underlying processes. Nallapati et al. (2008) extends Blei et al. (2003)'s LDA framework into Pairwise-Link-LDA and the Link-PLSA-LDA frameworks where they combine LDA and mixed membership stochastic blockmodels, and then combine them into a single graphical model respectively. Wang et al. (2009) develops the Group-Topic Model as a directed graphical model to cluster users by taking their relations and attributes into account. Ahmed and Xing (2008) introduce a temporal aspect to Dirichlet process mixture model to cluster data over each epoch where they assume exchangeability within the same epoch. Sayadi et al. (2015) incorporates a random forest classifier to Latent Dirichlet Allocation to categorize webpages hierarchically.

There are also non-generative models based on word similarities. For example, Burford et al. (2015) analyze the implicit document similarity by basing document links on n-gram overlaps. This paper however is using a generative model (hierarchical Dirichlet process mixture) in order to capture the latent states.

4.2.3 Topic-Based Community Detection Literature

Topic-based community detection models became popular with the availability of personal, user or author data on various platforms. Cucchiarelli et al. (2012) develop a model to detect community members from documents, based on relevant topics, and analyze how networks evolve over time. Ding (2011) applies a topic-based and a topology-based community detection approach to coauthorship networks and show that both approaches result in sub-communities within the detected communities. Duan et al. (2013) develops an algorithm to detect topics and communities within tex-augmented social networks. They model the communities and the topics separately by introducing different latent variables. They use a Dirichlet Process mixture model to for the community and a Hierarchical Dirichlet Process mixture model to for the topics. Ho and Do (2015) discovers communities of users in social network based on the topics and tracks how these communities change over time.

Hoff et al. (2002) develop algorithms where the probability of a relation between actors depends on the positions their social space as an alternative to stochastic blockmodeling approach. Nowicki and Snijders (2001) propose a probabilistic approach to blockmodeling where they assume that the vertices are made up of latent classes and these classes determine the probability distribution on the vertices. Jiang and Zhang (2016) develop an algorithm to automatically infer partitions over nodes in weighted networks without pre-specified number of clusters, using a Dirichlet process prior.

Li et al. (2012) not only identifies communities sharing similar topics, but also keeps track of the temporal dynamics of the communities based on Bernoulli distribution. They introduce a community topic model where they first select community, then topic under community and then word under topic. Under their dynamic model, they introduce the community distributions as Bernoulli trials.

This paper is among the very few studies that analyze Reddit users' behavior. The study of Wang et al. (2016) are among the few papers which analyze Reddit users' behavior over time by learning linguistic characteristics of users via an unsupervised neural model. Zhou et al. (2006) develop algorithms that discover semantic networks in online communities rather than only concentrating on the links. Zhao et al. (2012)takes both the links and the semantics into account in their community detection algorithm.

This paper is also the first to analyze the comment content and incorporate its sentiment into community detection as another layer. One paper that is close to this is that of Bi and Cho (2016), who propose a three layer Dirichlet process hierarchy to model retweet networks in Twitter. They take into account the tweet contents and whether it was retweeted by a user or not while creating the links between users. This paper takes their analysis one step further and takes the comment content into account as well instead of looking at whether a user commented on an entry or not (which would be equivalent to checking if a user retweeted). Incorporating the comment content enables getting a hierarchical semantic relationship from both entries and comments instead of getting the semantic from only the entries and the node links from comments without their semantics.

4.3 Data and Methodology

The data set in this paper comes from Reddit.[4] Reddit is a social news website where users can submit content such as text, photographs or URL's; and comment and vote on these posts. The website is organized into "subreddits" which depend on the categories of the content. Some important subreddit topics are news, politics, and gaming. For the scope of this study, only the entires in the world news subreddit are taken. The comment data are made available by Jason Baumgartner on his website pushshift.io.[5] Only the data which include comments on the website were used in this paper. [6]

This dataset contains all entries in Reddit from December 2005 to date.[7] Size of the data set is approximately 100GB zipped, and this contains all subreddits available. Only the comments under the worldnews subreddit are taken to use in this paper.

Similar to the previous chapter, this data set also required multiple stages of pre-processing. As the first step, the entries whose 'body' or 'author' appear as null or as '[deleted]' are removed from the data set. Although deleting accounts or comments by users themselves or by moderators may indicate that the discussion was controversial and it can carry some information about the thread in which the comments are contained, exploring this aspect was left as a possible extension for this paper. The final form of the data look as follows as given in Table-4.1

author	body	created_utc	id	link_id	name	parent_id
apotre	I appreciate your effort but turkeys leadi...	1370044819	ca9tb2u	t3_1ffo2c	t1_ca9tb2u	t1_ca9t7xr
Badummts	They do. Which channel have you been watching?...	1370044828	ca9tb70	t3_1feee1	t1_ca9tb70	t1_ca9svzi
erdemece	She is in hospital having brain surgery	1370044846	ca9tbd1	t3_1feee1	t1_ca9tbd1	t1_ca9pnha
Phycoz	Double the life of your aircraft wheels with t...	1370044851	ca9tbf4	t3_1ferge	t1_ca9tbf4	None
mxzrxp	had that idea years ago, though it would be pa...	1370044858	ca9tbi5	None	t1_ca9tbi5	t3_1ferge

Table 4.1: First 5 rows of the comments from the worldnews subreddit data.

[4] `https://www.reddit.com/`.

[5] `http://files.pushshift.io/reddit/`.

[6] `https://files.pushshift.io/reddit/comments/`.

[7] The website is being updated regularly and more datasets are uploaded as data become available.

Although there are many other fields available ‘author’, ‘body’, ‘created_utc’, ‘id’,‘link_id’,‘ name’, ‘parent_id’ are chosen as model features due to their relevance. Author corresponds to user name, body is the message content, created_utc shows the time the comment was posted at UTC (in Unix epoch timestamps), id is the identifier and name is the full name of the comment. Link_id is the id of the submission that the comment is in, and parent_id gives the id of comments or submission that are replied for.[8]

Using the unique link_id’s, the world news submissions were scraped by the author using Reddit’s API.[9] The submission_id corresponds to the link_id in the comments (excluding the first 3 characters), and the news headline and epoch timestamp are also taken, as shown in Table-4.2

submission_id	author	created_utc	title
13g5s9	anutensil	1353330303	Spain to Offer Residency to Foreign House Buyers -
14hn05	boyceheult	1354951152	Breast Augmentation
14w7uw	boyceheult	1355581348	Egyptians Vote on Islamist-Backed
157cpx	FreedomsPower	1356057083	UN calls for ban on ’grotesque practice’ of
15ld8w	Ozires	1356731923	BBC News - Delhi gang-rape victi

Table 4.2: First 5 rows of the submissions from the worldnews subreddit data.

Since all posts from Reddit are available, in this chapter the prediction is done per minute intervals. The same exchange rate data per tick for GBP/USD, EUR/USD, JPY/USD, CHF/USD as in previous chapter are used here as well: [10]

[8] For more details on the Reddit fields, see https://github.com/reddit/reddit/wiki/JSON.

[9] ‘praw’ package in Python was used to scrape submissions corresponding to the link_id’s taken from the comments data set. For more detail on praw, see: https://praw.readthedocs.io/en/latest/index.html.

[10] www.forex.co.uk.

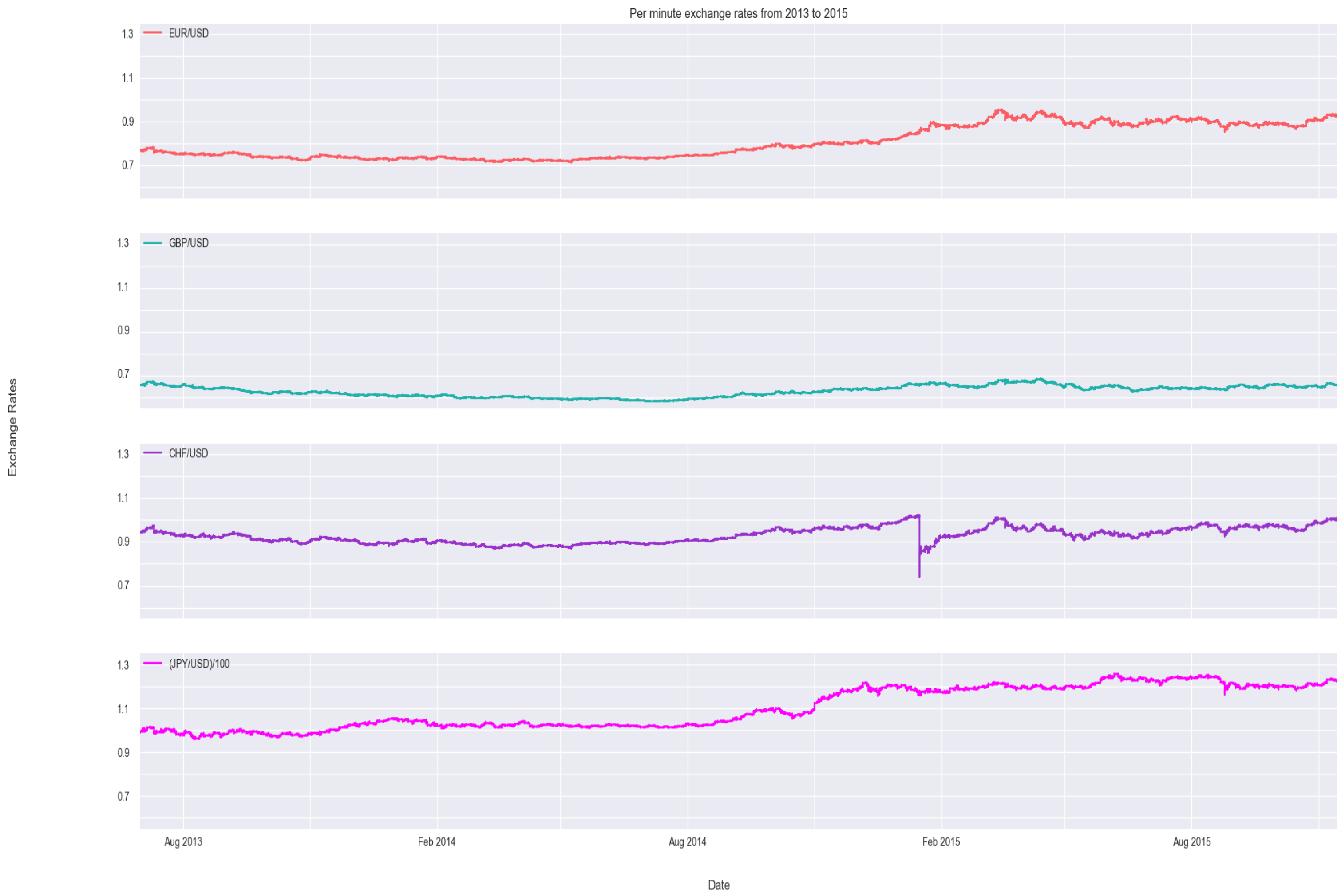


Figure 4.4: Exchange rates from 2013 to 2015.

4.4 Model

Ferguson (1973) defines Dirichlet process as a random process whose sample functions are almost surely probability measures. Sethuraman (1994) represents Dirichlet measure as a stick-breaking process as follows:"

1. The Dirichlet measure is a probability measure on the space of measures.
2. It gives probability one to the subset of discrete probability measures.
3. The posterior distribution is also a Dirichlet measure. "

With Dirichlet process mixture models, the conditional distribution of the random measure is a mixing distribution from a parameter (Antoniak (1974)), which makes them perfect for unsupervised topic modeling. Hierarchical modeling allows parameters to have distributions which can have new parameters (Teh and Jordan (2010)). Since observations arrive in a sequence in time series data, order of arrival matters. Hidden markov models (with discrete latent state) and autoregressive moving average systems (with continuous latent state) can help overcome this limitation with the regular Dirichlet process (Gershman and Blei (2012)).

The model in this paper closely follows from that of Bi and Cho (2016) where they discover communities in Twitter via a three-layer Dirichlet process model, taking the retweeting behavior into account. In the first two layers, they draw a global and then a personal topic distribution for the tweets. Then they draw a probability measure from a set of multiple topic probability measures to find the retweeters of a tweet. The model here takes this one step further and adds a fourth layer, taking their "retweeting" behavior analogous to "commenting" behavior in Reddit. After commenting users are drawn for each post, in the fourth layer of this model their opinions are drawn i.e, whether they are agreeing or disagreeing with the original post.

There are N users and M posts in total. Each user i has a total of M_i number of posts,

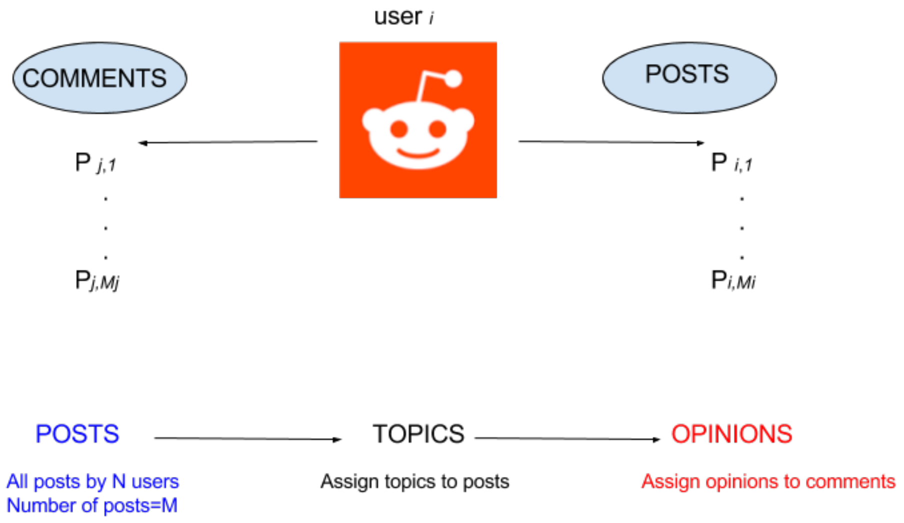


Figure 4.5: How model works.

denoted with the sequence $P_{i,1}, P_{i,2}, .., P_{i,M_i}$, where $\sum_{i \in N} M_i = M$. Each post corresponds to a news headline and assigned a topic l. Each post by user i also receives comments from users $-i$. Users are allowed to post as many comments as they wish, however only the first comment of each user under a post is taken since comments tend to get noisy as users respond to each other. Comments of user i under topic l are aggregated and assigned an opinion. Therefore for each user i, there is a vector of topic assignments P_i determined from their own posts, and a vector of opinion assignment $C_{i,l}$ for each topic l. If a user has not commented on a post from topic l, then for that topic, their opinion vector is null.

Each user and each comment are considered as a unique mixture model since users' post behavior and comment behavior might differ from each other. The global topic distribution (at the population-level) is given as follows:

$$G_0 \sim DP(\alpha_0, H) \tag{4.1}$$

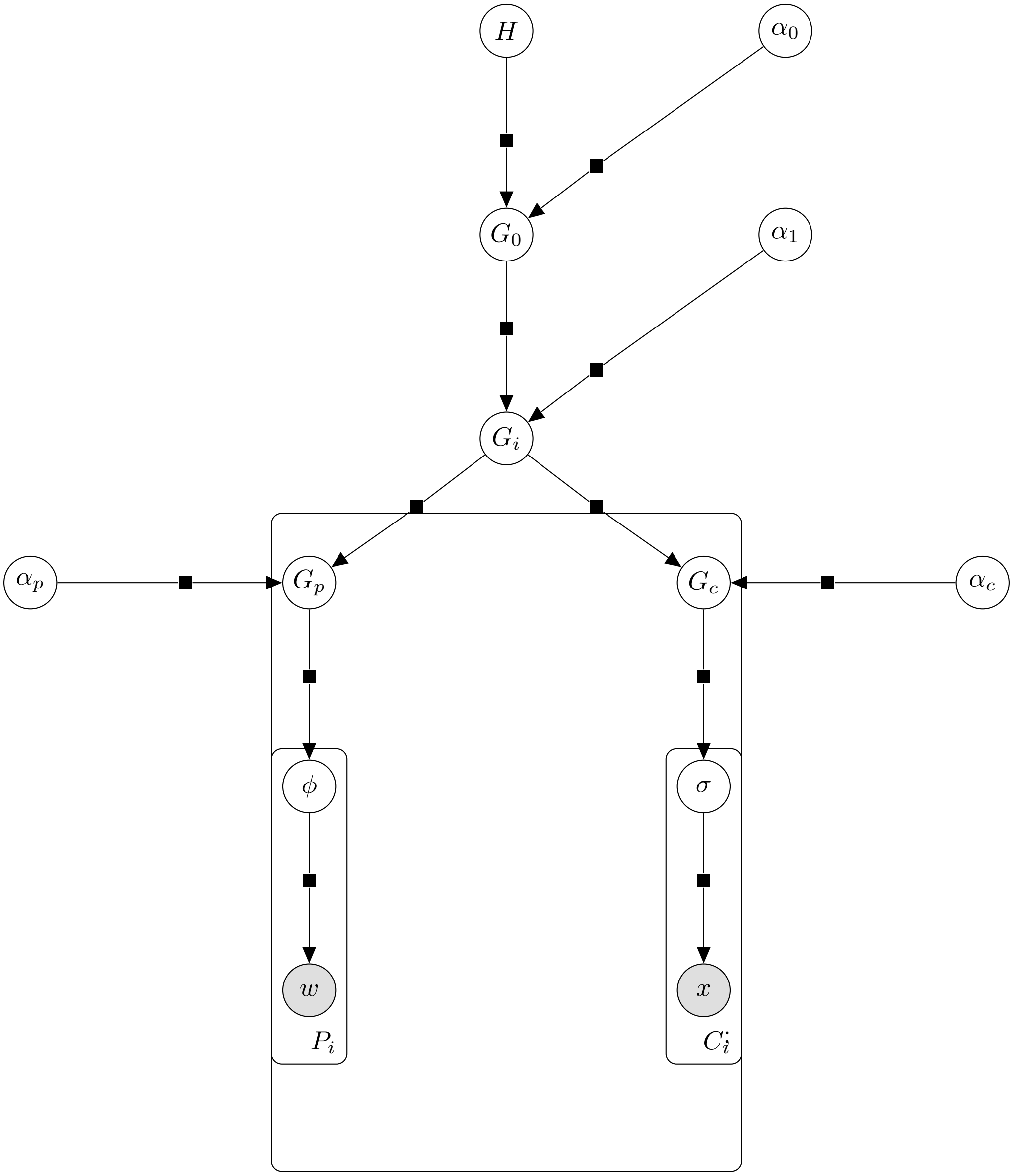


Figure 4.6: HDPMM representation for user i

where H is the top-level base measure and α_0 is the precision parameter. Individual-level topics are distributed as:

$$G_i|G_0 \sim DP(\alpha_1, G_0) \tag{4.2}$$

where $i = 1, .., N$. In terms of stick-breaking measure, $G_i = \sum_{k=1}^{\infty} \pi_{ik}\delta_{\phi_{ik}}$, where $\pi =$

$(\pi_{ik})_{k=1}^{\infty} \sim DP(\alpha_1, G_0)$ and π_i gives the mixing proportions that shows user i's interest in each topic. Next, the post behavior of users is drawn as a probability measure for each individual i:

$$G_p|G_i \sim DP(\alpha_p, G_i) \tag{4.3}$$

In terms of stick-breaking measure, $G_p = \sum_{k=1}^{\infty} \kappa_{ik}\delta_{\phi_{ik}}$, where $\kappa = (\kappa_{ik})_{k=1}^{\infty} \sim DP(\alpha_p, \pi)$ and it shows user i's posting interest in topics. The comment behavior of users is also drawn as a probability measure for each individual i:

$$G_c|G_i \sim DP(\alpha_c, G_i) \tag{4.4}$$

In terms of stick-breaking measure, $G_c = \sum_{\zeta=1}^{\infty} \zeta_{ik}\delta_{\phi_{ik}}$, where $\zeta = (\zeta_{ik})_{k=1}^{\infty} \sim DP(\alpha_c, \pi)$ and it represents user i's commenting interest in topics.

Figure-4.7 shows the stick-breaking process of the model graphically:

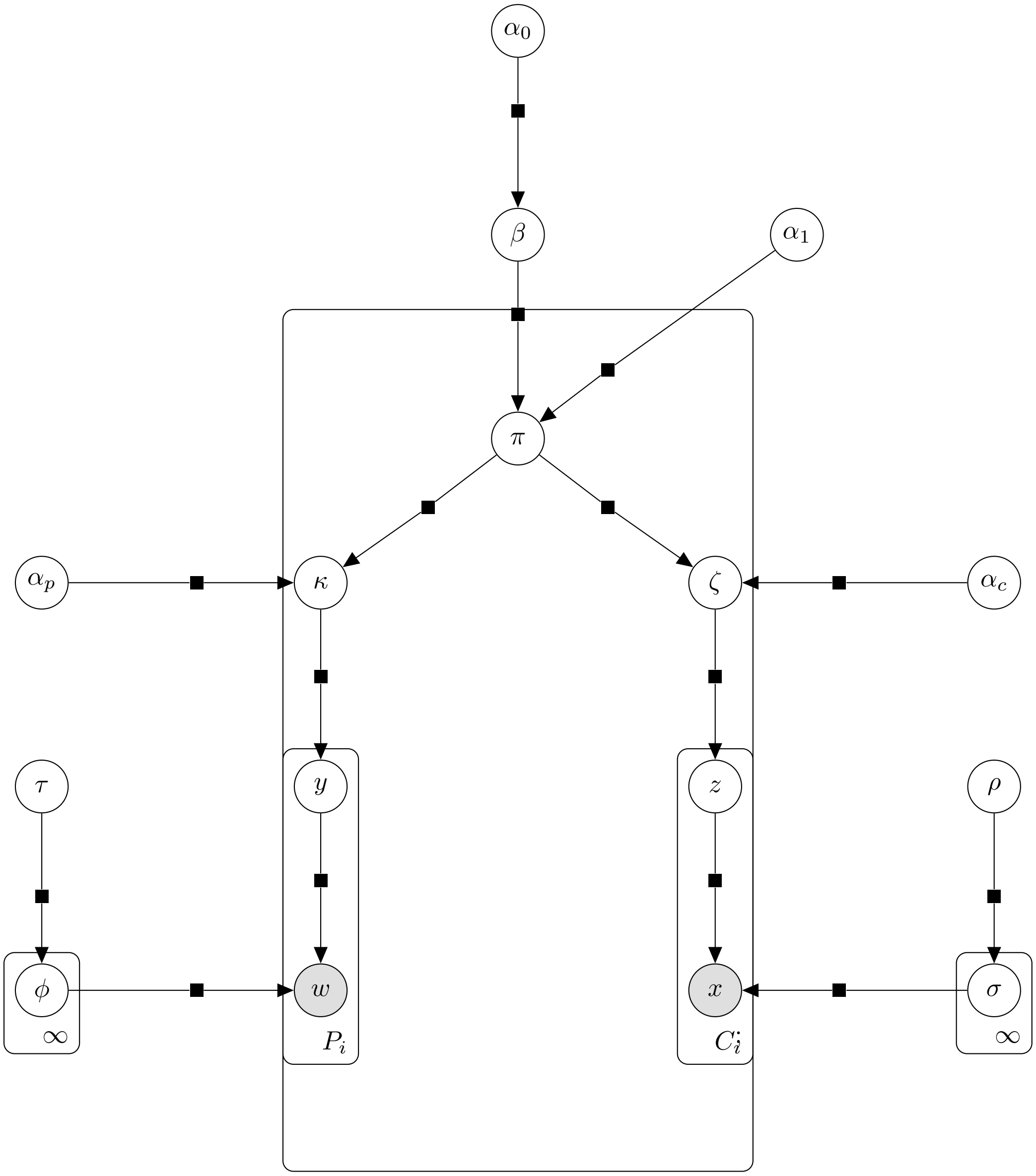


Figure 4.7: Stick-breaking process representation for user i

4.4.1 Model Inference

Posteriors for topic and opinion assignments are obtained via Gibbs sampling.

a^{th} word in user i's post is given by:

$$p(y_{ia} = k|.) \propto \left(s_{ik}^{-ia} + \alpha_p \pi_{ik}\right) \frac{e_{kw_{ia}}^{-ia} + \tau_{w_{ia}}}{e_{k_*}^{-ia} + \tau_*} \tag{4.5}$$

where s_{ik}^{-ia} gives the number of words in user i's posts assigned to topic k and e_{kw}^{-ia} is the number of times word w is assigned to topic k in all posts. A new topic k' is sampled for y_{ia} with probability:

$$p(y_{ia} = k'|.) \propto \frac{\alpha_p \pi_{ik'}}{S} \tag{4.6}$$

where S is the number of all unique words in the corpus.

b^{th} opinion in user i's post is given by:

$$p(z_{ib} = l|.) \propto \left(q_{il}^{-ib} + \alpha_c \pi_{il}\right) \frac{g_{lx_{ib}}^{-ib} + \rho_{x_{ib}}}{q_{l_*}^{-ib} + \rho_*} \tag{4.7}$$

where q_{il}^{-ib} gives the number of words in user i's opinions assigned to topic l and g_{lx}^{-ib} is the number of times opinion x is assigned to topic l in all posts. A new opinion l' is sampled for z_{ib} with probability:

$$p(z_{ib} = k'|.) \propto \frac{\alpha_c \pi_{il'}}{Q} \tag{4.8}$$

where Q is the number of all unique opinions in the corpus.

A latent topic can be given as a distribution over a fixed set of words in the corpus:

$$\phi_{kw} = p\left(w|y=k\right) = \frac{s_{kw} + \tau_w}{\sum_{w=1}^{S}\left(s_{kw} + \tau_w\right)} \tag{4.9}$$

where s_{kw} represents how many times word w is assigned to topic k in all posts. Posterior for a comment on topic l is given by:

$$\sigma_{lx} = p\left(x|z=l\right) = \frac{g_{lx} + \rho_x}{\sum_{x=1}^{Q}\left(g_{lx} + \rho_x\right)} \tag{4.10}$$

where g_{lx} represents how many times opinion x is assigned to topic l in all comments.

Figure-4.8 presents a heatmap of how frequency of importance of topics change over time, using 20 Newsgroups dataset labels.[11] This heatmap shows that the most important world news topic on Reddit is the news on Middle East, followed by politics.

Next subsection explores the community structure (in terms of sentiment) and how it changes over time for certain topics.

[11] This dataset is a collection of news articles classified into evenly across 20 newsgroups. http://qwone.com/~jason/20Newsgroups/.

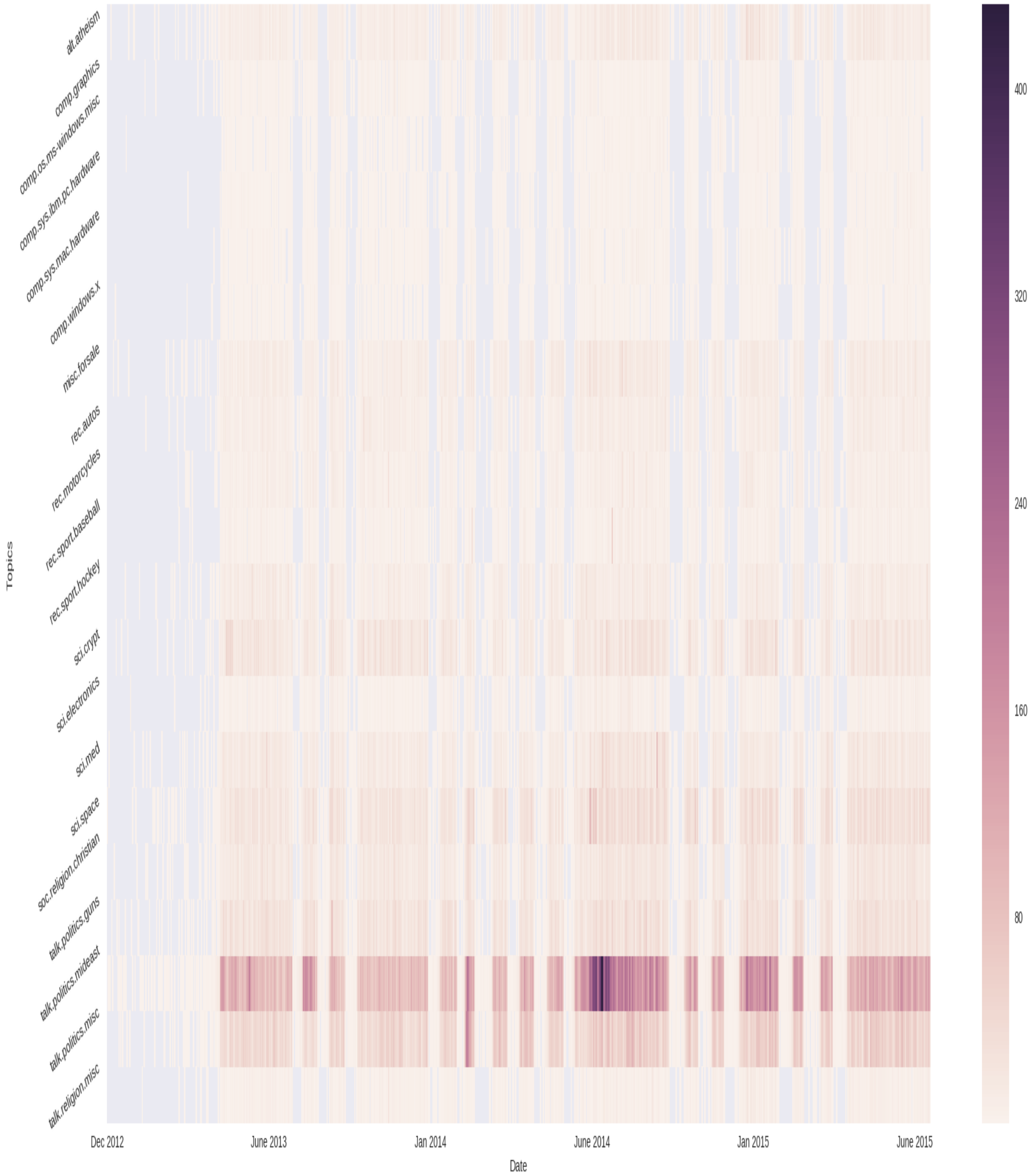


Figure 4.8: Frequency of topics- Heatmap shows the occurrence of each topic group for each timestamp.- Darker colors represent higher frequency.

4.4.2 Networks of users depending on their opinions

The comment networks change across topics and across time. Figure-4.9 and Figure-4.10 show how the sentiment under the Middle East and Religion networks change over time. Networks under other topics are presented in the Appendix.

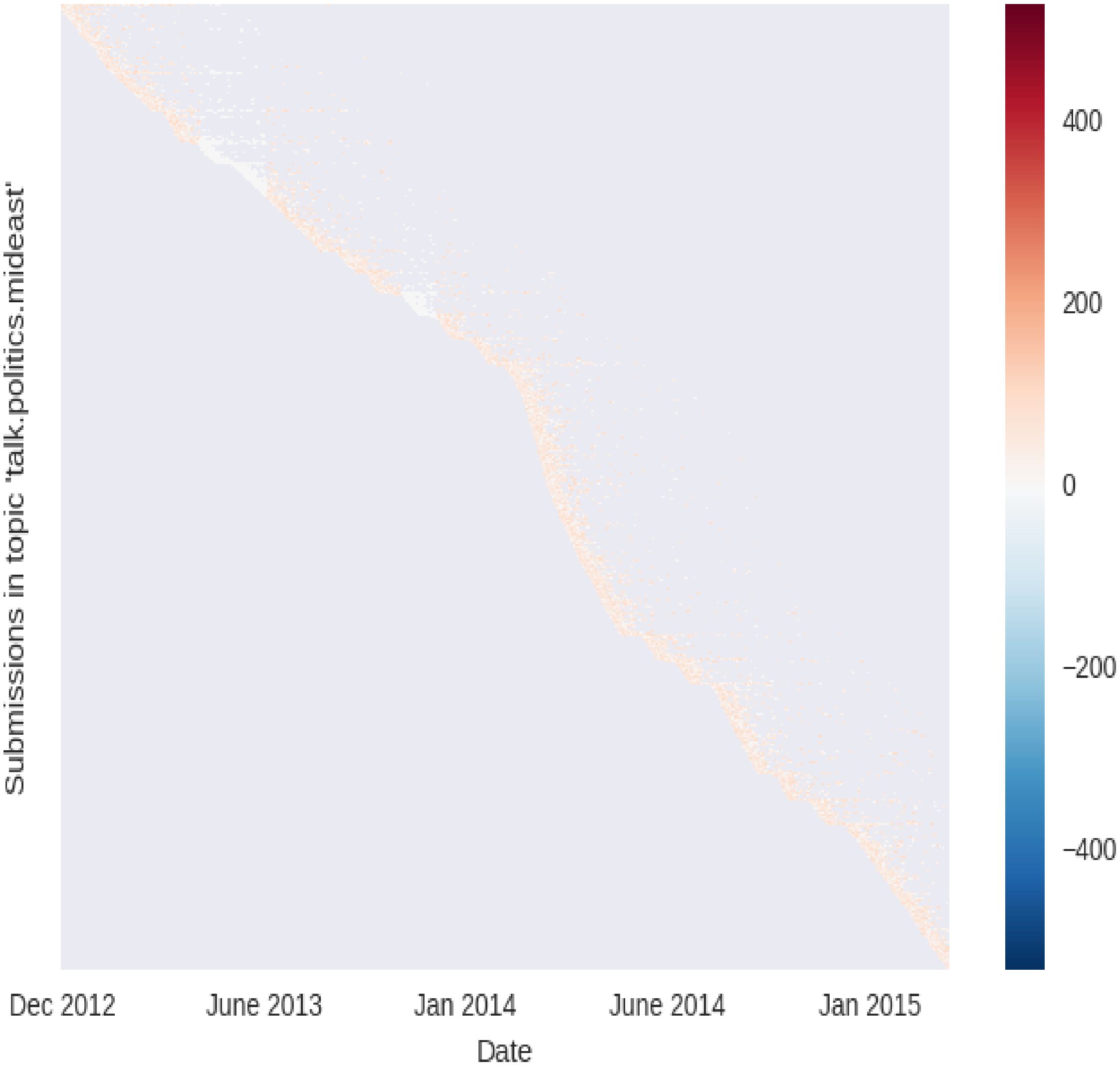


Figure 4.9: How sentiments of clusters change over time under topics related to Middle East- This figure shows how sentiments (color coded) on each submission on Middle East (on y-axis) change across time (on x-axis). Discussions are continued for several months and the sentiments vary a lot.

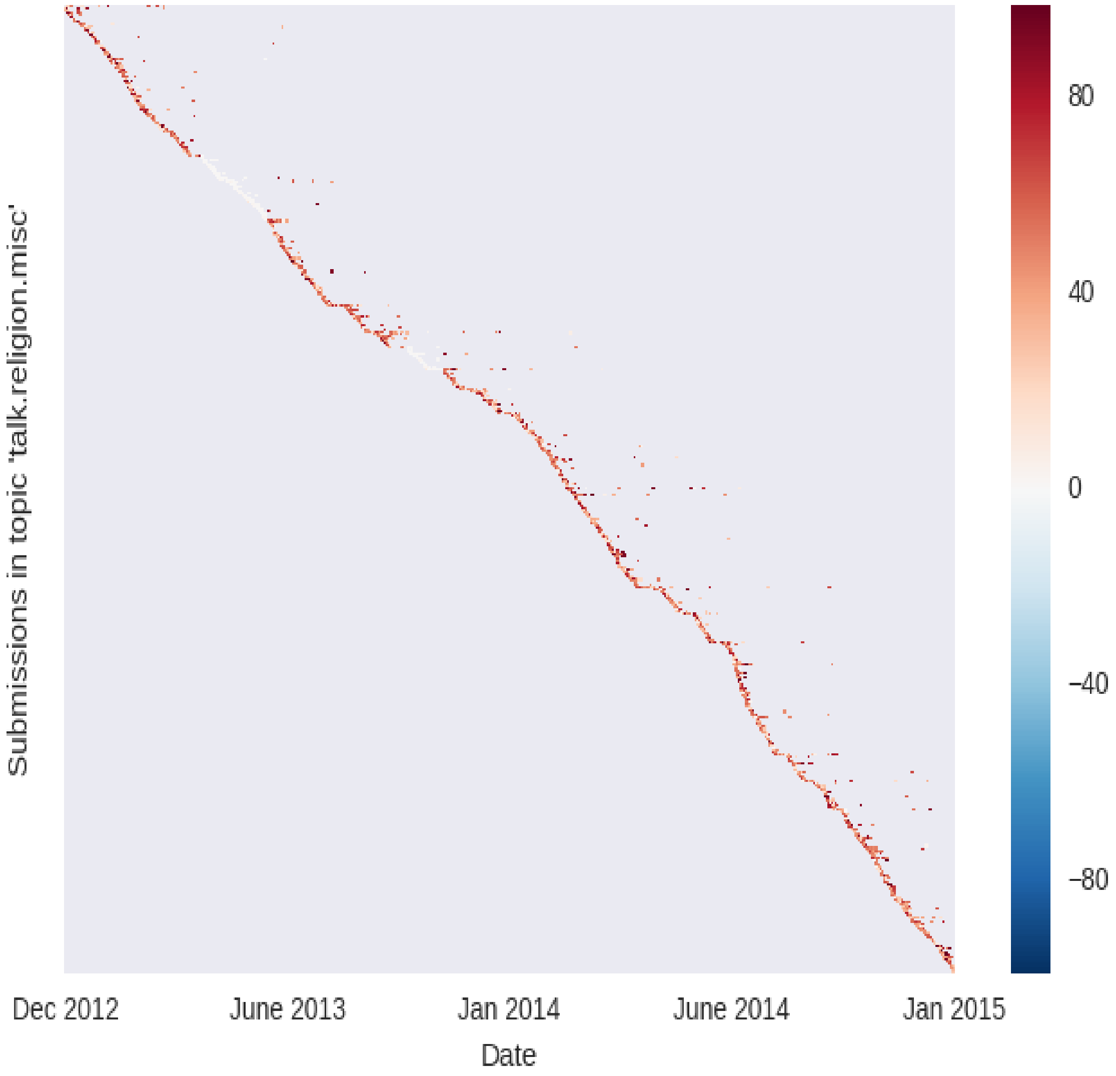


Figure 4.10: How sentiments of clusters change over time under topics related to Religion- This figure shows how sentiments (color coded) on each submission on religion (on y-axis) change across time (on x-axis). Similar to topics related to Middle East, discussions are continued for several months and the there is a higher fluctuation in the sentiments.

4.5 Exchange Rate Prediction

Following the previous chapter, sentiments on topic clusters are added as factors and fitted to the exchange rate series $\{z_t\}$ from July 2013 to June 2015 along with their lagged values:

$$z_t = \alpha_0 + \alpha_1 z_{t-k} + \sum_{k \in K} \sum_{s \in S} \alpha_{k,k_s,t} k_{s_t} + \epsilon_{z,t} \tag{4.11}$$

where $\{\alpha\}$ are the model coefficients, $\{k\}$'s are the topics, $\{s\}$'s sentiment at each topic at each timestamp and $\{\epsilon\}$ are white noises.

4.6 Results

Prediction errors are plotted in Figure-4.11 and their squared sum is given in Table-4.3:

This model shows a huge improvement over the previous chapter, which already had a better prediction than the baseline AR model. Since data is available per minute, spikes after shocks are not as big as in the previous model.

	Errors with sentiment clusters on networks
EUR/USD	0.00139
GBP/USD	0.000194
CHF/USD	0.00145
JPY/USD	10.70421

Table 4.3: Sum of squared errors.

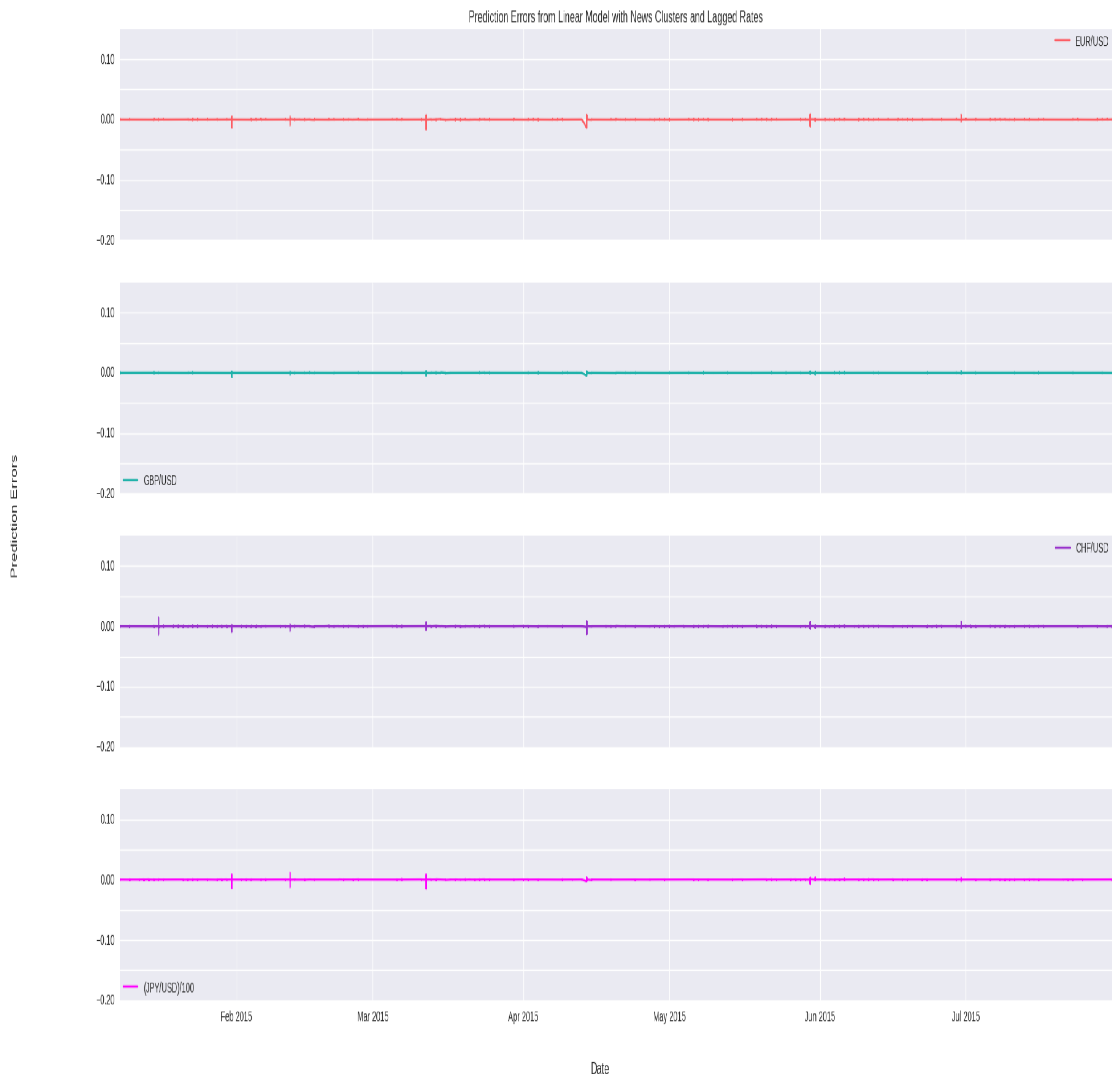


Figure 4.11: Prediction Errors from the model with networks

4.7 Conclusion

This paper is the first application of a community detection problem to an economic problem. By predicting the structure of communities under each topic and then using the community sentiments to predict exchange rates, this chapter showed that incorporating user networks and sentiment extracted from Reddit's worldnews subreddit improves the per minute exchange rate predictions.

This chapter presents a model that works better than the one in the previous chapter and is interpretable. One general conclusion from this dataset is that the news about Middle East and politics are the most important ones in proxying the general opinions.

In conclusion, this chapter provides an improvement in terms of errors and the time period in predicting exchange rates, over the model in the previous chapter, and over the baseline model. This model also successfully shows that as it becomes possible to capture more information from social media, the data start mimicking real world behaviors.

Chapter 5

Concluding Remarks

This thesis shows that capturing those high frequency variables from social media data and using them as a proxy for general sentiment provide great value to short term exchange rate prediction. Social media has the capability to capture ongoing events and how people react to those events at every instance, providing real time feedback with very high frequency. This makes it possible to do predictions for a very short time frame, hence acting as an early signal of where the currency will move given what is going on in the world and how people react.

Incorporating people's reactions as signals of their future behaviors will become important in almost every sector in the near future. As more variety of data (voice, navigation, health, etc.) become available; it will be possible to model, predict and imitate behaviors with very high accuracy within a very short time period. Increased variety and volume of data will introduce new computational challenges, which will attract more resources and more research in this field.

Bibliography

Acharya, V. V. and Pedersen, L. H. (2005). Asset pricing with liquidity risk. *Journal of financial Economics*, 77(2):375–410.

Ahmed, A. and Xing, E. (2008). Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 219–230. SIAM.

Aldous, D. J. (1985). Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII1983*, pages 1–198. Springer.

Amihud, Y. and Mendelson, H. (1980). Dealership market: Market-making with inventory. *Journal of Financial Economics*, 8(1):31–53.

Amihud, Y. and Mendelson, H. (1986). Asset pricing and the bid-ask spread. *Journal of financial Economics*, 17(2):223–249.

Amihud, Y. and Mendelson, H. (1991). Liquidity, maturity, and the yields on us treasury securities. *The Journal of Finance*, 46(4):1411–1425.

Amihud, Y., Mendelson, H., Pedersen, L. H., et al. (2006). Liquidity and asset prices. *Foundations and Trends® in Finance*, 1(4):269–364.

Anandkumar, A., Ge, R., Hsu, D. J., and Kakade, S. M. (2014). A tensor approach to learning mixed membership community models. *Journal of Machine Learning Research*, 15(1):2239–2312.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pages 1152–1174.

Atsalakis, G. S. and Valavanis, K. P. (2009a). Forecasting stock market short-term trends using a neuro-fuzzy based methodology. *Expert Systems with Applications*, 36(7):10696–10707.

Atsalakis, G. S. and Valavanis, K. P. (2009b). Surveying stock market forecasting techniques–part ii: Soft computing methods. *Expert Systems with Applications*, 36(3):5932–5941.

Backus, D. (1984). Empirical models of the exchange rate: Separating the wheat from the chaff. *Canadian Journal of Economics*, pages 824–846.

Bahrepour, M., Akbarzadeh-T, M.-R., Yaghoobi, M., and Naghibi-S, M.-B. (2011). An adaptive ordered fuzzy time series with application to forex. *Expert Systems with Applications*, 38(1):475–485.

Baldwin, T., Cook, P., Lui, M., MacKinlay, A., and Wang, L. (2013). How noisy social media text, how diffrnt social media sources? In *IJCNLP*, pages 356–364.

Beber, A., Brandt, M. W., and Kavajecz, K. A. (2009). Flight-to-quality or flight-to-liquidity? evidence from the euro-area bond market. *Review of Financial Studies*, 22(3):925–957.

Bi, B. and Cho, J. (2016). Modeling a retweet network via an adaptive bayesian approach. In *Proceedings of the 25th International Conference on World Wide Web*, pages 459–469. International World Wide Web Conferences Steering Committee.

Bishop, C. M. (2006). Pattern recognition. *Machine Learning*, 128:1–58.

Blei, D. M., Griffiths, T. L., and Jordan, M. I. (2010). The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):7.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Bollen, J., Mao, H., and Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of computational science*, 2(1):1–8.

Bollen, J. L. and Mao, H. (2013). Predicting economic trends via network communication mood tracking. US Patent 8,380,607.

Branson, W. H., Halttunen, H., and Masson, P. (1977). Exchange rates in the short run: The dollar-dentschemark rate. *European Economic Review*, 10(3):303–324.

Burford, C., Bird, S., and Baldwin, T. (2015). Collective document classification with implicit inter-document semantic relationships. *Lexical and Computational Semantics (* SEM 2015)*, page 106.

Chakravarty, S. and Sarkar, A. (1999). Liquidity in us fixed income markets: A comparison of the bid-ask spread in corporate, government and municipal bond markets.

Chen, R.-R., Chidambaran, N., Imerman, M. B., and Sopranzetti, B. J. (2014). Liquidity, leverage, and lehman: A structural analysis of financial institutions in crisis. *Journal of Banking & Finance*, 45:117–139.

Chen, Y., Wang, X., Bu, J., Tang, B., and Xiang, X. (2016). Network structure exploration in networks with node attributes. *Physica A: Statistical Mechanics and its Applications*, 449:240–253.

Chen, Y., Wang, X., Xiang, X., Tang, B., Chen, Q., Fan, S., and Bu, J. (2017). Overlapping community detection in weighted networks via a bayesian approach. *Physica A: Statistical Mechanics and its Applications*, 468:790–801.

Cheng, Q., Liu, Z., Huang, J., and Cheng, G. (2016). Community detection in hypernetwork via density-ordered tree partition. *Applied Mathematics and Computation*, 276:384–393.

Chowdhury, G. G. (2003). Natural language processing. *Annual review of information science and technology*, 37(1):51–89.

Cooper, W. (1999). The russian financial crisis of 1998: An analysis of trends. In *Causes, and Implications. Report for Congress available at: http://congressionalresearch. com/98-578/document. php.*

Cucchiarelli, A., DAntonio, F., and Velardi, P. (2012). Semantically interconnected social networks. *Social Network Analysis and Mining*, 2(1):69–95.

Davidov, D., Tsur, O., and Rappoport, A. (2010). Semi-supervised recognition of sarcastic sentences in twitter and amazon. In *Proceedings of the fourteenth conference on computational natural language learning*, pages 107–116. Association for Computational Linguistics.

De Finetti, B. (1989). Probabilism. *Erkenntnis*, 31(2):169–223.

De Jong, F. and Driessen, J. (2012). Liquidity risk premia in corporate bond markets. *The Quarterly Journal of Finance*, 2(02):1250006.

Díaz, A. and Escribano, A. (2012). Liquidity life cyclein us treasury bonds.

Ding, Y. (2011). Community detection: topological vs. topical. *Journal of Informetrics*, 5(4):498–514.

Dornbusch, R. (1976). Expectations and exchange rate dynamics. *Journal of political Economy*, 84(6):1161–1176.

Driessen, J. (2005). Is default event risk priced in corporate bonds? *Review of Financial Studies*, 18(1):165–195.

Duan, D., Li, Y., Li, R., Lu, Z., and Wen, A. (2013). Mei: Mutual enhanced infinite community–topic model for analyzing text-augmented social networks. *The Computer Journal*, 56(3):336–354.

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2005). Over-the-counter markets. *Econometrica*, 73(6):1815–1847.

Duffie, D., Gârleanu, N., and Pedersen, L. H. (2007). Valuation in over-the-counter markets. *Review of financial studies*, 20(6):1865–1900.

Eisenstein, J. (2013). What to do about bad language on the internet. In *HLT-NAACL*, pages 359–369.

Evans, C., Pappas, K., and Xhafa, F. (2013). Utilizing artificial neural networks and genetic algorithms to build an algo-trading model for intra-day foreign exchange speculation. *Mathematical and Computer Modelling*, 58(5):1249–1266.

Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

Ferguson, T. S. (1973). A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230.

Feuerriegel, S. and Prendinger, H. (2016). News-based trading strategies. *Decision Support Systems*, 90:65–74.

Fortunato, S. (2010). Community detection in graphs. *Physics reports*, 486(3):75–174.

Garbade, K. D. and Silber, W. L. (1979). Structural organization of secondary markets: Clearing frequency, dealer activity and liquidity risk. *The Journal of Finance*, 34(3):577–593.

Gârleanu, N. (2009). Portfolio choice and pricing in illiquid markets. *Journal of Economic Theory*, 144(2):532–564.

Gencay, R. (1999). Linear, non-linear and essential foreign exchange rate prediction with simple technical trading rules. *Journal of International Economics*, 47(1):91–107.

Geromichalos, A., Herrenbrueck, L., and Salyer, K. (2016). A search-theoretic model of the term premium. *Theoretical Economics*, 11(3):897–935.

Gershman, S. J. and Blei, D. M. (2012). A tutorial on bayesian nonparametric models. *Journal of Mathematical Psychology*, 56(1):1–12.

Girvan, M. and Newman, M. E. (2002). Community structure in social and biological networks. *Proceedings of the national academy of sciences*, 99(12):7821–7826.

Gopalan, P. K. and Blei, D. M. (2013). Efficient discovery of overlapping communities in massive networks. *Proceedings of the National Academy of Sciences*, 110(36):14534–14539.

Gopalan, P. K., Gerrish, S., Freedman, M., Blei, D. M., and Mimno, D. M. (2012). Scalable inference of overlapping communities. In *Advances in Neural Information Processing Systems*, pages 2249–2257.

Gregory, S. (2011). Fuzzy overlapping communities in networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(02):P02017.

Griffiths, T. L. and Ghahramani, Z. (2005). Infinite latent feature models and the indian buffet process. In *NIPS*, volume 18, pages 475–482.

Hadavandi, E., Shavandi, H., and Ghanbari, A. (2010). Integration of genetic fuzzy systems and artificial neural networks for stock price forecasting. *Knowledge-Based Systems*, 23(8):800–808.

Hagenau, M., Liebmann, M., Hedwig, M., and Neumann, D. (2012). Automated news reading: Stock price prediction based on financial news using context-specific features. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 1040–1049. IEEE.

Hann, T. H. and Steurer, E. (1996). Much ado about nothing? exchange rate forecasting: Neural networks vs. linear models using monthly and weekly data. *Neurocomputing*, 10(4):323–339.

He, Z. and Milbradt, K. (2014). Endogenous liquidity and defaultable bonds. *Econometrica*, 82(4):1443–1508.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the american Statistical association*, 97(460):1090–1098.

Hopper, G. P. et al. (1997). What determines the exchange rate: Economic factors or market sentiment. *Business Review*, 5:17–29.

Huang, J.-Z. and Huang, M. (2012). How much of the corporate-treasury yield spread is due to credit risk? *Review of Asset Pricing Studies*, 2(2):153–202.

Jameel, S., Lam, W., and Bing, L. (2015). Nonparametric topic modeling using chinese restaurant franchise with buddy customers. In *European Conference on Information Retrieval*, pages 648–659. Springer.

Janetzko, D. (2014). Using twitter to model the eur/usd exchange rate. *arXiv preprint arXiv:1402.1624*.

Jankowitsch, R., Nashikkar, A., and Subrahmanyam, M. G. (2011). Price dispersion in otc markets: A new measure of liquidity. *Journal of Banking & Finance*, 35(2):343–357.

Jiang, X. and Zhang, W. (2016). Structure learning for weighted networks based on bayesian nonparametric models. *International Journal of Machine Learning and Cybernetics*, 7(3):479–489.

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

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

Kamara, A. (1994). Liquidity, taxes, and short-term treasury yields. *Journal of Financial and Quantitative Analysis*, 29(03):403–417.

Kim, H.-j. and Shin, K.-s. (2007). A hybrid approach based on neural networks and genetic algorithms for detecting temporal patterns in stock markets. *Applied Soft Computing*, 7(2):569–576.

Kuan, C.-M. and Liu, T. (1995). Forecasting exchange rates using feedforward and recurrent neural networks. *Journal of applied econometrics*, 10(4):347–364.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010a). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591–600. ACM.

Lagos, R. (2010). Asset prices and liquidity in an exchange economy. *Journal of Monetary Economics*, 57(8):913–930.

Lagos, R. and Rocheteau, G. (2007). Search in asset markets: Market structure, liquidity, and welfare. *The American Economic Review*, 97(2):198–202.

Lagos, R. and Rocheteau, G. (2009). Liquidity in asset markets with search frictions. *Econometrica*, 77(2):403–426.

Lagos, R., Rocheteau, G., and Weill, P.-O. (2011). Crises and liquidity in over-the-counter markets. *Journal of Economic Theory*, 146(6):2169–2205.

Lancichinetti, A. and Fortunato, S. (2009). Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117.

Laniado, D., Tasso, R., Volkovich, Y., and Kaltenbrunner, A. (2011). When the wikipedians talk: Network and tree structure of wikipedia discussion pages. In *ICWSM*, pages 177–184.

Leinweber, D. and Sisk, J. (2011). Event-driven trading and the new news. *The Journal of Portfolio Management*, 38(1):110–124.

Li, D., Ding, Y., Shuai, X., Bollen, J., Tang, J., Chen, S., Zhu, J., and Rocha, G. (2012). Adding community and dynamic to topic models. *Journal of Informetrics*, 6(2):237–253.

Li, Y. (2016). Community detection with node attributes and its generalization. *arXiv preprint arXiv:1604.03601*.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

Luca, M. and Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, 62(12):3412–3427.

Lui, M. and Baldwin, T. (2012). langid. py: An off-the-shelf language identification tool. In *Proceedings of the ACL 2012 system demonstrations*, pages 25–30. Association for Computational Linguistics.

Meese, R. A. and Rogoff, K. (1983). Empirical exchange rate models of the seventies: Do they fit out of sample? *Journal of international economics*, 14(1-2):3–24.

Mendelson, H. (1982). Market behavior in a clearing house. *Econometrica: Journal of the Econometric Society*, pages 1505–1524.

Minev, M., Schommer, C., and Grammatikos, T. (2012). News and stock markets: A survey on abnormal returns and prediction models. Technical report, Technical Report, UL.

Mitra, L. and Mitra, G. (2011). Applications of news analytics in finance: A review. *The handbook of news analytics in finance*, 596:1.

Mittermayer, M.-A. and Knolmayer, G. F. (2006). Newscats: A news categorization and trading system. In *Data Mining, 2006. ICDM'06. Sixth International Conference on*, pages 1002–1007. Ieee.

Muhammad, A. and King, G. (1997). Foreign exchange market forecasting using evolutionary fuzzy networks. In *Computational Intelligence for Financial Engineering (CIFEr), 1997., Proceedings of the IEEE/IAFE 1997*, pages 213–219. IEEE.

Nallapati, R. M., Ahmed, A., Xing, E. P., and Cohen, W. W. (2008). Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 542–550. ACM.

Nassirtoussi, A. K., Aghabozorgi, S., Wah, T. Y., and Ngo, D. C. L. (2015). Text mining of news-headlines for forex market prediction: A multi-layer dimension reduction algorithm with semantics and sentiment. *Expert Systems with Applications*, 42(1):306–324.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

Newman, M. E. and Girvan, M. (2004). Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113.

Newman, M. E. and Leicht, E. A. (2007). Mixture models and exploratory analysis in networks. *Proceedings of the National Academy of Sciences*, 104(23):9564–9569.

Ni, H. and Yin, H. (2009). Exchange rate prediction using hybrid neural networks and trading indicators. *Neurocomputing*, 72(13):2815–2823.

Nowicki, K. and Snijders, T. A. B. (2001). Estimation and prediction for stochastic blockstructures. *Journal of the American Statistical Association*, 96(455):1077–1087.

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

Pagnotta, E. (2013). Speed, fragmentation, and asset prices.

Palit, A. K. and Popovic, D. (2006). Computational intelligence in time series forecasting.

Papaioannou, P., Russo, L., Papaioannou, G., and Siettos, C. I. (2013). Can social microblogging be used to forecast intraday exchange rates? *Netnomics: Economic Research and Electronic Networking*, 14(1-2):47–68.

Paul, M. J. and Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272.

Peramunetilleke, D. and Wong, R. K. (2002). Currency exchange rate forecasting from news headlines. *Australian Computer Science Communications*, 24(2):131–139.

Praz, R. (2014). Equilibrium asset pricing with both liquid and illiquid markets. *Browser Download This Paper.*

Refenes, A. N., Azema-Barac, M., Chen, L., and Karoussos, S. (1993). Currency exchange rate prediction and neural network design strategies. *Neural Computing & Applications*, 1(1):46–58.

Rehman, M., Khan, G. M., and Mahmud, S. A. (2014). Foreign currency exchange rates prediction using cgp and recurrent neural network. *IERI Procedia*, 10:239–244.

Ritter, A., Clark, S., Etzioni, O., et al. (2011). Named entity recognition in tweets: an experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1524–1534. Association for Computational Linguistics.

Ritter, A., Etzioni, O., et al. (2010). A latent dirichlet allocation method for selectional preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434. Association for Computational Linguistics.

Rocheteau, G. (2009). A monetary approach to asset liquidity.

Rojas, R. (2013). *Neural networks: a systematic introduction.* Springer Science & Business Media.

Ruan, Y., Fuhry, D., and Parthasarathy, S. (2013). Efficient community detection in large networks using content and links. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1089–1098. ACM.

Sarig, O. and Warga, A. (1989). Bond price data and bond market liquidity. *Journal of Financial and Quantitative Analysis*, 24(03):367–378.

Sayadi, K., Bui, Q. V., and Bui, M. (2015). Multilayer classification of web pages using random forest and semi-supervised latent dirichlet allocation. In *Innovations for Community Services (I4CS), 2015 15th International Conference on*, pages 1–7. IEEE.

Sermpinis, G., Dunis, C., Laws, J., and Stasinakis, C. (2012a). Forecasting and trading the eur/usd exchange rate with stochastic neural network combination and time-varying leverage. *Decision Support Systems*, 54(1):316–329.

Sermpinis, G., Laws, J., Karathanasopoulos, A., and Dunis, C. L. (2012b). Forecasting and trading the eur/usd exchange rate with gene expression and psi sigma neural networks. *Expert systems with applications*, 39(10):8865–8877.

Sermpinis, G., Theofilatos, K., Karathanasopoulos, A., Georgopoulos, E. F., and Dunis, C. (2013). Forecasting foreign exchange rates with adaptive neural networks using radial-basis functions and particle swarm optimization. *European Journal of Operational Research*, 225(3):528–540.

Sethuraman, J. (1994). A constructive definition of dirichlet priors. *Statistica sinica*, pages 639–650.

Shirota, Y., Hashimoto, T., and Sakura, T. (2014). Extraction of the financial policy topics by latent dirichlet allocation. In *TENCON 2014-2014 IEEE Region 10 Conference*, pages 1–5. IEEE.

Si, J., Mukherjee, A., Liu, B., Li, Q., Li, H., and Deng, X. (2013). Exploiting topic based twitter sentiment for stock prediction. *ACL (2)*, 2013:24–29.

Singer, P., Flöck, F., Meinhart, C., Zeitfogel, E., and Strohmaier, M. (2014). Evolution of reddit: from the front page of the internet to a self-referential community? In *Proceedings of the 23rd International Conference on World Wide Web*, pages 517–522. ACM.

Sun, Y., Tang, J., Han, J., Gupta, M., and Zhao, B. (2010). Community evolution detection in dynamic heterogeneous information networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 137–146. ACM.

Talebi, H., Hoang, W., and Gavrilova, M. L. (2014). Multi-scale foreign exchange rates ensemble for classification of trends in forex market. *Procedia Computer Science*, 29:2065–2075.

Tayal, A., Poupart, P., and Li, Y. (2012). Hierarchical double dirichlet process mixture of gaussian processes. In *AAAI*.

Teh, Y. W. and Jordan, M. I. (2010). Hierarchical bayesian nonparametric models with applications. *Bayesian nonparametrics*, 1.

Teh, Y. W., Newman, D., and Welling, M. (2006). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. In *NIPS*, volume 6, pages 1378–1385.

Tenti, P. (1996). Forecasting foreign exchange rates using recurrent neural networks. *Applied Artificial Intelligence*, 10(6):567–582.

Thibaux, R. and Jordan, M. I. (2007). Hierarchical beta processes and the indian buffet process. In *AISTATS*, volume 2, pages 564–571.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10(1):178–185.

Vayanos, D. and Wang, T. (2007). Search and endogenous concentration of liquidity in asset markets. *Journal of Economic Theory*, 136(1):66–104.

Vayanos, D. and Weill, P.-o. (2008). A search-based theory of the on-the-run phenomenon. *The Journal of Finance*, 63(3):1361–1398.

Wang, A., Hamilton, W. L., and Leskovec, J. (2016). Learning linguistic descriptors of user roles in online communities. *NLP+ CSS 2016*, page 76.

Wang, J. et al. (2008). Why are exchange rates so difficult to predict? *Economic Letter*, 3.

Wang, X., Jiao, L., and Wu, J. (2009). Adjusting from disjoint to overlapping community detection of complex networks. *Physica A: Statistical Mechanics and its Applications*, 388(24):5045–5056.

Warga, A. (1992). Bond returns, liquidity, and missing data. *Journal of Financial and Quantitative Analysis*, 27(04):605–617.

Weill, P.-O. (2008). Liquidity premia in dynamic bargaining markets. *Journal of Economic Theory*, 140(1):66–96.

Weninger, T. (2014). An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. *Social Network Analysis and Mining*, 4(1):1–19.

Wu, Z., Lin, Y., Wan, H., Tian, S., and Hu, K. (2012). Efficient overlapping community detection in huge real-world networks. *Physica A: Statistical Mechanics and its Applications*, 391(7):2475–2490.

Xia, Z. and Bu, Z. (2012). Community detection based on a semantic network. *Knowledge-Based Systems*, 26:30–39.

Xie, J., Kelley, S., and Szymanski, B. K. (2013). Overlapping community detection in networks: The state-of-the-art and comparative study. *Acm computing surveys (csur)*, 45(4):43.

Xu, Z., Ke, Y., Wang, Y., Cheng, H., and Cheng, J. (2012). A model-based approach to attributed graph clustering. In *Proceedings of the 2012 ACM SIGMOD international conference on management of data*, pages 505–516. ACM.

Yang, B., Di, J., Liu, J., and Liu, D. (2013). Hierarchical community detection with applications to real-world network analysis. *Data & Knowledge Engineering*, 83:20–38.

Yang, J. and Leskovec, J. (2012). Community-affiliation graph model for overlapping network community detection. In *Data Mining (ICDM), 2012 IEEE 12th International Conference on*, pages 1170–1175. IEEE.

Yao, J. and Tan, C. L. (2000). A case study on using neural networks to perform technical forecasting of forex. *Neurocomputing*, 34(1):79–98.

Yin, J. and Wang, J. (2014). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242. ACM.

Yu, G., Huang, R., and Wang, Z. (2010). Document clustering via dirichlet process mixture model with feature selection. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 763–772. ACM.

Zhang, D., Simoff, S., and Debenham, J. (2005). Exchange rate modelling using news articles and economic data. *AI 2005: Advances in Artificial Intelligence*, pages 467–476.

Zhang, D., Simoff, S., and Debenham, J. (2007). Exchange rate modelling for e-negotiators using text mining techniques. In *E-Service Intelligence*, pages 191–211. Springer.

Zhang, X., Fuehres, H., and Gloor, P. A. (2011). Predicting stock market indicators through twitter i hope it is not as bad as i fear. *Procedia-Social and Behavioral Sciences*, 26:55–62.

Zhao, Q. J. Y. (2013). *Risk Analysis for Corporate Bond Portfolios*. PhD thesis, Citeseer.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *European Conference on Information Retrieval*, pages 338–349. Springer.

Zhao, X. and Jiang, J. (2011). An empirical comparison of topics in twitter and traditional media. *Singapore Management University School of Information Systems Technical paper series. Retrieved November*, 10:2011.

Zhao, Z., Feng, S., Wang, Q., Huang, J. Z., Williams, G. J., and Fan, J. (2012). Topic oriented community detection through social objects and link analysis in social networks. *Knowledge-Based Systems*, 26:164–173.

Zhou, D., Manavoglu, E., Li, J., Giles, C. L., and Zha, H. (2006). Probabilistic models for discovering e-communities. In *Proceedings of the 15th international conference on World Wide Web*, pages 173–182. ACM.

Zhu, R. and Jiang, W. (2016). Combining random walks and nonparametric bayesian topic model for community detection. *arXiv preprint arXiv:1607.05573*.

Appendix A

Appendix of Chapter 3

Word	Sequence	Word	Word	Avg.	Std	Doc.	Neg-	Pos-	Uncer-	Liti-	Const-	Super-	Interes-	Mo-	Irr_	Harv-	Syllables
AARDVARK	1	81	5.69E-09	3.07E-09	5.78E-07	45	0	0	0	0	0	0	0	0	0	0	2
AARDVARKS	2	2	1.40E-10	8.22E-12	7.84E-09	1	0	0	0	0	0	0	0	0	0	0	2
ABACI	3	8	5.62E-10	1.69E-10	7.10E-08	7	0	0	0	0	0	0	0	0	0	0	3
ABACK	4	5	3.51E-10	1.73E-10	7.53E-08	5	0	0	0	0	0	0	0	0	0	0	2
ABACUS	5	1752	1.23E-07	1.20E-07	1.11E-05	465	0	0	0	0	0	0	0	0	0	0	3
ABACUSES	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
ABAFT	7	4	2.81E-10	3.25E-11	3.10E-08	1	0	0	0	0	0	0	0	0	0	0	2
ABALONE	8	80	5.62E-09	3.88E-09	1.04E-06	30	0	0	0	0	0	0	0	0	0	0	4
ABALONES	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
ABANDON	10	80492	5.65E-06	5.25E-06	4.12E-05	45941	2009	0	0	0	0	0	0	0	0	1	3
ABANDONED	11	174298	1.22E-05	1.22E-05	8.81E-05	83234	2009	0	0	0	0	0	0	0	0	2	3
ABANDONING	12	15926	1.12E-06	9.95E-07	1.45E-05	10125	2009	0	0	0	0	0	0	0	0	2	4
ABANDONMENT	13	177889	1.25E-05	1.19E-05	8.19E-05	65686	2009	0	0	0	0	0	0	0	0	1	4
ABANDONMENTS	14	7091	4.98E-07	6.85E-07	1.75E-05	3891	2009	0	0	0	0	0	0	0	0	2	4
ABANDONS	15	6430	4.52E-07	2.13E-07	4.48E-06	4625	2009	0	0	0	0	0	0	0	0	2	3 f
ABASE	16	46	3.23E-09	3.19E-09	7.35E-07	39	0	0	0	0	0	0	0	0	0	0	2
ABASED	17	72	5.06E-09	1.10E-08	2.33E-06	59	0	0	0	0	0	0	0	0	0	0	2
ABASEMENT	18	6	4.21E-10	9.29E-11	4.02E-08	6	0	0	0	0	0	0	0	0	0	0	3
ABASEMENTS	19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABASES	20	2	1.40E-10	1.64E-10	1.28E-07	2	0	0	0	0	0	0	0	0	0	0	3
ABASH	21	3	2.11E-10	8.76E-11	4.94E-08	3	0	0	0	0	0	0	0	0	0	0	2
ABASHED	22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ABASHEDLY	23	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
ABASHES	24	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABASHING	25	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABASHMENT	26	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABASHMENTS	27	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABASING	28	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABATE	29	22938	1.61E-06	9.55E-07	1.32E-05	12824	0	0	0	0	0	0	0	0	0	1	2
ABATED	30	26393	1.85E-06	9.47E-07	1.39E-05	12126	0	0	0	0	0	0	0	0	0	2	3
ABATEMENT	31	99853	7.01E-06	3.69E-06	4.10E-05	31115	0	0	0	0	0	0	0	0	0	2	3
ABATEMENTS	32	10159	7.14E-07	4.77E-07	1.50E-05	5543	0	0	0	0	0	0	0	0	0	2	3
ABATES	33	331	2.33E-08	1.47E-08	1.30E-06	263	0	0	0	0	0	0	0	0	0	2	2
ABATING	34	1821	1.28E-07	7.47E-08	3.16E-06	1256	0	0	0	0	0	0	0	0	0	2	3
ABATTOIR	35	125	8.78E-09	5.79E-09	8.25E-07	56	0	0	0	0	0	0	0	0	0	0	3
ABATTOIRS	36	136	9.55E-09	5.80E-09	1.24E-06	41	0	0	0	0	0	0	0	0	0	0	3
ABBE	37	180	1.26E-08	1.33E-08	1.92E-06	106	0	0	0	0	0	0	0	0	0	0	1
ABBES	38	4	2.81E-10	8.95E-11	6.04E-08	2	0	0	0	0	0	0	0	0	0	0	2
ABBESS	39	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ABBESSES	40	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
ABBEY	41	1163	8.17E-08	8.36E-08	1.18E-05	415	0	0	0	0	0	0	0	0	0	0	2
ABBEYS	42	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ABBOT	43	452	3.18E-08	3.64E-08	4.53E-06	201	0	0	0	0	0	0	0	0	0	0	2
ABBOTS	44	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
ABBREVIATE	45	896	6.29E-08	7.66E-08	6.35E-06	360	0	0	0	0	0	0	0	0	0	0	4
ABBREVIATED	46	14778	1.04E-06	9.82E-07	1.92E-05	8808	0	0	0	0	0	0	0	0	0	0	5
ABBREVIATES	47	62	4.36E-09	5.33E-09	1.17E-06	40	0	0	0	0	0	0	0	0	0	0	4
ABBREVIATING	48	17	1.19E-09	3.31E-09	9.40E-07	16	0	0	0	0	0	0	0	0	0	0	5
ABBREVIATION	49	7236	5.08E-07	2.46E-07	5.26E-06	4167	0	0	0	0	0	0	0	0	0	0	5
ABBREVIATIONS	50	25648	1.80E-06	8.53E-07	1.28E-05	12253	0	0	0	0	0	0	0	0	0	0	5
ABDICATE	51	46	3.23E-09	1.68E-09	4.43E-07	32	0	0	0	0	0	0	0	0	0	1	3
ABDICATED	52	112	7.87E-09	4.29E-09	6.24E-07	66	2009	0	0	0	0	0	0	0	0	2	4
ABDICATES	53	1	7.02E-11	1.88E-11	1.80E-08	1	2009	0	0	0	0	0	0	0	0	2	3
ABDICATING	54	29	2.04E-09	1.31E-09	3.02E-07	21	2009	0	0	0	0	0	0	0	0	2	4
ABDICATION	55	597	4.19E-08	2.68E-08	1.82E-06	505	2009	0	0	0	0	0	0	0	0	2	4
ABDICATIONS	56	0	0	0	0	0	2009	0	0	0	0	0	0	0	0	2	4
ABDOMEN	57	1588	1.12E-07	7.68E-08	3.70E-06	1015	0	0	0	0	0	0	0	0	0	0	3
ABDOMENS	58	6	4.21E-10	1.30E-10	6.78E-08	5	0	0	0	0	0	0	0	0	0	0	3
ABDOMINAL	59	6471	4.55E-07	3.34E-07	1.04E-05	2884	0	0	0	0	0	0	0	0	0	0	4
ABDOMINALS	60	23	1.62E-09	3.20E-09	6.87E-07	23	0	0	0	0	0	0	0	0	0	0	4

Table A.1: First rows of the sentiment lexicon used in Twitter paper.

Stemmed Word	Word	Avg.	Std	Doc.	Neg-	Pos-	Uncer-	Liti-	Const-	Super-	Interes-	Mo-	Irr_	Harv-	Syllables
0 aardvark	83	5.83E-09	3.08E-09	5.86E-07	46	0	0	0	0	0	0	0	0	0	0
1 ab	68898	4.84E-06	4.64E-06	0.000134153	8965	0	0	0	0	0	0	0	0	0	0
2 abaci	8	5.62E-10	1.69E-10	7.10E-08	7	0	0	0	0	0	0	0	0	0	0
3 aback	5	3.51E-10	1.73E-10	7.53E-08	5	0	0	0	0	0	0	0	0	0	0
4 abacus	1752	1.23E-07	1.20E-07	1.11E-05	465	0	0	0	0	0	0	0	0	0	0
5 abaft	4	2.81E-10	3.25E-11	3.10E-08	1	0	0	0	0	0	0	0	0	0	0
6 abalon	80	5.62E-09	3.88E-09	1.04E-06	30	0	0	0	0	0	0	0	0	0	0
7 abandon	462126	3.24E-05	3.12E-05	0.00024768	213502	2009	0	0	0	0	0	0	0	0	1.666666667
8 abas	126	8.85E-09	1.44E-08	3.23E-06	106	0	0	0	0	0	0	0	0	0	0
9 abash	3	2.11E-10	8.76E-11	4.94E-08	3	0	0	0	0	0	0	0	0	0	0
10 abat	161495	1.13E-05	6.16E-06	8.76E-05	63127	0	0	0	0	0	0	0	0	0	1.833333333
11 abattoir	261	1.83E-08	1.16E-08	2.07E-06	97	0	0	0	0	0	0	0	0	0	0
12 abb	184	1.29E-08	1.34E-08	1.98E-06	108	0	0	0	0	0	0	0	0	0	0
13 abbess	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14 abbey	1163	8.17E-08	8.36E-08	1.18E-05	415	0	0	0	0	0	0	0	0	0	0
15 abbot	452	3.18E-08	3.64E-08	4.53E-06	201	0	0	0	0	0	0	0	0	0	0
16 abbrevi	48637	3.42E-06	2.17E-06	4.57E-05	25644	0	0	0	0	0	0	0	0	0	0
17 abdic	785	5.51E-08	3.41E-08	3.21E-06	625	1674.166667	0	0	0	0	0	0	0	0	1.833333333
18 abdomen	1594	1.12E-07	7.69E-08	3.77E-06	1020	0	0	0	0	0	0	0	0	0	0
19 abdomin	6494	4.57E-07	3.37E-07	1.11E-05	2907	0	0	0	0	0	0	0	0	0	0
20 abduct	379	2.67E-08	3.26E-08	5.50E-06	243	0	0	0	0	0	0	0	0	0	0
21 abductor	4	2.81E-10	1.81E-10	8.85E-08	4	0	0	0	0	0	0	0	0	0	0
22 abe	17	1.19E-09	1.18E-09	3.40E-07	16	0	0	0	0	0	0	0	0	0	0
23 abeam	27	1.90E-09	2.16E-09	6.44E-07	15	0	0	0	0	0	0	0	0	0	0
24 aberr	2732	1.92E-07	1.74E-07	1.18E-05	2014	2009	0	0	0	0	0	0	0	0	0
25 abet	24212	1.70E-06	1.50E-06	3.19E-05	12630	502.25	0	0	0	0	0	0	0	0	0
26 abett	29	2.04E-09	3.75E-09	9.65E-07	21	0	0	0	0	0	0	0	0	0	0
27 abettor	170	1.19E-08	1.54E-08	2.72E-06	132	0	0	0	0	0	0	0	0	0	0
28 abey	9466	6.65E-07	4.62E-07	1.13E-05	4453	0	0	2009	0	0	0	0	0	0	0
29 abhor	10	7.02E-10	4.92E-10	1.74E-07	10	0	0	0	0	0	0	0	0	0	1.25
30 abhorr	36	2.53E-09	2.25E-09	5.84E-07	22	0	0	0	0	0	0	0	0	0	2
31 abid	55941	3.93E-06	2.22E-06	2.69E-05	33076	0	0	0	0	574	0	0	0	0	0
32 abil	8147058	0.000572326	0.000604568	0.000571784	803945	0	0	0	0	0	0	0	0	0	0
33 abject	29	2.04E-09	1.02E-09	3.23E-07	26	0	0	0	0	0	0	0	0	0	1.5
34 abjur	7	4.92E-10	4.17E-10	3.25E-07	4	0	0	0	0	0	0	0	0	0	0
35 abjuratori	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
36 abl	3253260	0.000228539	0.00023187	0.000343182	553588	0	2009	0	0	0	0	0	0	0	0
37 ablat	19167	1.35E-06	9.43E-07	4.54E-05	4021	0	0	0	0	0	0	0	0	0	0
38 ablaz	24	1.69E-09	2.18E-09	6.08E-07	14	0	0	0	0	0	0	0	0	0	0
39 abler	10	7.02E-10	2.93E-10	1.02E-07	10	0	0	0	0	0	0	0	0	0	0
40 ablest	108	7.59E-09	1.56E-08	5.16E-06	31	0	0	0	0	0	0	0	0	0	0
41 abli	750	5.27E-08	2.86E-08	2.28E-06	597	0	0	0	0	0	0	0	0	0	0
42 abloom	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43 ablut	16	1.12E-09	7.01E-10	1.95E-07	16	0	0	0	0	0	0	0	0	0	0
44 abneg	5	3.51E-10	2.24E-10	1.58E-07	3	0	0	0	0	0	0	0	0	0	0
45 abnorm	74653	5.24E-06	4.94E-06	5.97E-05	42194	2009	0	0	0	0	0	0	0	0	1.75
46 aboard	6205	4.36E-07	3.71E-07	1.09E-05	3247	0	0	0	0	0	0	0	0	0	0
47 abod	332	2.33E-08	9.44E-09	1.05E-06	207	0	0	0	0	0	0	0	0	0	0
48 abolish	13308	9.35E-07	5.75E-07	1.56E-05	10632	1607.2	0	0	0	0	0	0	0	0	1.8
49 abolit	616	4.33E-08	2.97E-08	1.84E-06	507	0	0	0	0	0	0	0	0	0	0
50 abolition	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
51 abolitionist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
52 abomin	28	1.96E-09	1.25E-09	5.01E-07	18	0	0	0	0	0	0	0	0	0	0.125
53 aborigin	2191	1.54E-07	1.17E-07	8.28E-06	762	0	0	0	0	0	0	0	0	0	0
54 aborn	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
55 abort	4218	2.97E-07	3.14E-07	1.90E-05	2464	0	0	0	0	0	0	0	0	0	0
56 abortionist	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
57 abound	495	3.48E-08	3.93E-08	4.33E-06	457	0	0	0	0	0	0	0	0	0	0
58 about	5418019	0.000380612	0.000482292	0.000492471	744462	0	0	0	0	0	0	0	0	0	0
59 abov	7409756	0.00052053	0.000452254	0.000483819	724896	0	0	0	0	0	0	0	0	0	0
60 aboveboard	7	4.92E-10	2.28E-10	9.09E-08	7	0	0	0	0	0	0	0	0	0	0
61 aboveground	4751	3.34E-07	2.19E-07	8.26E-06	2962	0	0	0	0	0	0	0	0	0	0

Table A.2: Processed Sentimen Lexicon -Dictionary is reduced to a smaller dataset by using the stemmed version of the words and grouping the ones with the same root.

	hashtag	low_nohash_trend	nohash_trend	spaced_trend	time
0	0	monumental	Monumental	monumental	10
1	0	andy murray	Andy Murray	andy murray	10
2	0	pri	PRI	pri	10
3	0	prep	PREP	prep	10
4	0	lanata	Lanata	lanata	10

	timestamp	trend	woeid	en	nostop_trend
0	1373234195	Monumental	1	1	monumental
1	1373236603	Andy Murray	1	1	andy murray
2	1373250449	PRI	1	1	pri
3	1373258875	PREP	1	1	prep
4	1373259477	Lanata	1	1	lanata

	tokenized_trend	str_trend	tokenlist_trend	stemmed_trend	stem
0	['monumental']	monumental	['monumental']	['monument']	monument
1	['andy', 'murray']	andy murray	['andy', 'murray']	['andi', 'murray']	['andi', 'murray']
2	['pri']	pri	['pri']	['pri']	pri
3	['prep']	prep	['prep']	['prep']	prep
4	['lanata']	lanata	['lanata']	['lanata']	0

	Word Count	Word Proportion	Average Proportion	Std Dev	Doc Count
0	14717	1.03E-06	2.66E-05	4909	0
1	0	0	0	0	0
2	250	1.76E-08	1.45E-07	0.000117722	171
3	5070	3.56E-07	6.19E-07	8.75E-05	1898
4	0	0	0	0	0

	Negative	Positive	Uncertainty		
0	0	0	0		
1	0	0	0		
2	0	0	0		
3	0	0	0		
4	0	0	0		

Table A.3: Final version of the Twitter data.

Appendix B

Appendix of Chapter 4

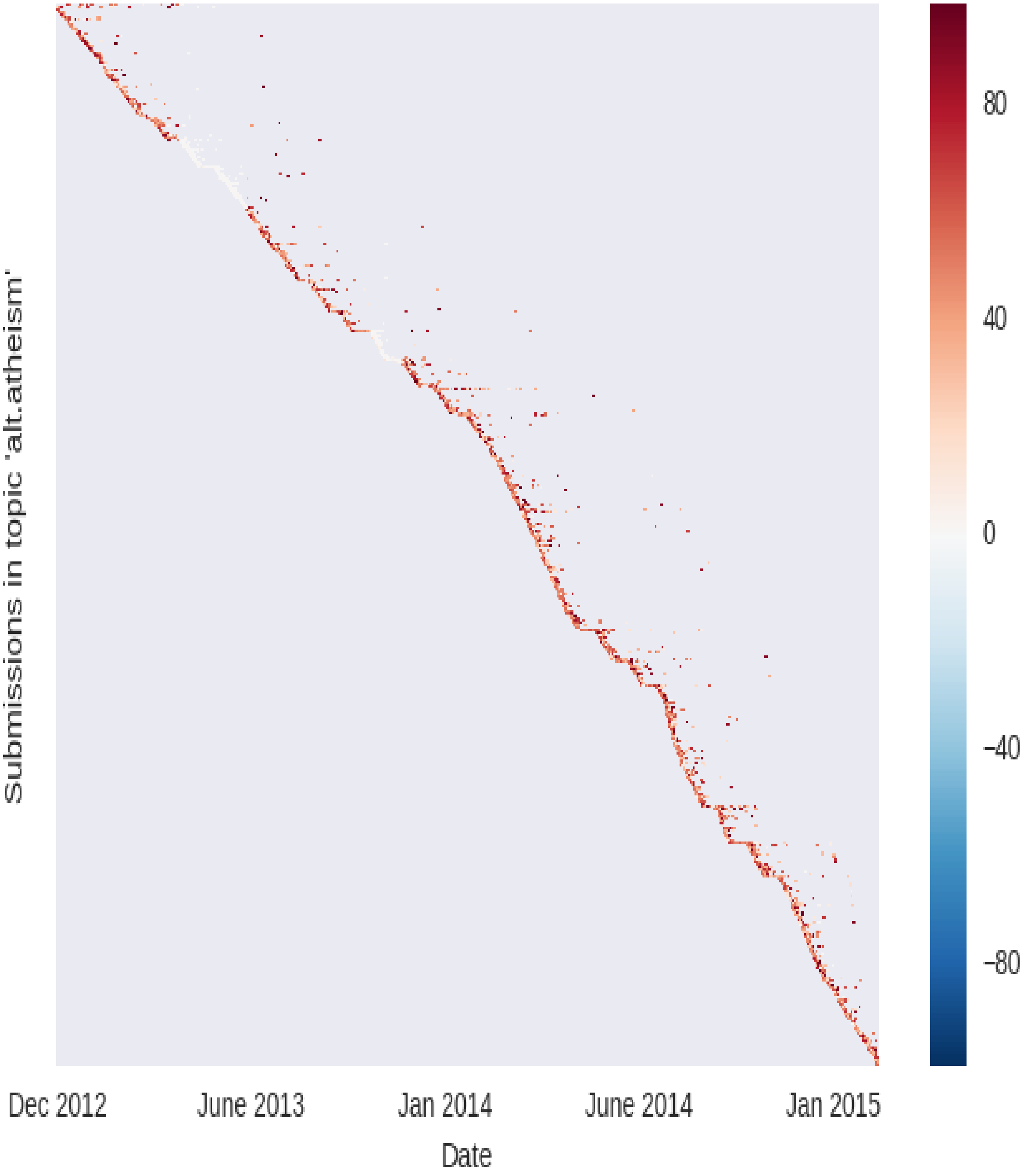


Figure B.1: How sentiments of clusters change over time under topics related to Atheism- This figure shows how sentiments (color coded) on each submission on atheism (on y-axis) change across time (on x-axis). On some submissions, discussions are continued for several months but the sentiments do not fluctuate much.

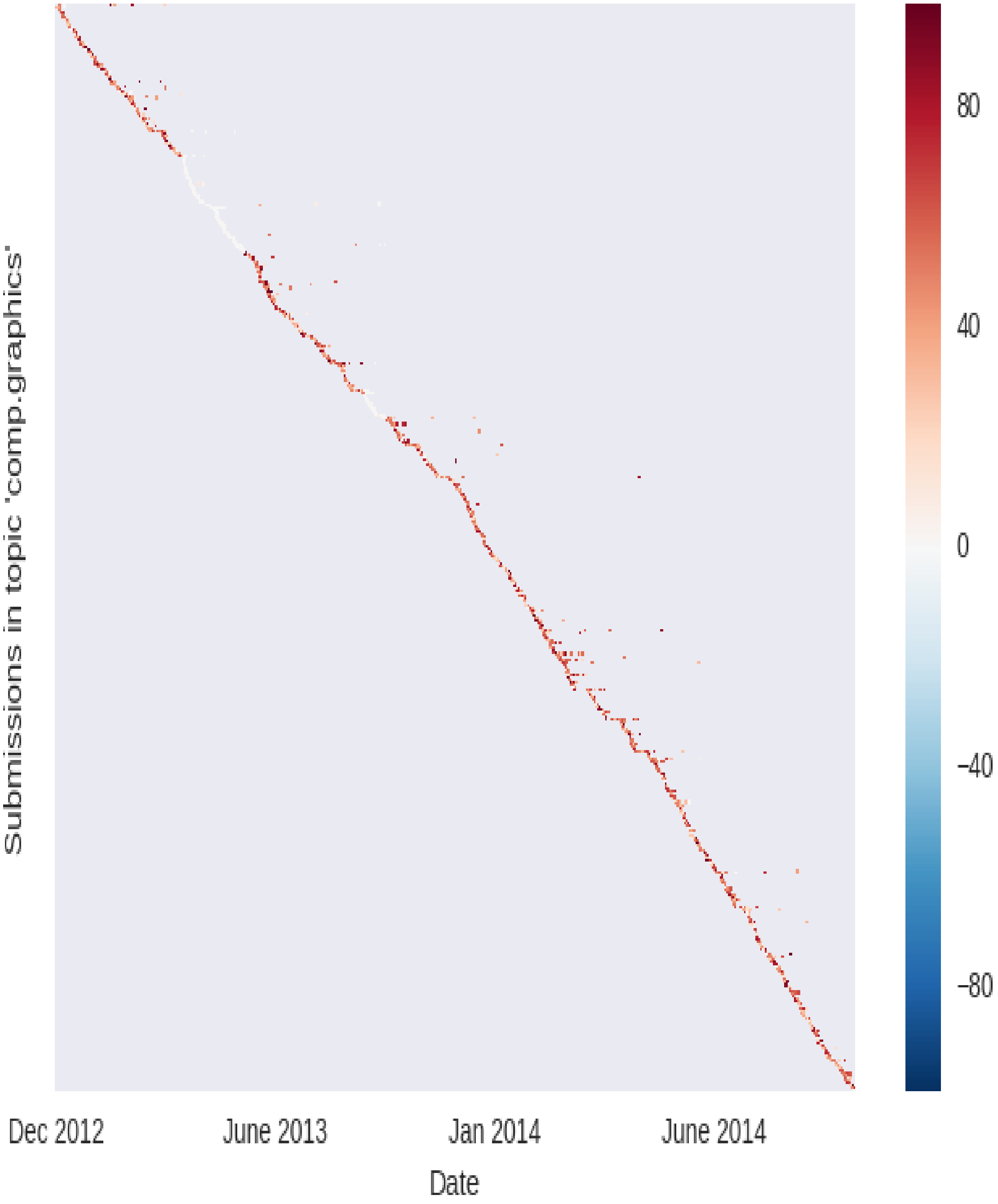


Figure B.2: How sentiments of clusters change over time under topics related to Graphics- This figure shows how sentiments (color coded) on each submission on graphics (on y-axis) change across time (on x-axis). On only a few submissions, discussions are continued for months but the sentiments do not fluctuate much.

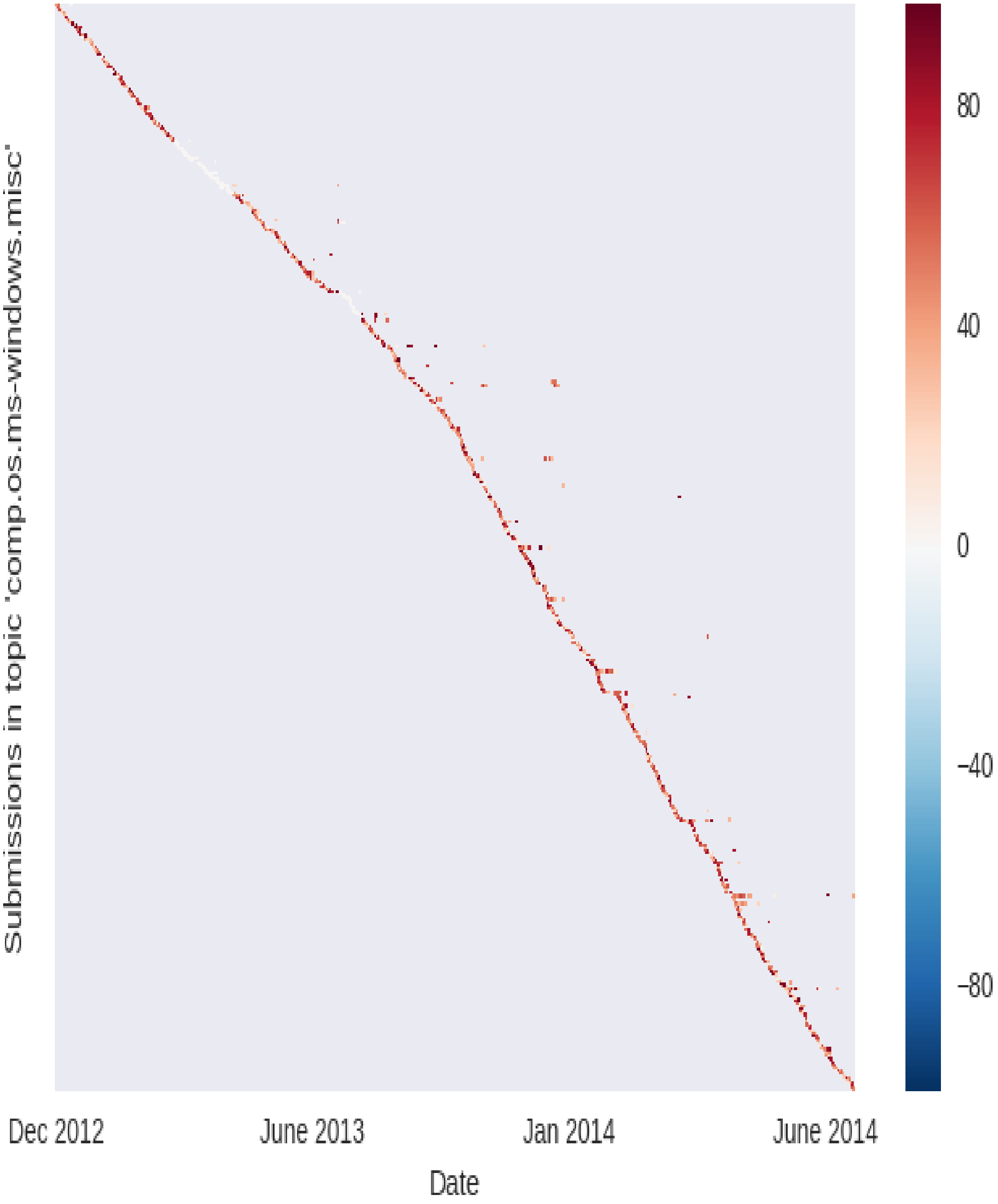


Figure B.3: How sentiments of clusters change over time under topics related to Windows- This figure shows how sentiments (color coded) on each submission on Windows (on y-axis) change across time (on x-axis). There are long discussions only during certain dates and on average the sentiments fluctuate between positive and neutral.

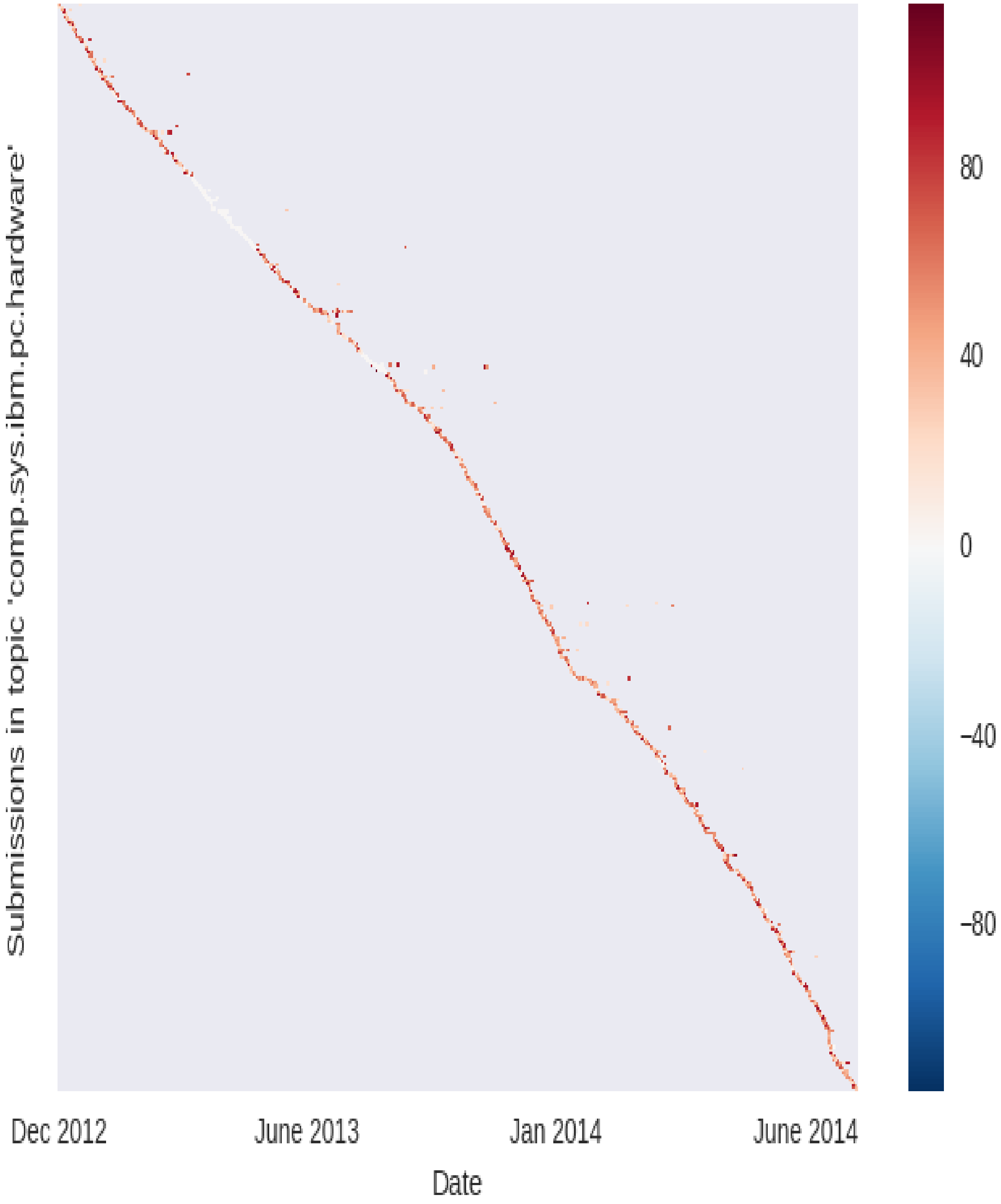


Figure B.4: How sentiments of clusters change over time under topics related to PC Hardware- This figure shows how sentiments (color coded) on each submission on PC hardware (on y-axis) change across time (on x-axis). Change in sentiment cluster is similar to topics on Windows.

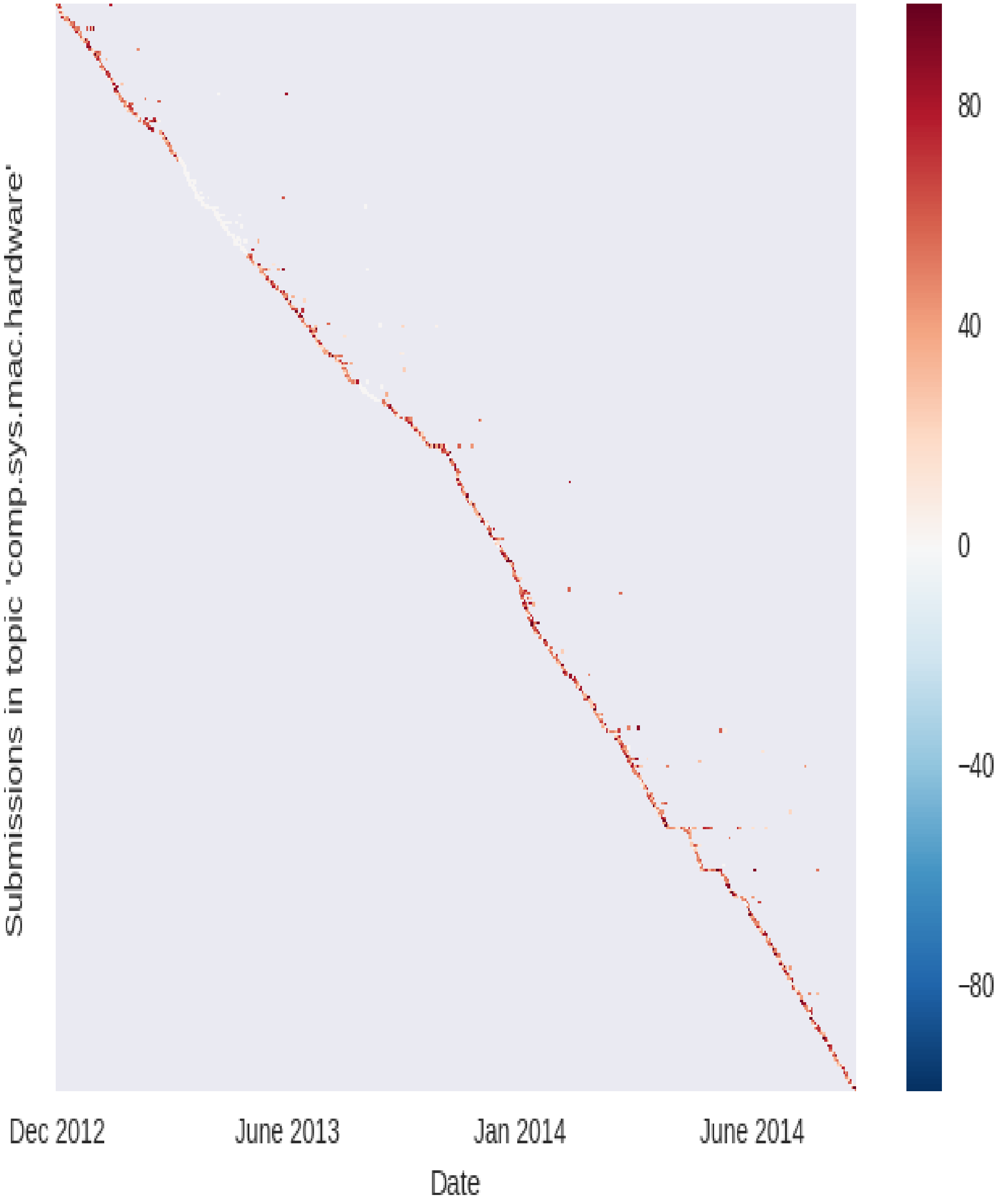


Figure B.5: How sentiments of clusters change over time under topics related to Mac Hardware- This figure shows how sentiments (color coded) on each submission on Mac hardware (on y-axis) change across time (on x-axis). The discussions last longer and there are more fluctuations in the sentiment compared to the submissions on PC hardware.

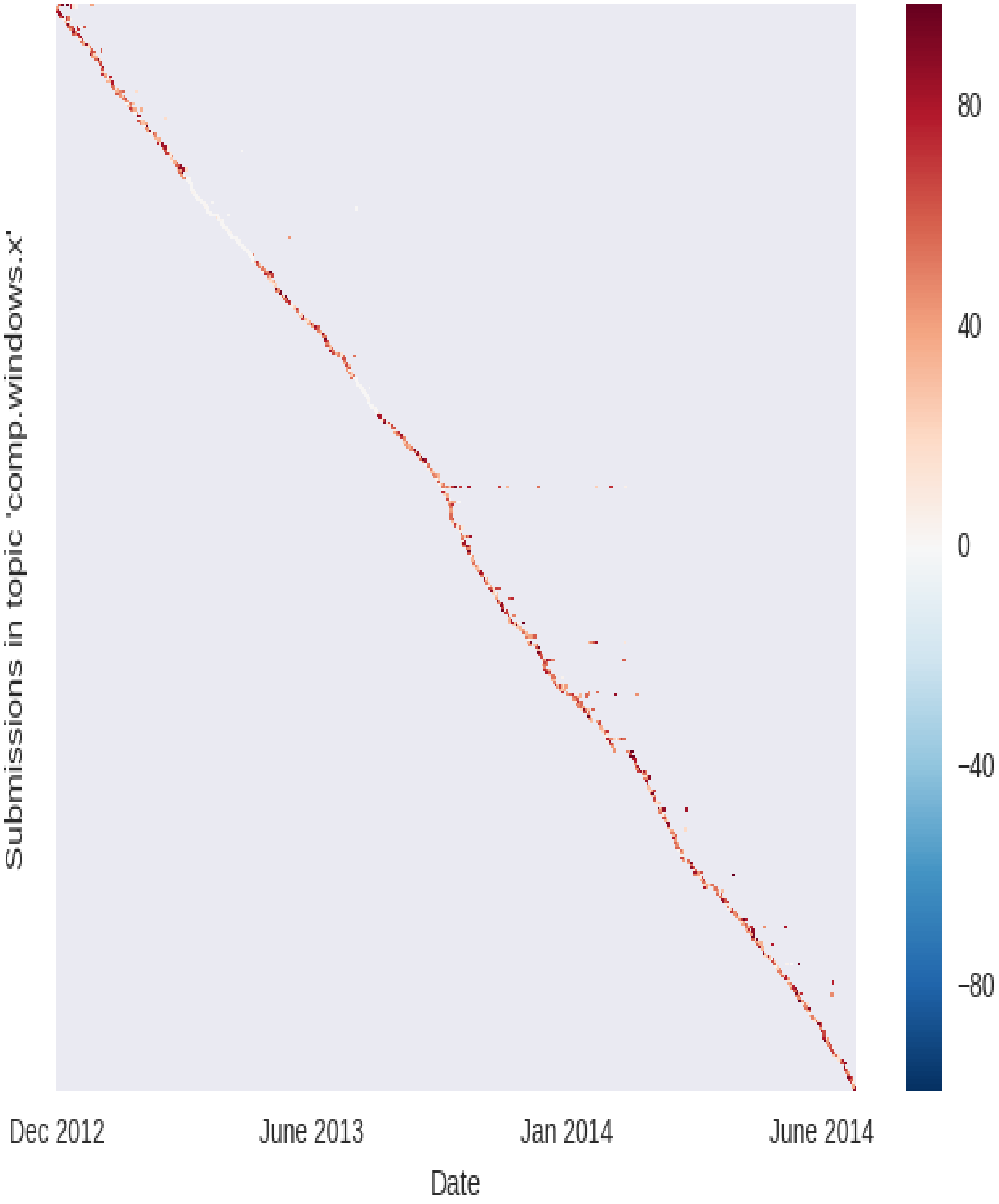


Figure B.6: How sentiments of clusters change over time under topics related to Windows X- This figure shows how sentiments (color coded) on each submission on Windows X (on y-axis) change across time (on x-axis). On a few submissions, discussions continue for a long time.

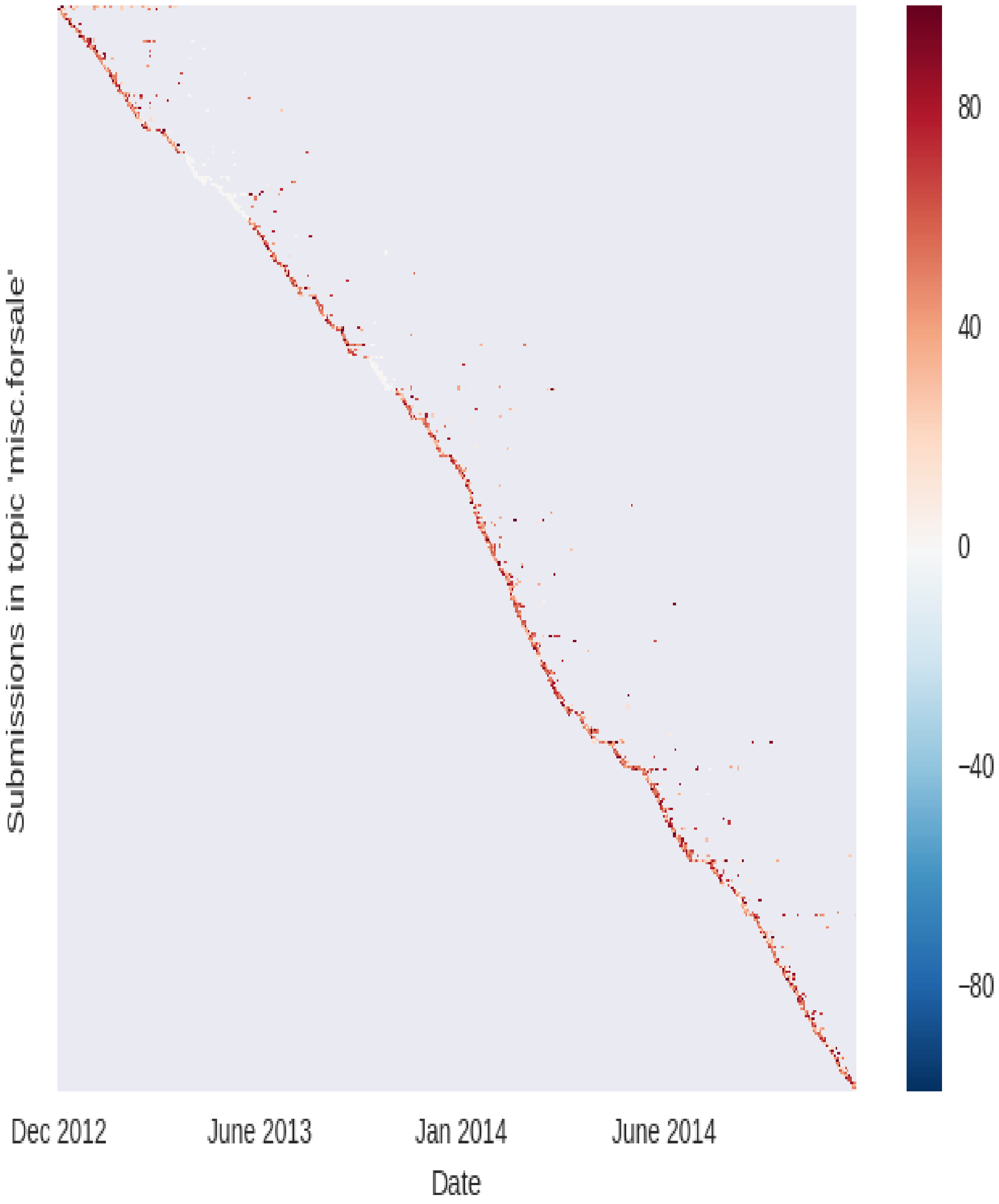


Figure B.7: How sentiments of clusters change over time under topics related to For sale- This figure shows how sentiments (color coded) on each submission on for sale (on y-axis) change across time (on x-axis). There are more fluctuations and longer discussions compared to previous categories.

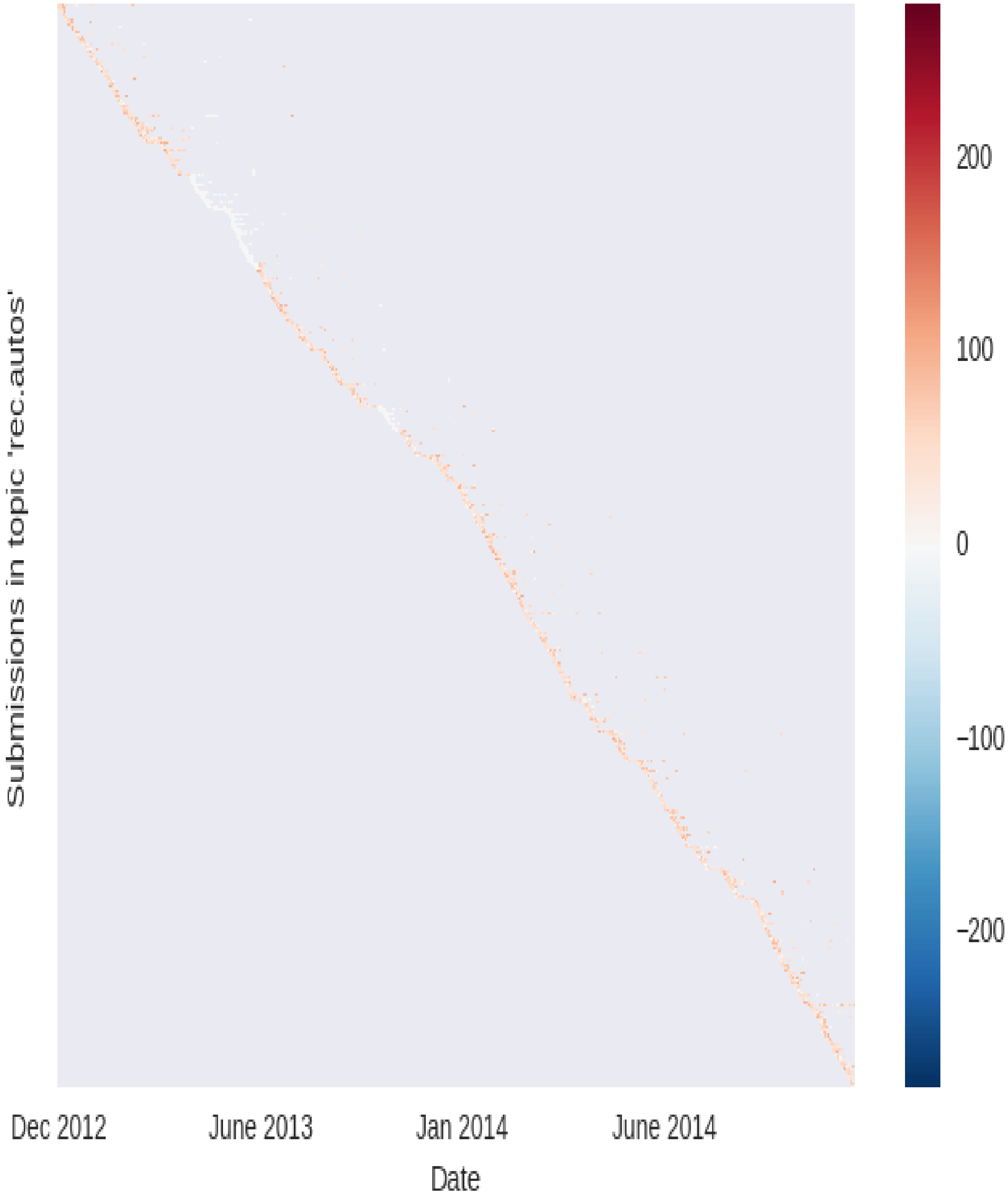


Figure B.8: How sentiments of clusters change over time under topics related to Autos- This figure shows how sentiments (color coded) on each submission on autos (on y-axis) change across time (on x-axis).

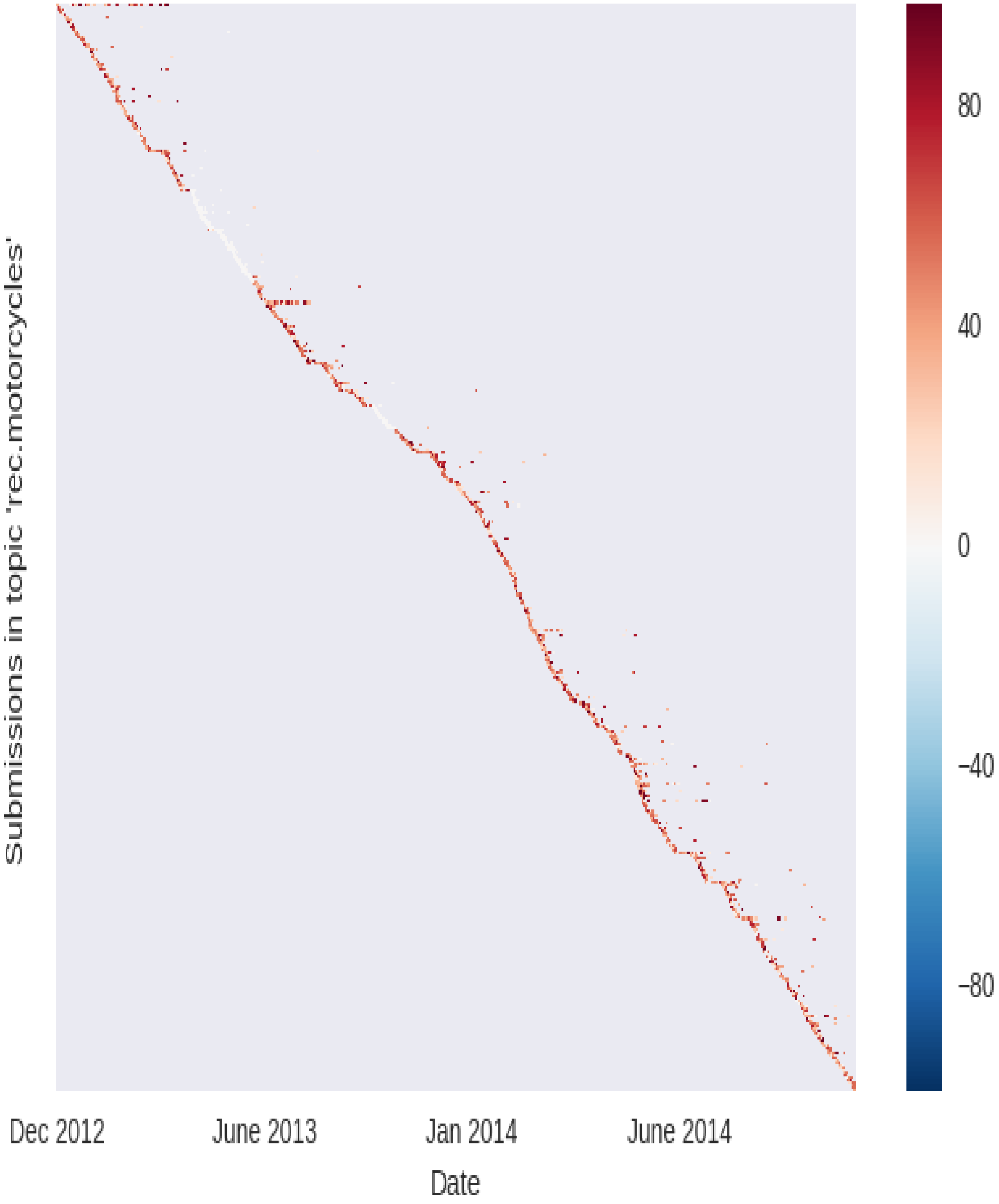


Figure B.9: How sentiments of clusters change over time under topics related to Motorcycles- This figure shows how sentiments (color coded) on each submission on motorcycles (on y-axis) change across time (on x-axis).

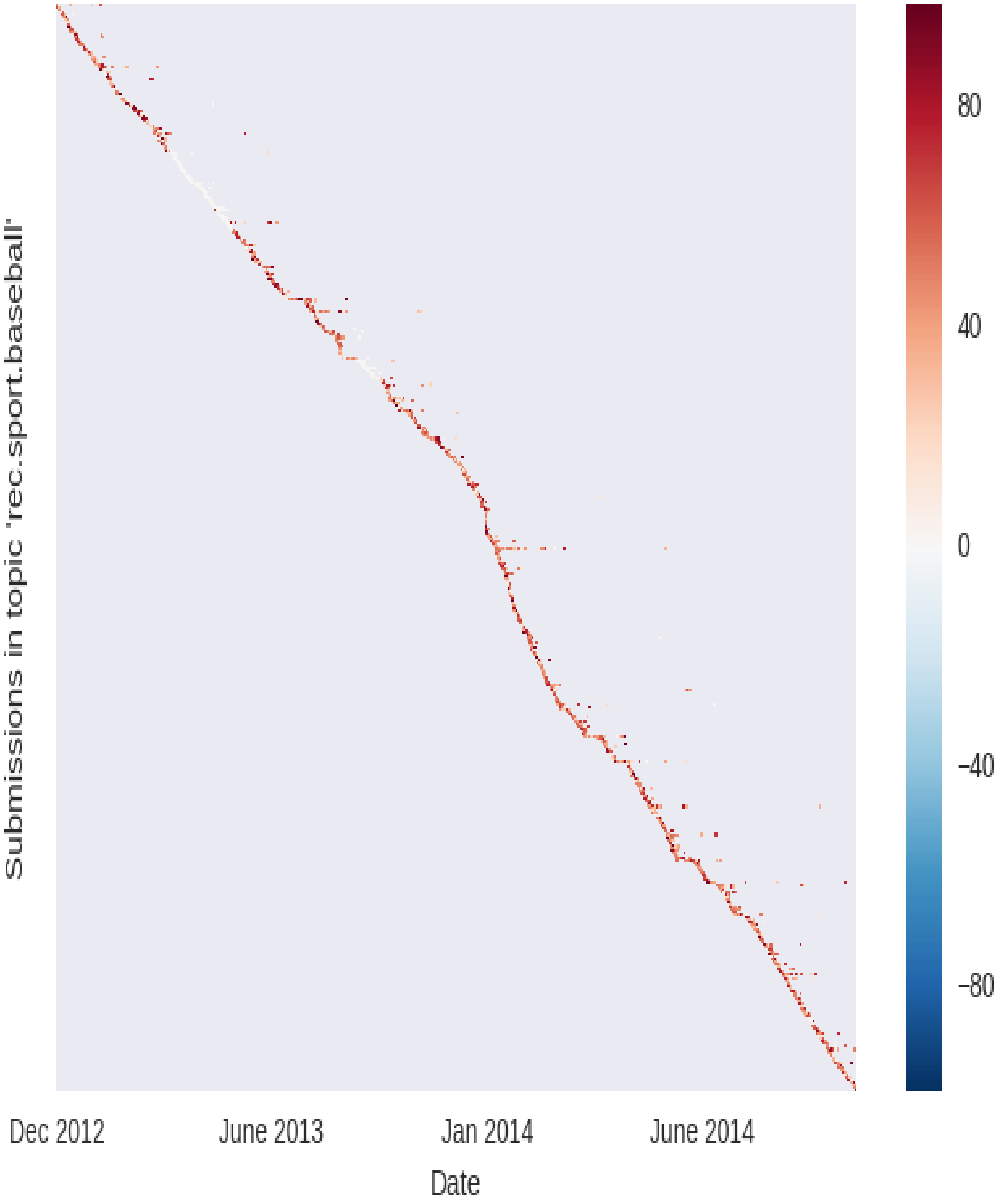


Figure B.10: How sentiments of clusters change over time under topics related to Baseball- This figure shows how sentiments (color coded) on each submission on baseball (on y-axis) change across time (on x-axis).

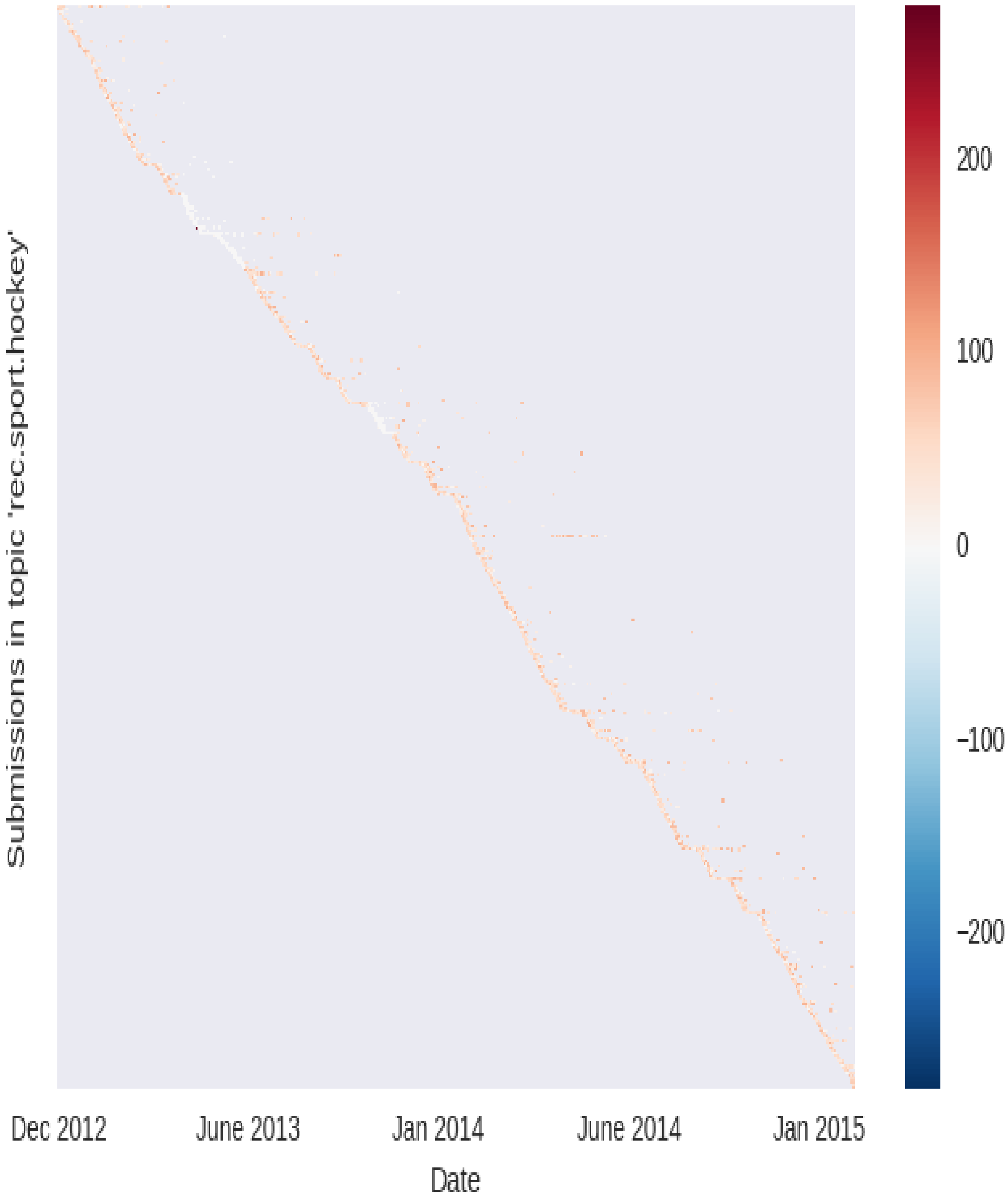


Figure B.11: How sentiments of clusters change over time under topics related to Hockey -This figure shows how sentiments (color coded) on each submission on hockey (on y-axis) change across time (on x-axis).

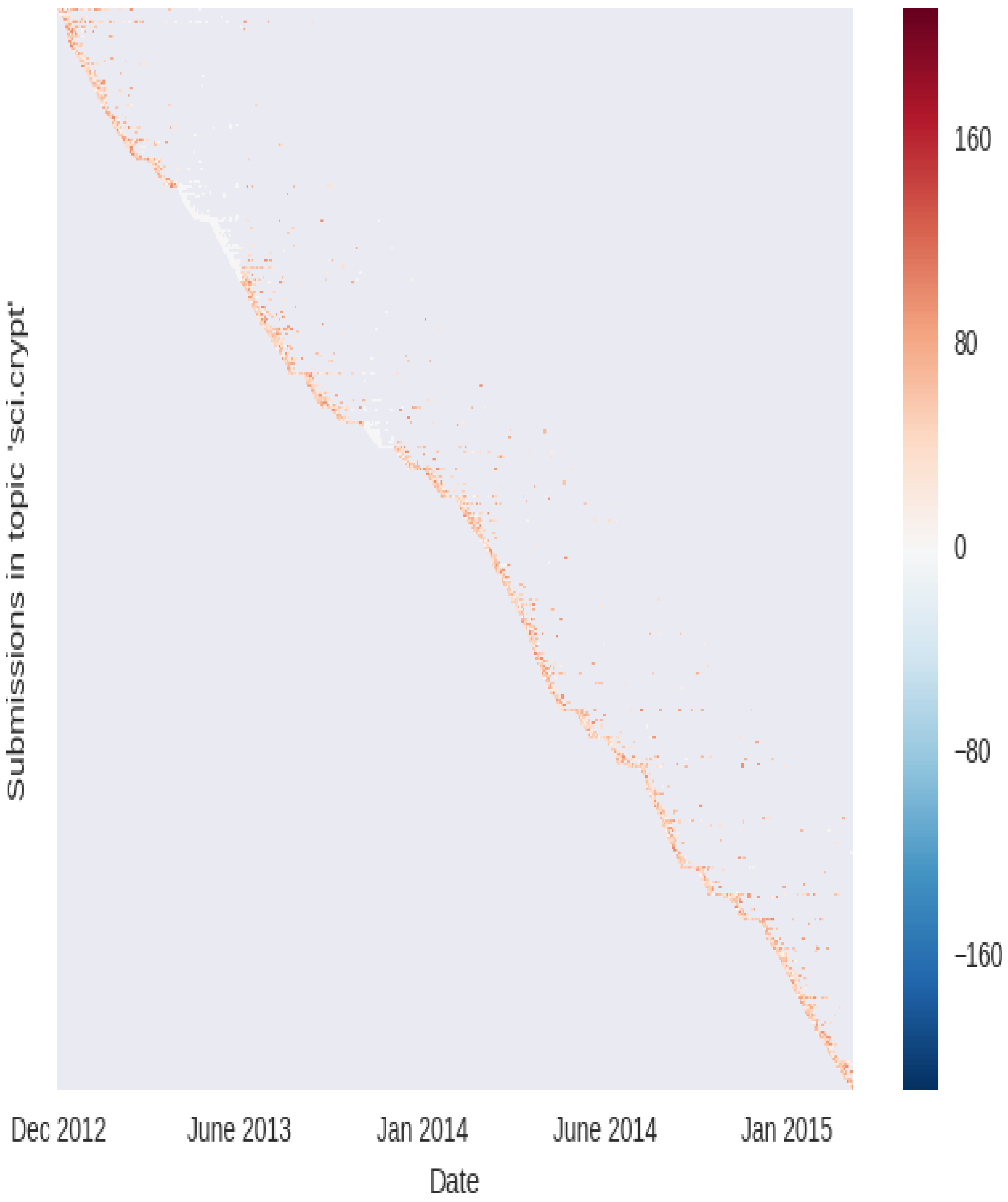


Figure B.12: How sentiments of clusters change over time under topics related to Crypt -This figure shows how sentiments (color coded) on each submission on atheism (on y-axis) change across time (on x-axis).

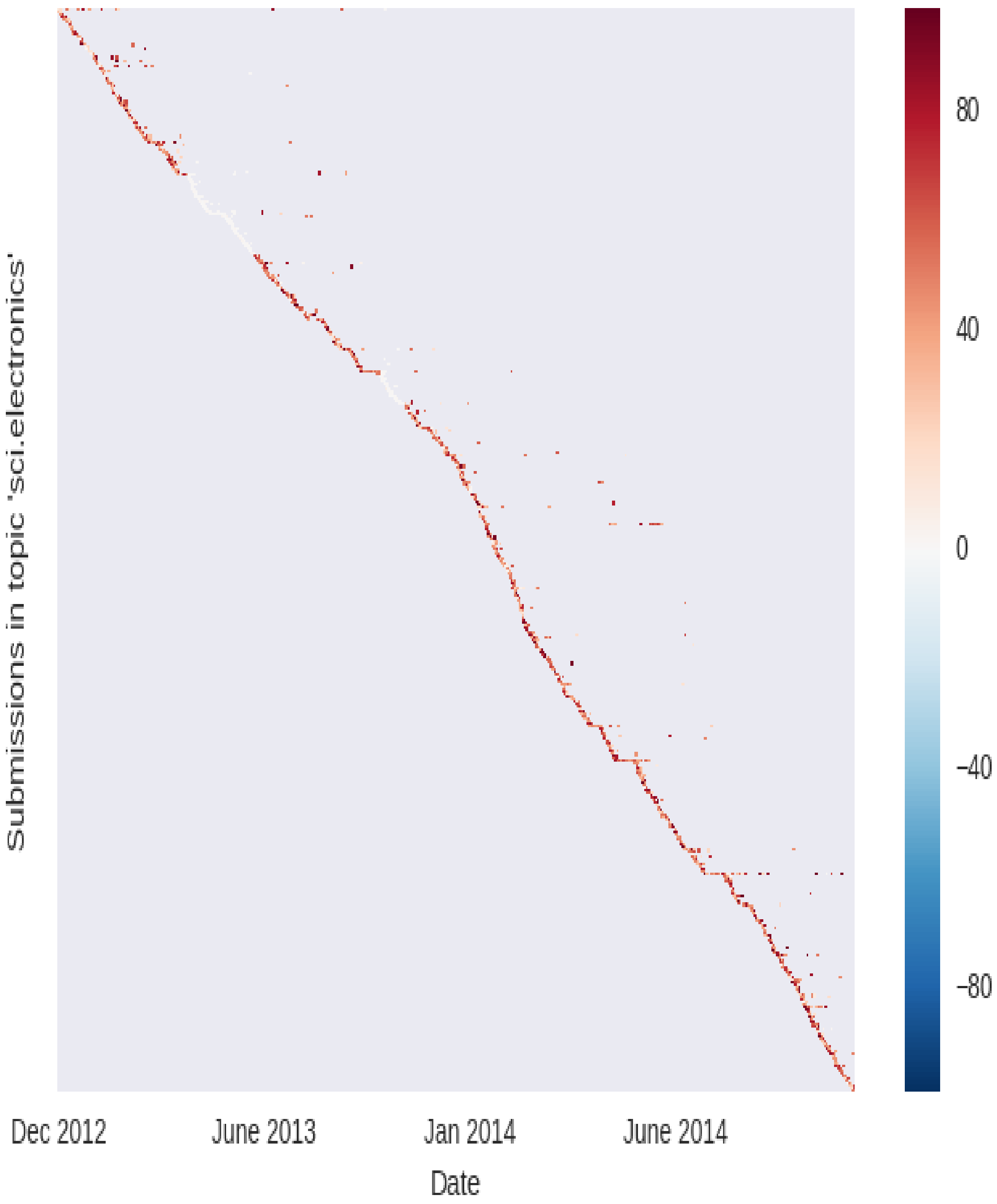


Figure B.13: How sentiments of clusters change over time under topics related to Electronics-
This figure shows how sentiments (color coded) on each submission on electronics (on y-axis) change across time (on x-axis).

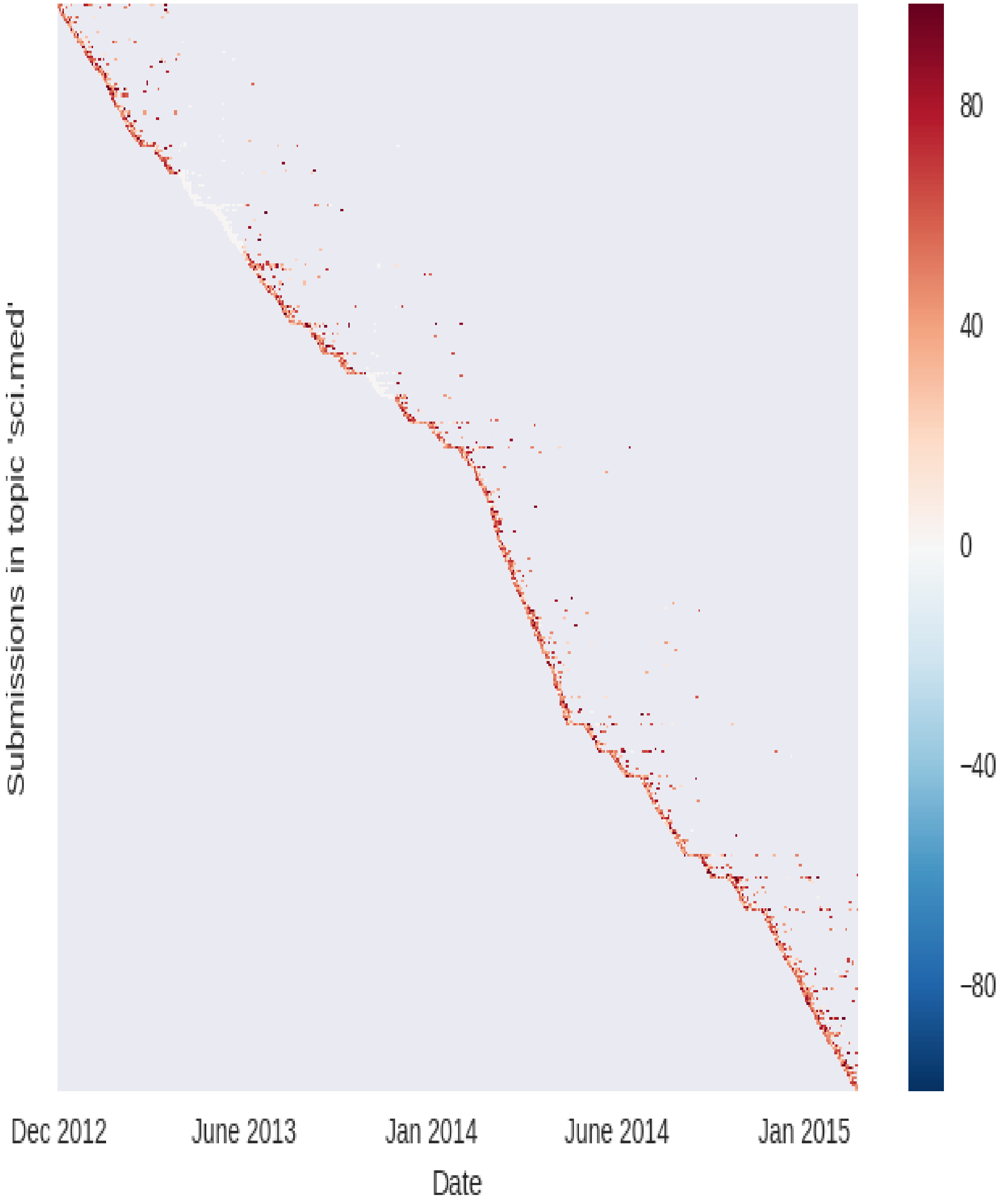


Figure B.14: How sentiments of clusters change over time under topics related to Med. -This figure shows how sentiments (color coded) on each submission on Med. (on y-axis) change across time (on x-axis). Sentiments are more positive compared to previous topics.

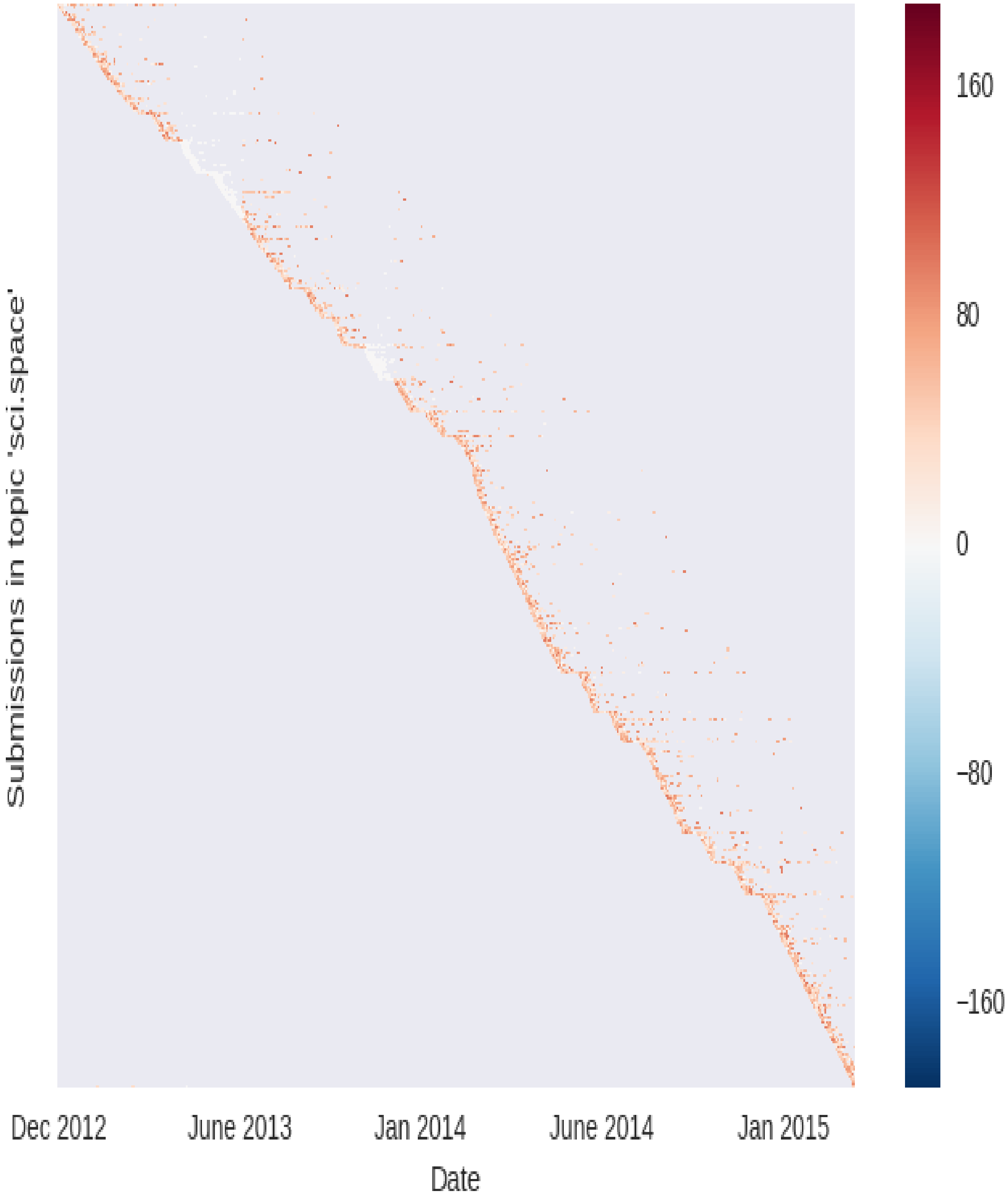


Figure B.15: How sentiments of clusters change over time under topics related to Space -This figure shows how sentiments (color coded) on each submission on space (on y-axis) change across time (on x-axis).

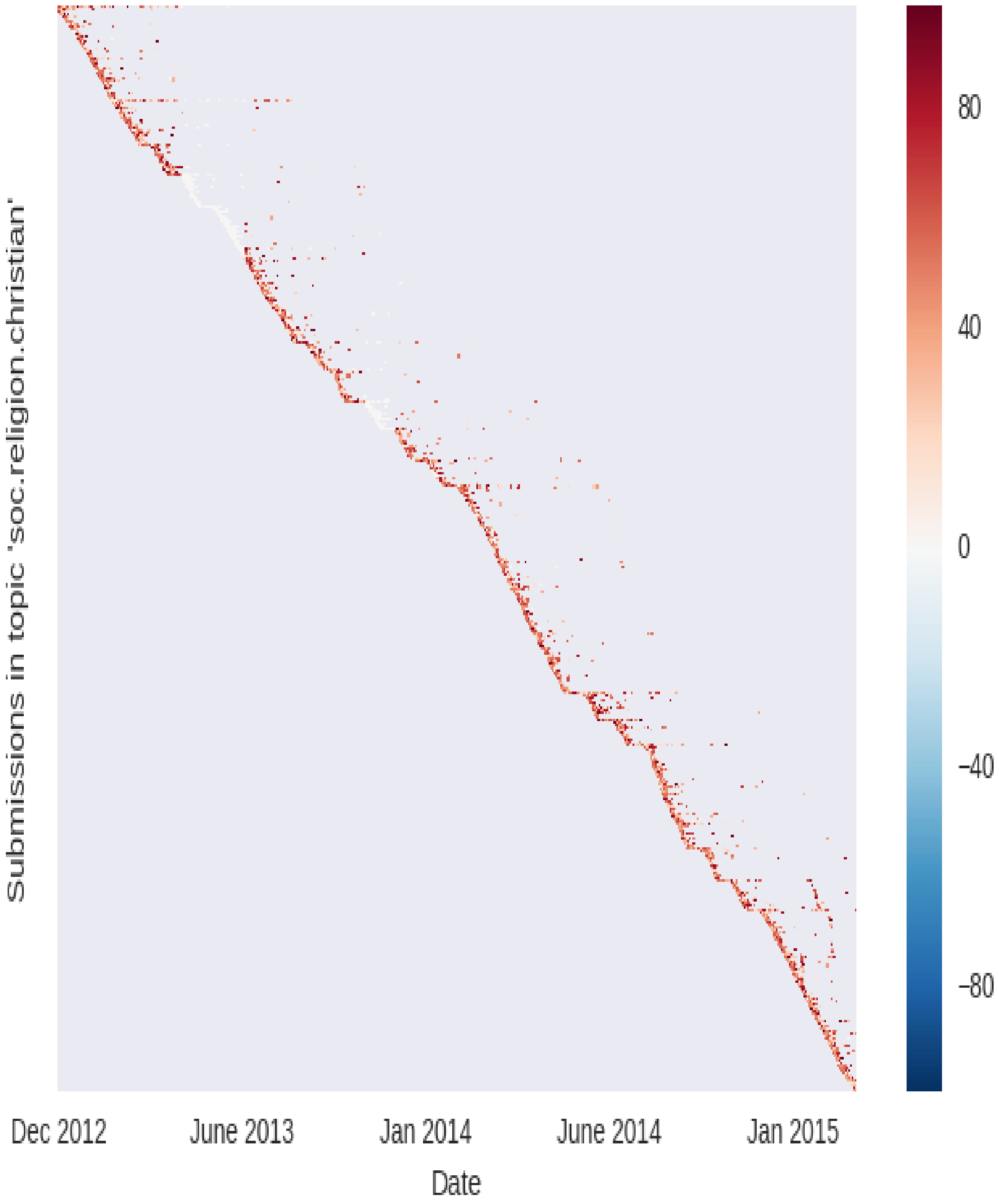


Figure B.16: How sentiments of clusters change over time under topics related to Christianity -This figure shows how sentiments (color coded) on each submission on Christianity (on y-axis) change across time (on x-axis). For most submissions, discussions are continued for several months and the sentiments fluctuate.

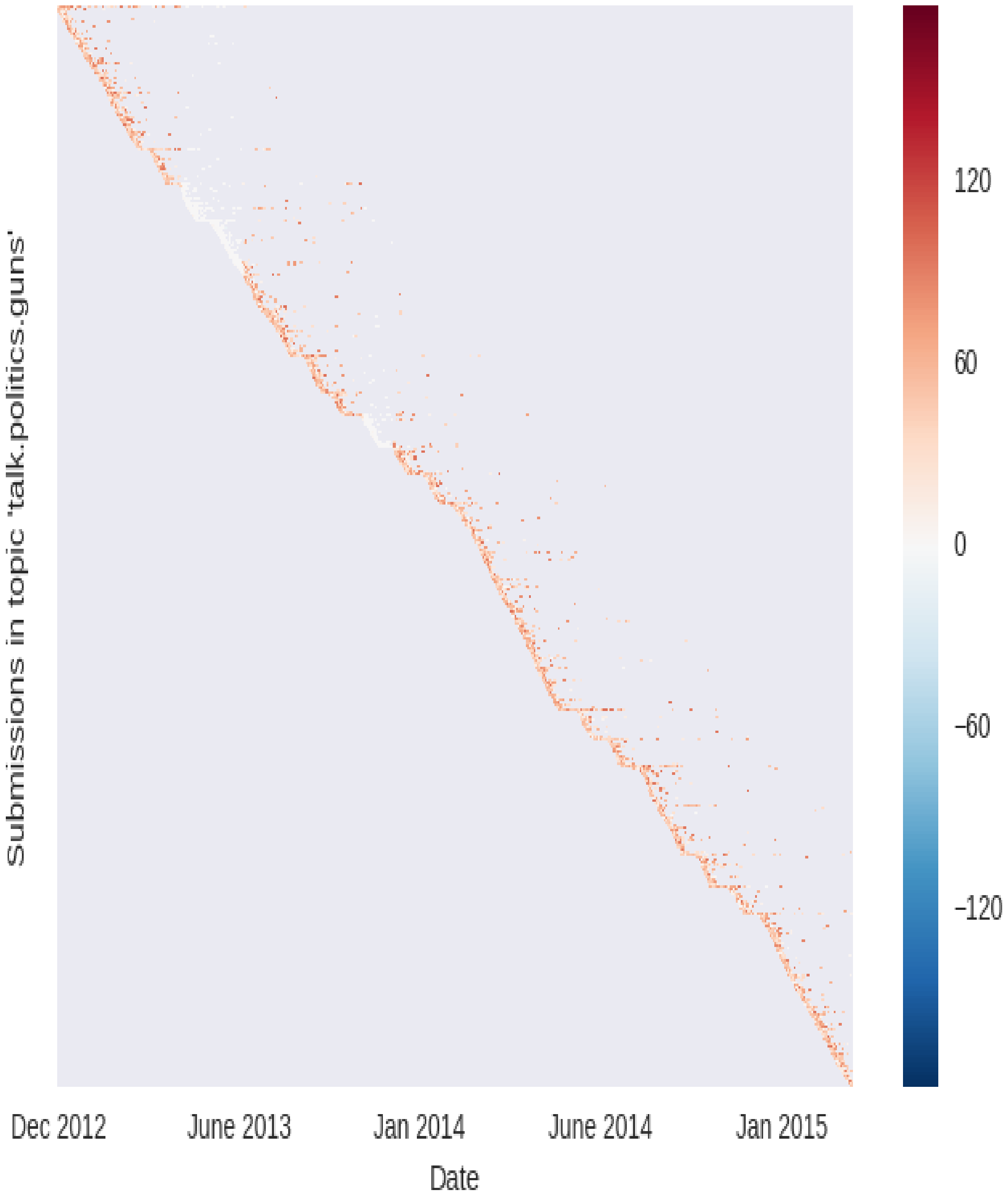


Figure B.17: How sentiments of clusters change over time under topics related to Guns -This figure shows how sentiments (color coded) on each submission on guns (on y-axis) change across time (on x-axis). For most submissions, discussions are continued for several months and the sentiments fluctuate.

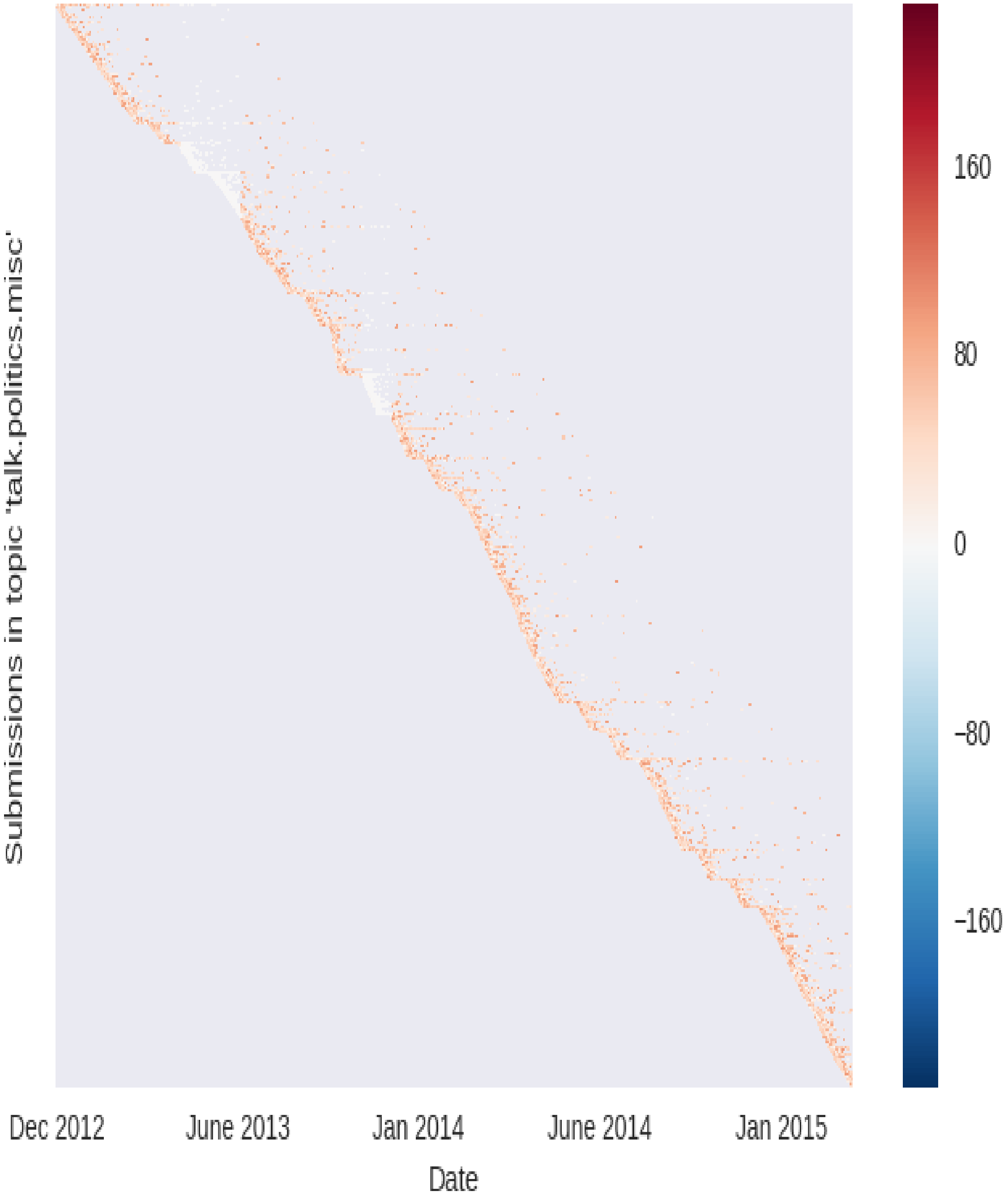


Figure B.18: How sentiments of clusters change over time under topics related to Politics -This figure shows how sentiments (color coded) on each submission on politics (on y-axis) change across time (on x-axis). Discussions are continued for several months and the sentiments vary a lot.