CHALMERS

Twitter Topic Modeling

Master's Thesis in the Master's programme in Algorithms Languages and Logic

KARINA BUNYIK

Department of Computer Science and Software Engineering

CHALMERS UNIVERSITY OF TECHNOLOGY
Göteborg, Sweden 2014

Twitter Topic Modeling

Karina Bunyik

Department of Computer Science and Engineering
Chalmers University of Technology

Gotheburg, Sweden 2014

Twitter Topic Modeling

Karina Bunyik

Examiner: DEVDATT DUBHASHI

Supervisor: NINA TAHMASEBI

Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
Göteborg, Sweden June 2014

Abstract

Following social media discussions related to real life events, has been a great topic of interest. There is no general method for deciding whether the social media discussions reflect the dynamics of the events or if they lead a life on their own. Existing methods for analyzing social media discussions rely on extensive manual work from domain experts and do not generalize well to discussions on languages other than English nor to various events. Combining the domain expert's knowledge with data driven approaches can lead to models that are applicable to different domains, and the same time are capable of handling large data amount from social media. In this research, we modeled the Twitter discussions about the Swedish party leader debate held on October 2013. We constructed a semi-automatic model based on *Term Frequency-Inverse Document Frequency* in order to identify and measure the debate topics on Twitter. For discovering other discussions, we made use of *Latent Dirichlet Allocation* - an unsupervised learning algorithm. We evaluated the models manually with the help of a domain expert. We compared the Twitter discussions to the topics the politicians were talking about on the debate. The correlation between the Twitter discussions and the debate topic corresponds to the results from a still ongoing political science research.

The political science domain expert Linn Sandberg from The University of Gothenburg, Department of Political Science contributed to the research by defining the research-question and evaluating the models.

Keywords. topic modeling, Twitter, LDA, tf-idf

Acknowledgements

I would like to give my thanks to my supervisor Nina Tahmasebi for providing continuous support and guide during this research.

I would also like to thank my examiner Devdatt Dubhashi, for giving me the opportunity to be part of an exciting and challenging study and for his advices that he provided whenever it was required.

Linn Sandberg, researcher at the Political Science Department of Gothenburg University of Technology, was able to offer me knowledge and instingts in political science and helped formulate the hypohtesis and research questions used in this thesis. Wihtout her, this reseach would not have been possible, thank you very much for your help!

Without Språkbanken's linguistic expertise and data, that I otherwise would not have been able to acquire, this work would not have gotten the attention and weight. Your help will not be forgotten.

Finally, I want to thank my family. My partner Steffen Strätz provided encouragement and support when the times got rough, I appreciate that very much. Many thanks to my parents for the financial help.

Contents

1 Introduction **6**
- 1.1 Motivation . . . 6
- 1.2 Hypothesis . . . 7
- 1.3 Similar Work . . . 8

2 Background **10**
- 2.1 Twitter . . . 10
- 2.2 Politics on Twitter . . . 10
- 2.3 Swedish Politics . . . 11
  - 2.3.1 Party Leader Debate . . . 12
- 2.4 Topic Modeling . . . 13
  - 2.4.1 Probabilistic Modeling . . . 13
  - 2.4.2 Latent Dirichlet Allocation . . . 13
  - 2.4.3 Probabilistic Latent Semantic Indexing . . . 14
- 2.5 Domain Specific Topic Modeling . . . 15
- 2.6 Term Frequency − Inverse Document Frequency . . . 15
- 2.7 Classification . . . 16
  - 2.7.1 Sparse Text Models in NLP . . . 17
  - 2.7.2 Sparse Matrix Representation . . . 17
- 2.8 Support Vector Machine . . . 17
- 2.9 Stochastic Gradient Descent . . . 19

3 Data **22**
- 3.1 Collection . . . 22
  - 3.1.1 Annotation . . . 23
  - 3.1.2 Data Description . . . 24

4 Pre-processing **27**
- 4.1 Pre-processing for the Unsupervised Algorithms . . . 27
- 4.2 Pre-processing for the Semi-supervised Algorithms . . . 27

5 Methodology **29**
- 5.1 Hashtag Classification . . . 29
- 5.2 Identifying Agenda Topics . . . 30
  - 5.2.1 Initial Topic List . . . 31
  - 5.2.2 Tf-idf Topic Extraction . . . 32
- 5.3 Finding Latent Topics . . . 34
  - 5.3.1 Filtering . . . 35
  - 5.3.2 Aggregation . . . 35

6 Results **38**
6.1 Hashtag Classification 38
6.2 Identifying Agenda Topics 39
6.3 Finding Latent Topics 41

7 Discussions **48**

8 Conclusion **50**
8.1 Limitation and Future Work 50
8.1.1 Identifying Agenda Topics 51
8.1.2 Finding Latent Topics 51

Appendices **60**

Appendix A JSON **60**

Appendix B Initial topic lists **62**

Appendix C Stochastic Gradient Descent **71**

Appendix D Negative Training Sample List **73**

Appendix E LDA Result With 50 topics **74**

Appendix F Words Filtering Swedish Language Tweets **75**

Appendix G Data Storage **76**

1 Introduction

Twitter is a convenient social media platform for discussing topics, which gives politically interested people the opportunity to interact with the politicians or spread ideas. In order to analyze political discussions on Twitter, we used data mining techniques. The vast amount of data and irregular language usage are one of the main challenges of doing Twitter data analysis. The goal driving this research is to compare the discussions on Twitter connected with the Swedish party leader debate held on October year 2013 to the actual debate broadcast. The methods we used in our work to answer the research-question are applicable to other fields of study. There is similar work done examining the Norwegian party leader debate related discussions on Twitter [58].

1.1 Motivation

Digital revolution is taking place in western politics. As social media appeared in the daily lives of politically interested people, the power of communication has shifted from central party administration to party leaders, staff, representatives, members and supporters. Now virtually anyone in the we tern world has the possibility to seed ideas, spread information and declare their opinion [38]. As social media spreads, political decision-making and interaction with the public are changing. The significant role of social media is bringing transparency, speed, interactivity and sharing to the political sector. The fact that all political parties have Twitter accounts shows the growing pressure on the party leaders to take part in the digital revolution [39]. If political parties are unable to maintain interests of social groups, it is likely they will lose their function as a party and they will be undermined.

Twitter is a large source of public opinion: at least 1% of its total data is available for data mining [2] purposes of this research[1], it is widely adapted: has 500 million users [4], and it provides information in a real-time short messaging format. All these features make Twitter more suitable for grand scale data analysis than other social media services. For example, the data from Facebook and Google+ is hard to access compared to Twitter. Some high impact studies have been made using Twitter data for building predictive models for the Arab Spring [6], predictive policing for law enforcement [7], identifying terrorist attacks [9] and presidential election forecasting [8]. These studies show the potential of using the data from Twitter for modeling opinions in national politics.

Nevertheless data mining Twitter also has some challenges. First of all many natural language techniques fail because of grammatically incorrect text, abbreviations and slang in tweets. Furthermore, compared to blogs and news media, tweets do not have clearly

defined context since each tweet is up to 140 characters long. Some context can be extracted from the retweet-reply network structure of a tweet, hashtags used in a tweet and the network of user subscriptions, but that is a rather complex task. Finally, the difficulty in using natural language tools can arise from the various usages in different cultures. For example, in the US, Twitter is mostly used for informative purposes, in the UK Twitter users are more interconnected, and in Indonesia and Canada users tend to form small clusters [62]. There has been various research about extracting knowledge from tweets created in English language [16], but those methods do not always generalize well to other languages. Another challenge of analyzing political discussions on Twitter is the problem of the lack of off-the-shelf methods for different domains. For example the sport domain is wide in the number of topics, it's not related to a specific event and might include various cultures. In contrast Swedish party leader debate domain has less topics, is connected to the debate event and is limited to the Swedish culture. The role of the political domain in the data might require adjustments of existing data mining techniques. If the domain is too specific and the current methods are not applicable, custom techniques should be used in order to build the models that are able to answer the research questions.

1.2 Hypothesis

This work is part of an ongoing research that explores the connection between mass media event and its reflection on Twitter. That study was performed by exploring the Swedish party leader debate and the corresponding discussions on Twitter. Firstly the study is interested in checking that Twitter users actively discuss the party leader debate on Twitter. Secondly, it measures to what extent the debate topics are discussed and thirdly, it is discovering which other politics related discussions come up on Twitter.

The hypothesis used in this research, stated by the political scientists in the study Linn Sandberg:
$\mathbf{H}_0$: *Twitter reflects the party leader debate.*

$\mathbf{H}_1$: *The party leader debate related dialogs differ from the party leader debate topics themselves.*

Proving the $\mathbf{H}_1$ hypothesis true or false might give the opportunity of using Twitter to extract public opinion. Not only the topics can be identified, but sentiment could be added to see how people feel about the issues they discuss. The hypothesis testing and interpretation of the results will be perform in Linn Sandbergs research.

The goal of this thesis is to build models using data mining techniques in order for the

political scientists to prove their hypothesis true or false. These models should provide quantitative measurements for concepts like party leader debate related tweets, agenda topic related tweets and non-agenda topic related tweets, in order to answer the following research questions from our domain expert:

RQ$_1$: *What's the magnitude of the party leader debate discussions in the Swedish Twitter stream?*

RQ$_2$: *How large are Twitter discussions about the debate topics?*

RQ$_3$: *What other political discussions are emerging on Twitter at the time of the debate?*

1.3 Similar Work

A recent study about the Norwagian party leader debate was exploring the related conversations on Twitter [58]. The goal of the work was to decide if the party leader debate TV broadcast mirrors the discussions on Twitter. Tweets from two debates in 2011 were analyzed with qualitative and quantitative methods. Multiple step analysis using the *IMSC* model (Issue, Meta, Sentiment, Close Reading) was used in the quantitative research. The magnitude of party leader related tweets was measured and the debate topics were identified. Sentiment of the discussions was classified to *supportive*, *critical* and *neutral*. The results of the study showed that Twitter mirrors the Norwegian party leader debate. The sentiment analysis showed critical sentiment pointed towards politicians. Our research differs in the language of the tweets and the data driven method. Since no comparison was made between the methods used in our research and the mentioned research, their performance cannot be compared.

Another research was comparing Twitter and news media topics [17]. The goal of the study was to discover Twitter topics which do not appear in the media. The authors compared the content of Twitter with the content of news media, in particular New York Times. For discovering the topics on Twitter, Twitter-LDA model was used, and natural language processing techniques were used for discovering topics on New York Times. The topics were classified into categories and types. The results showed that interesting entity-oriented topics can be found on Twitter, since there is low coverage of these topics in traditional news media. They also discovered that Twitter users are actively helping to spread news of important events. The mentioned study differers from our work in the domain of the research. While our research modeled tweets of the Swedish part leader domain, the mentioned work was performed on English Tweets. Moreover, we worked on data from 6 days time-span, the mentioned research had a 3 month time-span. Finally,

the main goal of our research also differs from the one mentioned.

2 Background

2.1 Twitter

Twitter is an online social networking and microblogging service. People use Twitter for connecting to people of similar interest, spreading information and opinions. Tweets, the publicly visible instant messages used in the microblogging services, have size limit up to 140 characters per message. Users of Twitter are able to subscribe to streams of other users and receive broadcasted messages from them. Each user's tweets are broadcasted to his or hers followers. Users may also send messages to single Twitter users. These messages are not accessible and therefore we did not include them in our research.

Tweets can be replies to other tweets. When a user wants to answer a tweet he or she can create a tweet and mark it as a reply. Replies are visible to all users who follow the author of the reply. Retweets are already posted tweets that are re-posted to the users own followers. Reply tweets and re-tweets are forming a graph structured network of messages on Twitter.

Additionally, tweets can be flagged with so called *hashtags*. Preceded by a "hash" symbol (#), a hashtag is a keyword assigned to information that describes a tweet. When looking for tweets of a particular topic, hashtags aid in searching, since related tweets are likely to be tagged with the same hashtags. With 50 million tweets per day being posted on Twitter, hashtags are central to organizing information. Hashtags organize discussion around specific topics or events.

2.2 Politics on Twitter

Twitter is said to be reshaping politics. The modern literature suggests that Twitter is a democratic media because it allows for instant reporting of breaking news and democratic activism. The political aspects of Twitter are under research and getting more attention in the research community. This work was performed with the focus on the party leader debates, but politicians, election and citizen opinions are trending topics as well.

Hashtag Political hashtags are gaining attention in the research community because they offer an easy way to cluster tweets into themes based on their meaning. They came to prominence in events like the 2009 Iran presidential election. #iranelection was the number one news topic on Twitter in 2009 [13]. The greatest value of a political hashtags is that it relates summarized or associated information to the tweet.

Election The influence of Twitter in national elections is a common topic of interest for political scientists. There is a study about microblogging under the 2010 Swedish elections. User types were identified in order to model opinions [40]. In another study, the forecasting of the 2009 German election output using Twitter was found to accurately reflect the election outcome [41]. The correlation of the 2011 Spanish presidential election result and Twitter was also a topic of interest [42].

Politicians The way politicians use Twitter is another leading question for political scientists. The association between the political candidates salience and the engagement level of the candidate in Twitter was measured for the impact on the 2012 US elections [43] with the result that high level social activity on Twitter does not result in increased online public attention. In another study the model of how politicians use Twitter was used to determine whether Twitter is used as a tool for communication or deliberation [44]. The results showed that politicians used Twitter to communicate with fellow politicians more than to have discussions with their opponents. The 2010 US congress members were analyzed for whether Twitter is used to disperse information [45]. The article concluded that Twitter is mostly used for self-promotion by the congress members.

Network of Actors There is also interest in the network of political connections on Twitter for extracting information about roles and relations of the social media users and politicians. The communication between Twitter users who take tutelage roles in order to provide help for information seeking users was modeled and analyzed in a research [49] which also modeled the elite/non-elite interactions of Twitter users with political interests. Finally, the way Australian political journalism has evolved around Twitter has also been studied by political scientists [50].

Cognitive Aspects The psychological aspects of Twitter communication is also of interest. A study found correlation between high interactivity and the sense of direct conversation with political candidates [51]. The same research group also studied the Twitter users with weak party identification and their candidate evaluations [52]. Modeling political polarization on Twitter is an upcoming topic in political sciences. The US congress and its Twitter followers were measured for polarization, in a recent research [53]. The outcome of the work indicated that most citizens with political interests might not be polarized, except a minority that has fanatic interest in politics.

2.3 Swedish Politics

Sweden is a parliamentary representative democratic constitutional monarchy. Parlamentarism got introduced in Sweden in the first decade of the 20th century. The Diet

of the Four Estates was replaced by a parliament in 1865, but only an economic elite could take part in elections. The elected parliament in Sweden is called Riksdag and it is led by a Prime Minister [34]. The universal right to vote was introduced in 1907 as part of voting reform, however women were still not allowed to vote. In the early stage of Sweden's democracy most influential political parties were the following: Social Democrats, the Conservatives and the Liberals.

During the 1980s and 1990s, the party structure changed: in 1988, the Green Party was the first additional party that got into the parliament since 1921. In 1991, the Christian Democrats managed to cross the 4% barrier and get into the parliament. From 1994, the parliament consisted of these parties: Sveriges socialdemokratiska arbetareparti (S), Moderata samlingspartiet (M), Miljöpartiet de Gröna (MP), Folkpartiet Liberalerna (FP), Centerpartiet (C), Sverigedemokraterna (SD), Kristdemokraterna (KD) and Vänsterpartiet (V). Since the 2010, some of the largest parties outside of the parliament are Feministiskt initiativ (FI), Sveriges Pensionärers Intresseparti (spi), Junilistan (JI) and the Piratpartiet (PP). Parties registered for the general election are also applicable for local and municipal elections across the country for the European elections. Historically, five to eight parties are represented in the Parliament. In order to form the government, the parties congregate in two political blocks called coalitions along the left-right scale.

The four fundamental laws are: Instrument of Government (since 1974), Act of Succession (since 1809), Freedom of the Press Act (since 1766) and Fundamental Law on Freedom of Expression (since 1991) [35].

Sweden's parliamentary parties have had a high levels of membership, but in recent years, the party system has stagnated because voters' mobility has increased, and a host small parties have been added as a choice. The media is of a great importance for agenda initiation, control of the agenda and in determining what is considered to be a social problem for the party leader debate. Because of its power, media is sometimes considered as the fourth estate.

2.3.1 Party Leader Debate

The party leader debate gives the party leaders the chance to publicly argue about the policies they want to pursue. There are three debates per year, hosted by the Agenda Swedish TV program broadcast on the public TV station *SVT*. Agenda is in charge of the topics that will be discussed by the party leaders under the debate. These topics will be referred to in this paper as *agenda topics*. The parliamentary year's first party leader debate takes place in October when the general exercise period has expired, but party leader debates are also held around January and in June.

Some of the rules of the party leader debate are: each speaker has the right to a statement of no more than ten minutes, the Prime Minister launches the debate followed by parties in magnitude order, for each speech there is the right to reply for the notified speakers, reply time is more than two minutes for the first reply and not more than one minute for the next one, the speeches and the replies are held in the pulpits in front of the podium.

2.4 Topic Modeling

The idea of topic modeling emerged from the need of searching for scientific articles in large digitalized collections, which are not indexed. It can be used for identifying articles that are similar to those of interest [15].

Manual indexing is not always a possibility because of the large amount and the growth of the collection. Topic modeling can be also applied to other documents that contain a mixture of topics in a similar way as articles. Topic modeling is an automated statistical method for discovering underlying topics in order to organize, manage and provide documents based on their content. Some of the first topic modeling methods are Probabilistic Latent Semantic Indexing and Latent Dirichlet Allocation.

2.4.1 Probabilistic Modeling

A probabilistic model describes the structure and relation of random variables. The probabilistic nature of the variables relation makes a model probabilistic. In a generative probabilistic model the observed variables are suspected to come from a generative process that includes hidden variables and creates a joint random distribution over the observed and hidden variables. The posterior distribution is gained by calculating the conditional distribution of the hidden variables assuming the observed variables are then retrieved from the joint distribution.

2.4.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) [15] is a generative probabilistic model of a corpus. The computation of inferring the latent topic structure from documents is the task of computing the posterior distribution, which is the conditional distribution of the hidden variables given the documents [68].

The basic idea is that documents are represented as random mixtures over hidden topics, where each topic is a distribution over words. The goal of topic modeling is automatically to discover the topics from a collection of documents. Its concept is easily captured by its generative process: the imaginary random process during which the documents are created.

LDA is the simplest of so called "bag of words" topic models. The main difference between LDA and other topic models like for example LSI is that in LDA documents exhibit multiple topics. A topic is formally defined to be a distribution over a fixed vocabulary, while a vocabulary is the set of unique words. The documents are the only observed variables. The per-topic word distributions, per-document topic, the per-document topic word choices are all hidden variables. The way that the algorithm works is, it uses the observed variables, which are the documents and infer from them all the hidden variables. The inference is the reversing of the generative process, or in other words finding the most probable hidden variable values that are generating the observed variable.

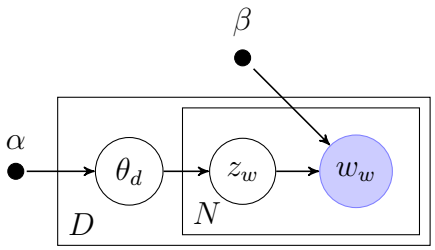

Figure 1: Plate Diagram of LDA [26].

α is the Dirichlet prior parameter on the per-document topic distributions,
β is the parameter of the Dirichlet prior on the per-topic word distribution,
θ_d is the topic distribution for document d,
z_w is the topic for the word w in document d, and
w_w is the specific word.

Generative Process: See [15].

2.4.3 Probabilistic Latent Semantic Indexing

Probabilistic Latent Semantic Indexing (PLSI) is an improved version of Latent Semantic Indexing (LSI) [24] and a method for automatically indexing and retrieving information. It maps documents and terms to a latent semantic space. It applies a linear projection to reduce the dimension of the document vector space representation based on the frequencies of the terms.

1: **for** document d_d in corpus D **do**
2: Choose $\theta_d \sim$ Dirichlet(α)
3: **for** position w in d_d **do**
4: Choose a topic $z_w \sim$ Multinomial(θ_d)
5: Choose a word w_w from $p(w_w|z_w, \beta)$, a multinomial distribution over words conditioned on the topic and the prior β.
6: **end for**
7: **end for**

2.5 Domain Specific Topic Modeling

Domain specific topic modeling refers to the domain or theme of the text the topic model is applied to. There are various aspects of domains on textual data. The language the text is written in can be considered to be a domain of the text. However, English is widely spread and, therefore, not considered as a domain, but Swedish text, for example, is. The theme of the content of the textual data is often viewed as a domain. For example, data from Rotten Tomatoes has a movie review domain and tweets connected to a party leader debate have the party leader debate domain. In many cases, documents such as news articles, product reviews and Tweets belong to a specific domains as genetics, computer science or literature. In contrast, the data from Twitter is not considered to have the Twitter domain since Twitter is a type of corpus and not a characteristic of the text itself.

When looking at Twitter data many themes, and many languages can be found in the non-filtered stream. Researchers from various fields are interested in how their subjects of research are reflected on Twitter, therefore domain specific models need to be created in order to extract the relevant information from the data.

A domain topic model is specific for a domain and, as a result, it provides answers to hypothesis defined by a domain expert. General topic models are often not able to capture the information that is required in order to answer the research questions. The use of topic models to analyze domain-specific texts often requires manual validation from a domain expert of the latent topics to ensure that they are meaningful.

2.6 Term Frequency – Inverse Document Frequency

Term frequency-inverse document frequency (tf-idf) is a statistical method to determine how important a word or term is to a document in a collection [30]. It is commonly for

analyzing textual data. The tf-idf weight is the normalized term frequency which is the term frequency times inverse document frequency. The term frequency is the number of times a term appears in a document. The inverse document frequency is the logarithm of the number of the documents in the corpus divided by the number of documents where the specific term appears. Let t be a term, d be a document and D be all documents then:

$$tfidf(t, d, D) = tf(t, D)$$

Where the term frequency and the inverse document frequency are calculated the following way:

$$tf(t, d) = frequency(t, d)$$

$$idf(t, D) = log\frac{N}{|\{d \in D : t \in d\}|}$$

The tf-idf increases proportionally to the number of times a term appears in the document and decreases by the frequency of the word in the corpus. This captures the idea that some words are more common than others and hence less important. Tf-idf is used for finding words which can act as unique identifiers of documents.

2.7 Classification

Machine learning as a sub-field of Artificial Intelligence, defines methods that learn some of their parameters from a provided input. The parameters are set based on the information extracted from the input and are used to retrieve information from data similar to the input in the future.

An example of classification is email spam detection: given a number of emails labeled as spam and not-spam, the classification method learns the relevant features of spam emails and is then able to process new email messages to mark them as spam or not-spam. The previous example is a supervised statistical classification. The method is supervised because it first needs to be trained on labeled data, where the expected output of the method on the data is known, as opposed to an unsupervised method where the output of the training data is not available and that is why other techniques are applied. A supervised learning algorithm that performs classification is called classifier. The classifier is first trained with data, where each item is labeled with the known output. This data is used to train the learning algorithm, which models the data and the model can be used to classify similar data. In contrast to this behavior the unsupervised algorithm has unlabeled training data which it tries cluster in groups based on some similarity measure.

2.7.1 Sparse Text Models in NLP

In information retrieval, natural language processing and some machine learning contexts work with large spaces of words or n-grams. In this case, the text is often represented as "a bag of its words", disregarding grammar and the word order while keeping the frequency of words. We refer to this as *bag-of-words* representation. A common model for this representation is a vector space model or term vector model. Algebraic models will be used describing the representation of texts as vectors of identifiers. In the following formula j is the jth document, d stands for document and $w_{i,j}$ is the frequency of ith term in the jth document.

$$d_j = (w_{1,j}, w_{2,j}, ..., w_{t,j})$$

In this model a document is a point in a high-dimensional space. d_j is a sparse vector in which the vector entries correspond to the terms that the document contains [36]. When creating this corpus model the terms are assigned a coordinate in the order in which that word appears in the document. For instance, the sentence "I prefer sunny weather to rainy weather" after tokenization, stopword filtering, word-frequency counting will be modeled as the vector $[1, 2, 1]$, where the ordered dimensions correspond to sunny, weather, rainy.

2.7.2 Sparse Matrix Representation

When a corpus consisting of documents is modeled using a vector space model, the result is a large sparse matrix whose size depends on the number of distinct words in the corpus. It is computationally expensive to store and perform operations on sparse matrixes, which is why programming languages usually implement sparse matrix types in order to make the algorithms using them run faster.

Python has a module called Scipy Sparse that contain different implementations of sparse data representations. This module also supports the usage on relevant linear algebraic methods on the representations. One of the most common implementation is the dictionary of keys (DOK). In DOK implementation dictionary mapping represents the non-zero values as (row, column)-pair keys to values. This implementation is efficient in construction of sparse data but not efficient in looping over non-zero values. DOK represents non-zero values as a dictionary mapping (row, column)-pairs to values.

2.8 Support Vector Machine

The following introduction is based on Andrew Ng's lecture notes [55]. The Support Vector Machine (SVM) is a supervised learning algorithm used for classification. It

is one of the most commonly used classifiers. For example in Kaggle [37], a platform for predictive modeling and analytics competitions, it is often the first choice for a classification model as an "off-the-shelf" classifier.

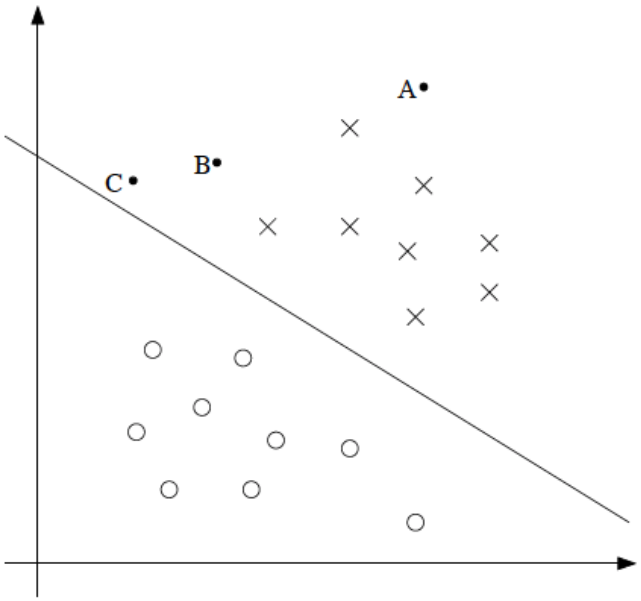

Figure 2: Intuition for SVM separation [55]

An intuition for how SVM works is the following example, on *figure* **??** the sign X represents positive training samples and O represents negative training samples of a two-dimensional dataset. The job of the classifier is to draw an imaginary line called separating hyperplane that would separate the positive training samples from the negative ones. On the figure 2 the separating hyperplane is a linear line in the form of $\theta^T x = 0$ where $x = [x_1, x_2]$, all points for which $\theta^T x < 0$ is true will be predicted as negative training data and points for which $\theta^T x > 0$ is true will be predicted as positive training samples. point A is expected to be predicted as a positive training sample since it is very far from the decision boundary. The model will predict a positive label to point C since it lies on the positive side of the decision boundary, however compared to point A, it is much closer to the decision boundary. If a small change is made to the decision boundary, like rotation in a point near point C or translation to the separating hyperplane in the positive direction, the predicted label of point C would change to negative. Point B lies between the points A and C and, therefore, the confidence of predicting point B's label would be more than point C and less than point A.

For the definition of the classification problem the following notation will be used for

defining the classifier:

$$h_{w,b}(x) = g(w^T + b) \tag{1}$$

The parameters w and b are used instead of θ and the function g will transfer the output to of the classifier to 1 and -1. $g(z) = 1$ if $z \geq 0$, otherwise $g(z) = -1$. Let $\gamma^{(i)}$ define the geometric margin which is the distance of each training sample $x^{(i)}$ to the decision boundary. γ will be the smallest geometric margin of all in respect to the training samples:

$$\gamma = \min_{i=1,\ldots,m} \gamma^{(i)} \tag{2}$$

The optimization problem which defines the classifier is the following:

$$\min_{\gamma,w,b} \frac{1}{2}||w||^2 \tag{3}$$

having $y^{(i)}(w^T x^{(i)} + b) \geq 1$ where i is a training sample [55].

2.9 Stochastic Gradient Descent

The following introduction is based on Andrew Ng's lecture notes [56]. For simplicity the gradient descent will be introduced on logistical regression instead of SVM. For logistic regression the classifier will look as follows:

$$h(x) = \sum_{i=0}^{n} \theta_i x_i = \theta^T x \tag{4}$$

where x_i are the training samples and θ is the parameter of the classifier also called as weight.

The problem will be presented as an optimization function like in the previous chapter, a cost function will be defined as follows:

$$J(\theta) = \frac{1}{2} \sum_{i=1}^{m} (h_\theta(c_{(i)}) - y_{(i)})^2 \tag{5}$$

The task is to minimize the $J(\theta)$ by choosing a θ. So the optimization problem is as follows:

$$\min_{\theta} \frac{1}{2} \sum_{i=1}^{m} (h_\theta(c_{(i)}) - y_{(i)})^2 \tag{6}$$

In order to solve this minimization problem, the gradient descent algorithm will keep changing the θ parameter in order to minimize the $J(\theta)$. The gradient of the function $J(\theta)$ represents the magnitude of the slope of the tangent that is drawn at each value of the θ. The θ parameter will be updated in a loop with the gradient in order to get closer to the minimum:

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta) \tag{7}$$

$$\theta_j := \theta_j + \alpha (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)} \tag{8}$$

The α corresponds to the learning rate of the update. Setting α influences the convergence of the update to the global minimum. If α is too small, the update will not converge to the global minimum because changes to θ will be small after each update. On the other hand, having a large value of α might also cause the update not to converge to the global minimum, because the updates might overshoot the optimal θ.
The algorithm for gradient descent will look as follows:

`repeat until` θ `changes:`

$$\theta_j := \theta_j + \alpha \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)})) x_j^{(i)} \forall j \tag{9}$$

The function $J(\theta)$ is quadratic and therefore convex. That means it has one global minimum. If the learning rate is set up correctly, the algorithm will always converge. A plot of gradian descent algorithm running and converging is shown on figure 3

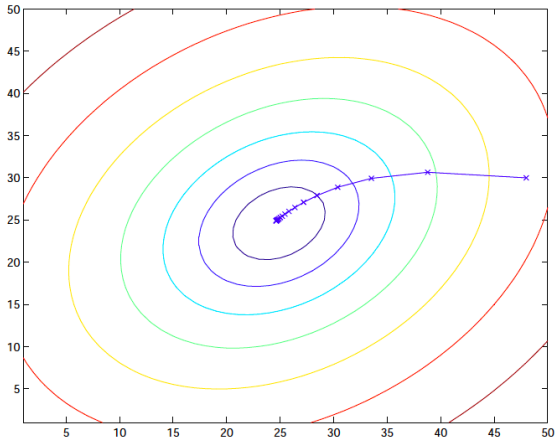

Figure 3: Intuition for SVM separation [56]

The drawback of the presented gradient descent algorithm, also called a batch gradient descent, is that in each step of the loop the update uses all training data. Another version of stochastic gradient descent which has better time performance is *stochastic gradient descent*. The algorithm follows:

```
repeat until converges or step limit reached:
  for i=1 to m:
```

$$\theta_j := \theta_j + \alpha(y^{(i)} - h_\theta(x^{(i)}))x_j^{(i)} : \forall j \tag{10}$$

Instead of updating all parameters using the whole training set, stochastic gradient descent updates each parameter with the gradient of the cost function with respect to one training sample. This makes stochastic gradient descent updating parameters more often and having a half-way solution faster while the algorithm is still running. This is a useful feature when dealing with large amounts of data, and that is why stochastic gradient descent is preferable to batch gradient descent. On the other hand, stochastic gradient descent might never converge to the global minimum. It might oscillate around it giving an estimate of the global minimum. This estimate is reasonably good for using it as the solution.

The convergence of stochastic gradient descent can be improved by decreasing the learning rate α while the algorithm is still running.

3 Data

The data collection has been performed by Språkbanken [5] using filters for selecting the data of interest from Twitter. After gathering, the data was annotated using Språkbankens tools and lexical information was added to the tweets. We built an infrastructure for loading and storing the data from Språkbanken in order to make further processing convenient.

3.1 Collection

Twitter has several APIs for making their data accessible. The most popular API for data scientists is the Streaming API for gathering public statuses in near real-time. The data provided by the Streaming API is sampled or filtered by track keyword or username. The Streaming API consists of three specific APIs, User Stream and Site Stream and Public Streaming API which was used in this research. The dataset of all tweets tweeted is called the Twitter Firehouse. It is expensive to gain access to the Firehouse data. However, 1% of the Firehouse data is called Sprinkle and is available for anyone.

The goal of Språkbanken was to gather all Swedish tweets during a 5 day period using the Public Streaming API. Identifying the language in which the tweets is written is not as a simple task as it seems. Twitter has high language diversity; a study has shown that 10 million tweets contain 65 languages [63]. Twitter has added its own language filter in March 2013, but the quality of the filter output is discussable. Since Twitter messages are relatively short, they provide only a few words for classifying the corresponding tweets language, making it a hard task. The register of languages on Twitter has a rather informal style: mostly conversational, including slang, figurative language, acronyms, incomplete sentences, symbols and abbreviations. These characteristics make standard NLP tools perform poorly on Twitter. Wide range of lexical variations of in-vocabulary (IV) words are making the dictionary based methods less efficient. Languages appear with different dialects; non-native speakers write many tweets. This adds to the difficulty of language detection. Finally, the limited labeled data available at this time is not enough for training classifiers from the data. Moreover, Twitter specific language usage styles just add to the problem: it is not uncommon to switch language in the middle of a tweet and non-lexical tokens like URLs, hashtags, and usernames are not necessary language specific and, therefore, should be handled with care.

The overall accuracy of detecting the language of a formal text is 99.4% [64], however the accuracy for microblogs is 89.5% [65]. Two recent methods achieved the best result for now: a model enhanced with semi-supervised priors [64] that overcomes the short message problem and is remarkably good on the 5 languages it was trained on, Swedish is not one of them. However mixed language tweets are still an issue for this model and

vast amount of data is need to train the models, which is not available for all languages.

Språkbanken used a simple language detection: the most frequent Swedish words were gathered to a list, all those tweets that contain at least one word from the list were labeled as Swedish. This method has lower accuracy then the methods mentioned above, but it fits to the task. The common NLP tasks like tokenization and part-of-speech tagging were not used during topic modeling so there was no need for a high accuracy language detector.

Språkbanken used a list of most common Swedish words [5] to filter out Swedish tweets using the Public Stream API. The amount of tweets gathered using this filter does not exceed 1% of all tweets, therefore Twitter does not include limitations to the results: all tweets matching the filter criteria were gathered [3]. Tweets were gathered two days ahead and three days after the debate: from 4th of October to 8th of October 2013.

3.1.1 Annotation

After gathering the data and aggregating it by user, Språkbanken performed lexical annotation and added the extracted information to the data.

Their first task was to define and separate sentences in tweets, which is called sentence segmentation. Languages like English and Swedish use punctuation, so the full stop character is a good indication of the end of a sentence. But even in these languages the problem is not so simple, since the the full stop character is also used for abbreviations.

In order to tokenize the tweets Språkbanken divided parts of continuous text into separate words. For Swedish, this is simple, because words are usually separated by space characters. However, some written languages like Chinese and Korean do not mark word boundaries with space characters and, therefore, complex algorithms using vocabulary and morphology of words in the language is used. There are some user behaviors on Twitter that make tokenization difficult even for Swedish. Words are often not separated by a space character; they are simply concatenated together. In some cases CamelCase [66] is used where each separate word starts with a character. This behavior appears on Twitter because the 140 character limit encourages users to leave out unimportant characters in order to compress more information in the tweet. Words part of non-lexical tokens like hashtags and URLs add to the problem since they cannot be separated by space characters.

The annotating information comes from parts-of-speech tagging. This is performed by determining the part of speech for each word in a sentence of a tweet. Common words can serve as multiple parts of speech. The word “flies”, for example, can be a verb

("Time flies like an arrow.") or a noun ("Fruit flies like a banana"); "set" can be a noun, verb or adjective; and "out" can be any of five different parts of speech. Some languages have more such ambiguity than others. Languages with little inflectional morphology, such as Swedish are especially prone to such ambiguity. In contrary, languages like Chinese are not so ambiguous due to inflectional morphology.

3.1.2 Data Description

Even though, the data was collected using Twitter Public Stream API, during the lexical annotation performed by Språkbanken the structure of the data changed in order for the lexical information to fit in the data. We received the data and lexical information from Språkbanken in a specific format which differs from the Twitter common data format.

The main components of the data are Twitter users. All those users who have at least one tweet that was labeled as a Swedish tweet will be in the data as a user. *Figure 4* shows an example tweet in JSON format. MongoDB has a unique identifier with the key `_id` for each JSON element which is used by MongoDB. The same identification value is available for usage with the `id` name. Further keys connected to the user element are `username` which has for value the Twitter username of the user, the `name` has for value the name that the user gave when creating the account, `created` JSON key has for value the date when the users account was created, `following` has for value the number of other users that are followed by the user, `followers` has for value the number of other users following the user, `tweets` has for value the total number of tweets the user has tweeted since being on Twitter, `description` has for value a textual description the user gave for the account and `text` has for value a list of elements that represent the tweets.


```
{ _id : 1000007083,
  username : felicianansi,
  name : Felicia Nanasi ♡,
  created : 2012-12-09,
  following : 58,
  followers : 115,
  tweets : 478,
  id : 1000007083,
  description : its just me ♡,
  text : [ ]
}
```


Figure 4: User element example in JSON format

The element that represents one tweet in the data looks like the following a*fugure 5*, and it is an embedded in the user element described above. The JSON key `dateto` has the value of the date the tweet was sent to Twitter from the user's device, `datefrom`

has the value of the date the tweet was received by Twitter, `datetime` has for value the date and time the tweet arrived at Twitter, the JSON key `hashtags` has a value of all hashtags contained in the tweet, separated by the character `|`. In case the tweet has no hashtags the `hashtags` key will have the value `|`. Furthermore, the JSON key `retweets` has for value the number of times the tweet has been retweeted, `replies` has for value the ids of other tweets that are replies of the tweet separated by `|` having the empty value `|`, `mentions` has for value the users who were mentioned in the tweet separated by `|` having the empty value `|`, the JSON key `id` has for value a unique id given to the tweet by Twitter, `weekday` has for value `true` if the tweet was sent on a weekday otherwise it has the value `false` and finally the `sentence` JSON key is itself a list of JSON objects that represent a sentence from the tweet it is embedded to. In the case, the tweet is a reply of another tweet there is a key `replytostatus` that has the value of the other tweet id that it replies to. If the tweet is not a reply to another tweet, this key will not be present as seen on figure *figure 5*.

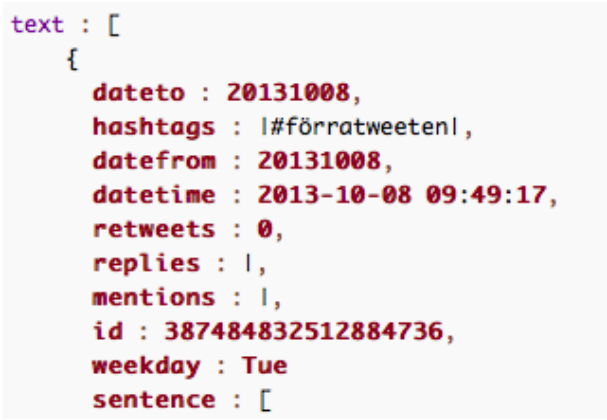

Figure 5: Text element example in JSON format

The element sentence seen on *figure 6* is simpler compared to the previous elements. It consists of the JSON key `id` which is a unique id of the sentence and key `w` which has a list of words from the sentence as a value. The exact word contained in the tweet is stored under the key `val`, *figure 6*. The other keys like `suffix`, `prefix` and `lemma` represent lexical annotation of the word under `val` in the context of the whole sentence.

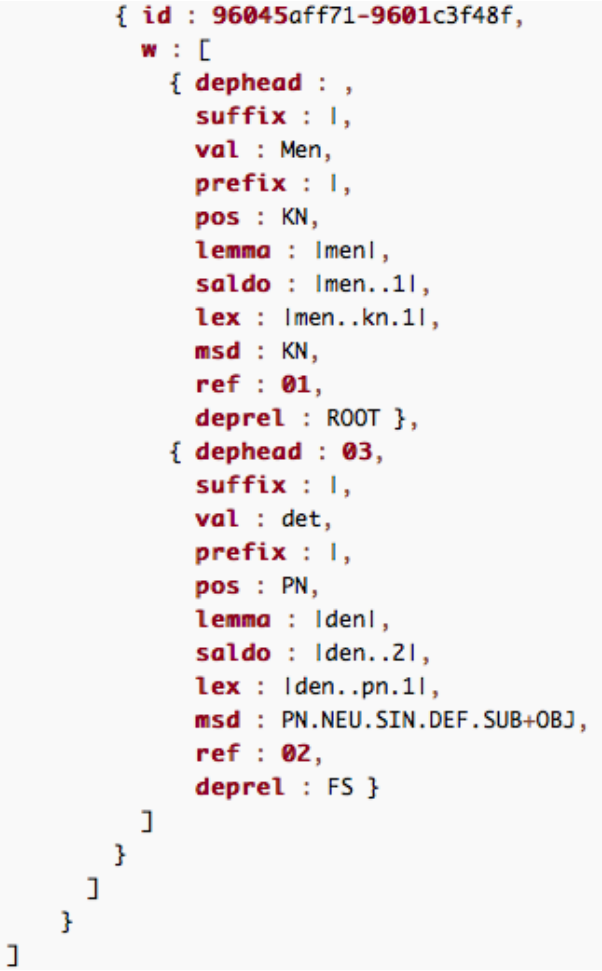

Figure 6: Sentence element example in JSON format

The data described above is nested. It has four levels: the user level, the tweet level with `text` key, the sentence level with `sentence` key and the word level with `w` key. The main part the data, is the textual part of the tweets that lies on the fourth layer of the data under the key `val`. We made the use of the `lemma` of the words in this research.

4 Pre-processing

We pre-processed the data received from Språkbanken. The data was partially processed and lexically annotated by Språkbanken. We performed further pre-processing in order to remove the unnecessary information from the data and adjust its format to the one required by the algorithms that were applied on it. Different pre-processing for the algorithms in the unsupervised approach and for the semi-supervised approach.

4.1 Pre-processing for the Unsupervised Algorithms

Common words do not add to the value if topics. Before applying LDA to the data, we removed all common words from the data. Primarily we excluded the common Swedish words, but because Swedish tweets often contain English words, the most common English words were also considered for exclusion. The most frequent Swedish words, received from Språkbanken [5] and very frequent English words from Mallet English stop-word list were removed from each tweet.

In order to filter out links, and special characters from the data, Mallet's regular expressions were used.

```
'[\p{L}\p{M}]+'
```

This is a standard regular expression form for non-English textual data, that means the words can be formed of Unicode letters and marks.

The mentions and hashtags contained in the tweets are separated from their first special character in the data. They, therefore, act as valid words. They are expected to add information to the tweet about which topic might it be related to. Mentions, which are usernames, in contrast add little value for the topic determination and are removed from the tweets.

4.2 Pre-processing for the Semi-supervised Algorithms

In order to be able to compare words in their different lexical forms stemming or lemmatization can be used on the words. Stemming is the process when the word is reduced to its root form. For example *farming* and *farmer* will be reduced to *farm*. Lemmatization is the process where the word is converted to its grammatical lemma form. For example *better* and *best* will have the lemma *good*.

Usually stemming is being done in the pre-processing phase of text processing in English language [29]. If available lemmas of the words are preferred to the stems in Swedish since there is less collision of distinct words [28].

5 Methodology

The research questions of this work are: to what extent do party leader debate tweets appear in the twitter stream debate, how represented are the agenda topics on twitter and finally what other topics, outside of the agenda topics, do Twitter users discuss.

In order to answer the first research question, the number of party leader debated tweets were measured against all Swedish tweets from the data. The proportion of the party leader debate tweets and all tweets presents the amount of party leader debate tweets in the Twitter stream. The classifier used to answer this question will be introduced under the *Classification subsection 5.1*.

The second research question states: to what extent are individual agenda topics from the party leader debate discussed on Twitter. We implemented a semi-supervised approach to model the data and answer this research question, *subsection 5.2*. The model consists of a pipeline of algorithms that were applied on the data in order to identify the tweets related to each agenda topic. Counting the tweets related to the agenda topics the model will provide a proportion of tweets which discuss each agenda topic on Twitter.

Finally, for answering the last research question that states: what other unknown topics do originate from Twitter. We applied an unsupervised approach in order to model topics in the data, where topics here differ from the known agenda topics, *subsection 5.3*. Our solution consists of a pipeline of algorithms for filtering data, aggregating data and building a probabilistic topic model. The unsupervised approach identifies topic clusters that form in the tweets based on word concurrence in single tweets. These clusters are later validated with the domain expert and the topics relevant to the clusters are identified.

5.1 Hashtag Classification

The algorithm for the classification of the *#pldebatt* hashtag was used for expanding the *#pldebatt* context. The core data of the *#pldebatt* context are all tweets that have the hashtag *#pldebatt*. The classification algorithm expanded the data by classifying tweets without hashtags as *#pldebatt* tweets. We used this algorithm for answering the first research question about how big part of the Swedish tweets will be related to the party leader debate. The extension of the *#pldebatt* context also improved the model used in the semi supervised approach for identifying and measuring the agenda topics.

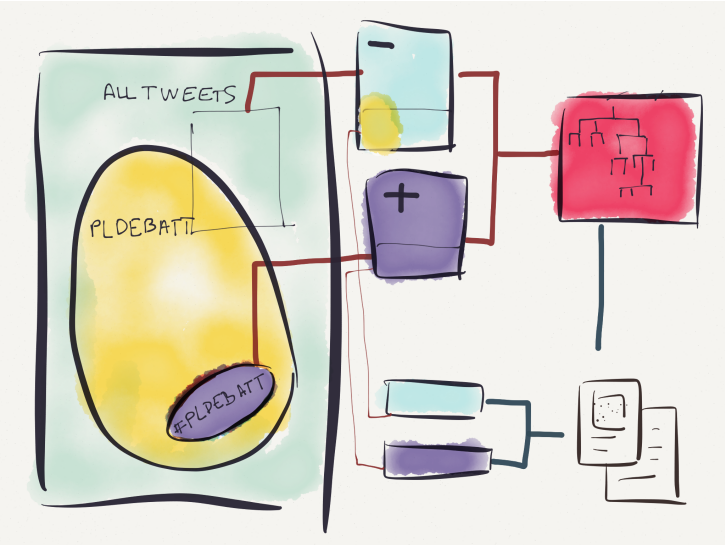

Figure 7: Steps in training the *#pldebatt* classifier: The yellow bubble represents all tweets that are related to the *#pldebatt* but do not contain the hashtag. The aim is to use the tweets in the purple bubble (that do contain the *#pldebatt* hashtag), to train a classifier that can correctly find the tweets in the yellow bubble.

In the data, we had 27657 tweets which contained the hashtag *#pldebatt*. Tweets that are containing the hashtag do not model well the tweets that are related to the party leader debate. It is a common practice in Twitter for users to omit using the hashtag. This can save character space in the tweet, so the users are able to tweet faster or to decrease information redundancy in their personal tweet stream. However, when these tweets are observed individually the context will be missing, and the relation of that tweet to a hashtag might be unclear.

In order to catch tweets related to the party leader debate which have no hashtag, shows as the yellow bubble in *figure 7*, we trained a classifier for deciding if a tweet with no hashtag is related to the party leader debate or not. The linear SVM classifier was used trained by stochastic gradient descent. The regularization term and hyperparameters were chosen using cross-validation. The python Scikit Learn library was used for the classifier described in detail in *appendix C* using parameters: hinge loss and $L2$ regularization and hyperparameters $\alpha = 0.00001$ and the learning rate of 0.1.

5.2 Identifying Agenda Topics

To model the data appearance of the six topics *healthcare*, *crime&punishment*, *refugees*, *job&tax*, *school* and *climate* on Twitter, a semi supervised approach was used. The model of the semi-supervised approach defined the agenda topics on Twitter and iden-

tified the tweets corresponding to the agenda topics. This was done using a pipeline of algorithms and manual processes. The first step was to manually create a list of words that correspond to the agenda topics. We refer to this list as *initial topic list*. Since the agenda topics are not precisely defined it was not possible to use the domain knowledge in the model. To be able to do that a list of words for each agenda topic was created which represents the corresponding agenda topic without a direct link to Twitter. This list is referred to as initial topic list.

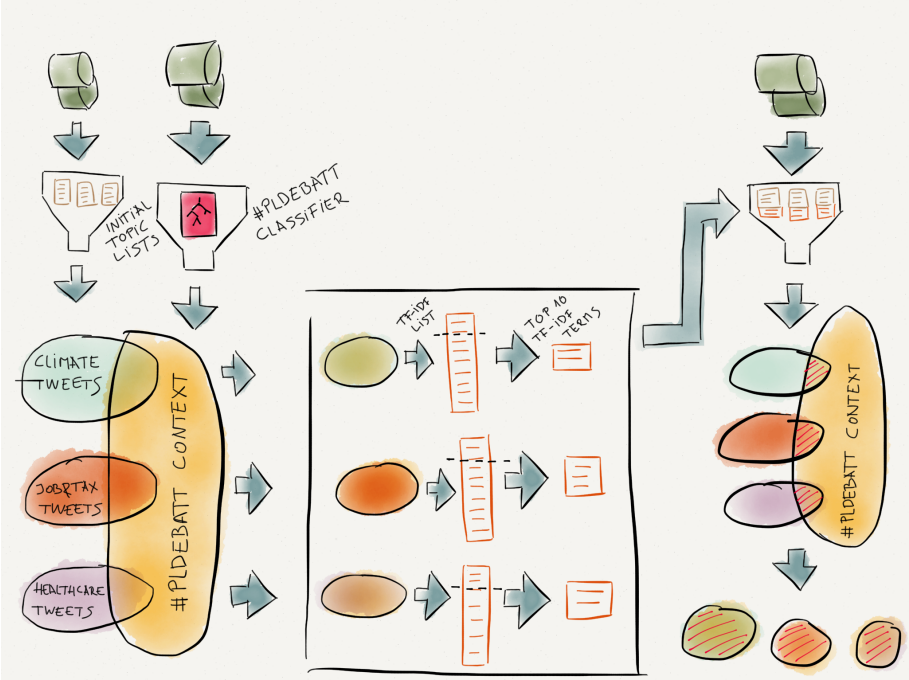

Figure 8: The pipeline of filters and algorithms of the semi-supervised model

5.2.1 Initial Topic List

The initial topic lists in the semi-supervised model include a list of words and are built manually according to the information from the TV broadcast of the party leader debate. These lists contain the words in their lemma form. Karp [5] was used to improve the lists with synonyms and related words. Finally the Government's (*Riksdagen's*) webpage was used to gather more words related to the agenda topic but which might have not appeared in the party leader debate. For some topics like *job&tax* it was very easy to find relevant words. For other topics like *climate*, fewer representative terms were found. In order to measure those topics on Twitter without introducing bias, we limited the length of the initial topic lists to the shortest topic list for which we gathered representative words. After gathering representative words we ended up having around 150 words per

list. Some examples from the initial topic lists can be found in *table 1.* For the complete initial topic list see *appendix B.*

Healthcare	Crime	Tax	School	Immigration	Climate
sjukvård	rån	jobb	skola	flykting	klimat
akutvård	brott	avdrag	klass	flykt	klimatkris
vårdtid	straff	skatt	pisa	irak	utsläpp
remiss	häkte	bolag	tenta	asyl	reaktor
carema	tjuv	firma	läsår	asylrätt	miljöfråga
doktor	olaglig	fas3	komvux	krig	grön
landsting	snut	bidrag	lärare	visum	kärnkraft
kötid	sexbrott	koncern	rektor	syrien	vindkraft
apotek	svindel	arbeta	kunskap	amnesti	miljöbil
diagnos	bov	pension	betyg	asylprocess	växthusgaser

Table 1: 10 of the words representing the initial topic list

5.2.2 Tf-idf Topic Extraction

In the second step of the semi-supervised pipeline the initial topic lists were enriched with domain specific words gathered from Twitter.

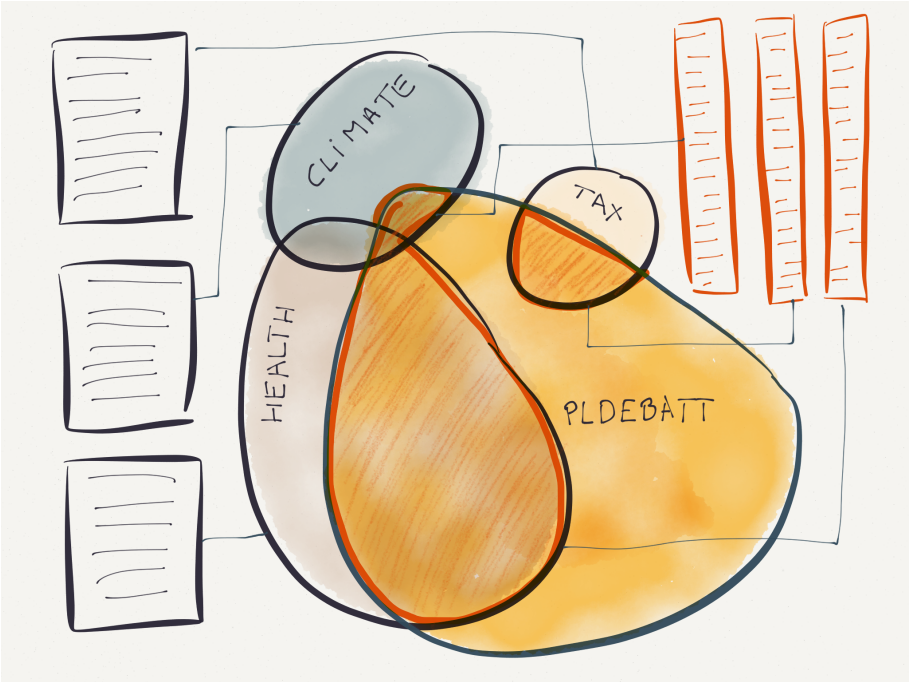

Figure 9: Selecting tweets based on the initial list and calculating the sorted list of terms and tf-idf values

The intuition is the following; the initial topics lists might not be sufficient to capture agenda topics because they were created without Twitter in mind. Different words could be used on Twitter in agenda discussions than in the official areas like TV broadcast and Riksdagen website. In order to solve this issue, the Twitter data was mined for Twitter specific representative words of the party leader debate, which might catch the possibly different context the party leader debate is discussed in on Twitter. The tf-idf measurement was used to identify relevant words for each agenda topic. This was done by grouping all tweets to a document that contain at least one word of the initial topic list. For each agenda topic, one document was formed. The 10 highest tf-idf words for each document were added to the initial topic list creating an extended initial topic list. The domain expert chose which words were added to the initial topic list. In some cases, noisy words appeared in the top to tf-idf terms and these were, therefore, discarded and not added to the initial topic list. To have 10 words selected from the top tf-idf valued words was chosen because the smaller the tf-idf value is the more noisy the words are and therefore add less value to the model.

Having the extended initial topic lists referred to as the *topic lists*, all tweets that contain at least one word from this list will be considered to discuss the corresponding agenda topic. The model at this step will give a number of tweets that discussed each agenda topic.

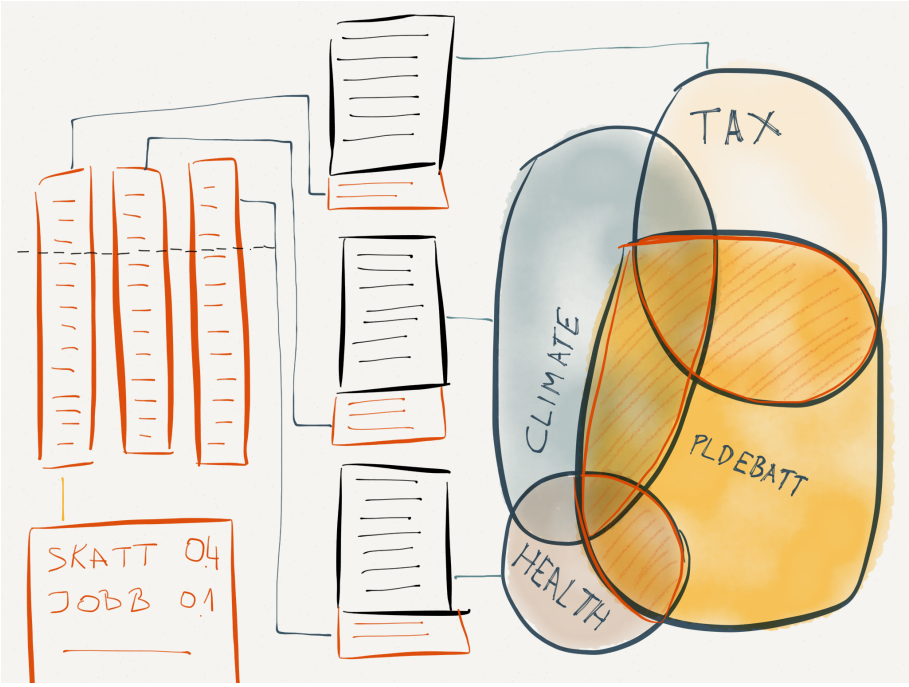

Figure 10: Updating the initial topic list with the top 10 tf-idf values. Selecting the agenda topic tweets based on the updated topic list.

5.3 Finding Latent Topics

We applied an unsupervised approach in order to answer the third research question which is to discover unknown topics, that users discuss on Twitter. In order to do that the unsupervised approach will model topics discussed on Twitter. Topic modeling is not directly applicable to Twitter. The intuition would be to take the tweets as articles or documents and apply topic modeling to them [16]. The advantage of using LDA is to model documents as a mixture of topics, so short documents, tweets containing maximum 140 characters do not perform well with LDA. Modified version of LDA are more suitable for Twitter topic modeling.

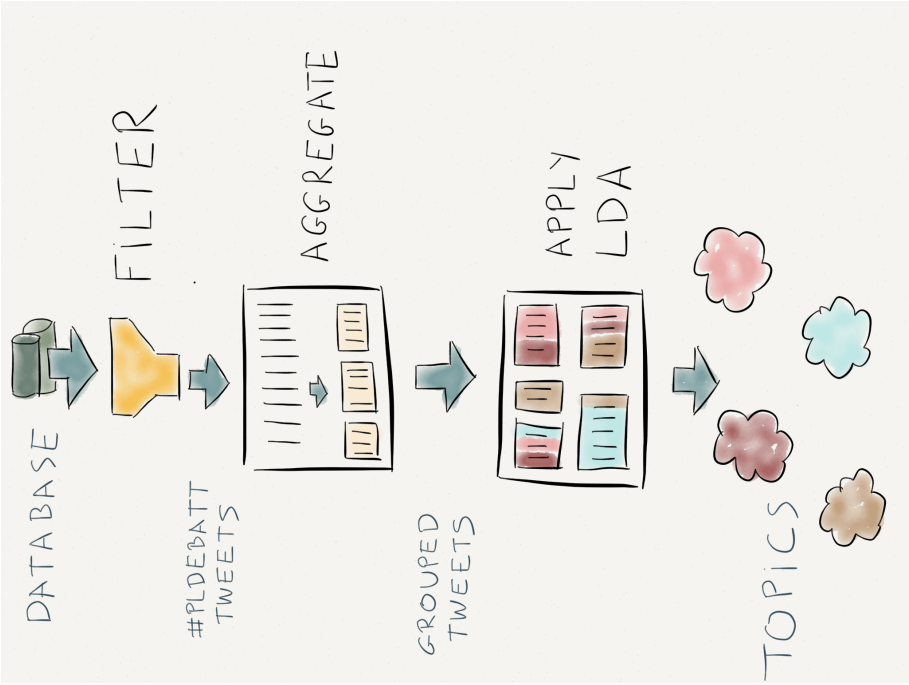

Figure 11: The pipeline of filters and algorithms of the unsupervised model

5.3.1 Filtering

The data contained all Swedish tweets in a time frame so in order to get domain specific tweets from the model a filtering on the party leader debate domain was performed. All tweets are filtered using *#pldebatt* as a label for a tweet related to the political debate. The domain experts choose *#pldebatt* as the most suitable hashtag for discovering the political debate related tweets. The fact that the tweets are gathered tightly around the time of the debate guarantees that the hashtag refers to the actual event and not to some future or past debates. There was also a consideration about the *#svpol* hashtag, which is the most famous political hashtag used in the data. Based on the domain experts knowledge this hashtag turned to have a suitable use of the political debate.

After performing semi-supervised topic modeling described in *subsection 5.2*, LDA was applied on all tweets assigned to one Agenda topic in order to find finer grained topic within, e.g,. a discussion on the teachers education within the *school* topic.

5.3.2 Aggregation

The modified versions of LDA where each document corresponds to more than one tweet, perform better on Twitter [23]. The difference in applying LDA on tweets and articles would be to aggregate the tweets with some aggregation method and use the

text of aggregated tweets as a document. This work implemented tweet aggregations by author, hashtag, hashtag cluster and replies. The aggregations were manually evaluated.

#pldebatt skapa polarisera tweets derbysupportrar . total ointressant läsa ni subjektiv tweets .
@emanuelkarlsten ja , överhetta . och aggressiv . #pldebatt #twitter
@emanuelkarlsten - tappa sting ... läsa rad #pldebatt
@emanuelkarlsten en objektiv ärlig subjektiv . de bry #pldebatt twitter övertygad .
#pldebatt skapa polarisera tweets derbysupportrar . total ointressant läsa ni subjektiv tweets . @emanuelkarlsten ja , överhetta . och aggressiv . #pldebatt #twitter @emanuelkarlsten - tappa sting ... läsa rad #pldebatt @emanuelkarlsten en objektiv ärlig subjektiv . de bry #pldebatt twitter

Table 2: Example of reply aggregation. The grey text is the aggregated document

sluta beskriva hälsa sjukvård beståläkare sjuksköterska @goranhagglund #fysioterapi #pldebatt
tänka hälsa sjukvård , sjukvård #fysioterapi #pldebatt
#pldebatt gammal övermedicinerade undertränade . använda #fysioterapi #StefanLoefven @asaromson @Jonas_Sjostedt `http://t.co/dnPNqgWo8X`
sluta beskriva hälsa sjukvård beståläkare sjuksköterska @goranhagglund #fysioterapi #pldebatt tänka hälsa sjukvård , sjukvård #fysioterapi #pldebatt #pldebatt gammal övermedicinerade undertränade . använda #fysioterapi #StefanLoefven @asaromson @Jonas_Sjostedt `http://t.co/dnPNqgWo8X`

Table 3: Example of hashtag aggregation. The grey text is the aggregated document based on the #fysioterapi hashtag

Table 4: Example of user aggregation. The grey text is the aggregated document

RT @Einerstam : redan nu kunna jag säga att åsa romson vinna debatt och den säga jag i egenskap av oberoende socialdemokrat #pldebatt
RT @Schandorff : självmål av annie lööf #pldebatt `http://t.co/yzgG2vVWYQ` ” den vara synd att skola ha bli en politisk slagträ i debatt ” . bara man kunna yttra såmeningsläs påstående . al #pldebatt
äntligen klimatdebatt #pldebatt
RT @Einerstam : redan nu kunna jag säga att åsa romson vinna debatt och den säga jag i egenskap av oberoende socialdemokrat #pldebatt RT @Schandorff : självmål av annie lööf #pldebatt `http://t.co/yzgG2vVWYQ` ” den vara synd att skola ha bli en politisk slagträ i debatt ” . bara man kunna yttra såmeningsläs påstående . al #pldebatt äntligen klimatdebatt #pldebatt `http://t.co/dnPNqgWo8X`

Aggregation by author goes through all users and takes all tweets from that user and concatenates them to one document, see *table 4* on page 37. We use these documents in LDA. Similarly in hashtag aggregation tweets are concatenated by having the same hashtag. If the tweet has multiple hashtags the aggregation adds it multiple times to different document, see *table 3* on page 36. Because of the limited tweets size, the number of hashtags found in a tweet is also limited. The tweet is going to be added multiple times only to limited amounts of documents, it cannot happen that a tweet gets added to a large percent of all documents. Based on this fact this study assumes that having some tweets multiple times, in different documents, will not have a noticeable impact on the topics found by LDA. The reply aggregation places a tweet and all its replies in the same document, see *table 2* on page 36. Finally, the hashtag cluster aggregation takes all tweets having the same hashtag or co-occurring with the same hashtag into one document.

After manual evaluation, it was concluded that user aggregation performed best. However all aggregations failed to create larger document sizes. Some larger documents did appear, but majority of the documents contained only one tweet after the aggregation. For the user aggregation that means most users tweeted only once. For hashtag cluster aggregation created one large document from a super cluster of co-occurring tweets and all other document remained one-tweet documents.

6 Results

The following chapter will describe the results of this study that includes both successes and failed attempts. The results will include the unsuccessful application of the unsupervised method on the data in order to retrieve the agenda topics as well as the success of retrieving new unknown topics from the data using the same method. The section will further include the second attempt of extracting the agenda topics using the semi-supervised method on the data as well the use of classification in order to improve the semi-supervised methods accuracy.

6.1 Hashtag Classification

Three models were considered based on the length of the tweets seen on *table 5*. The first model trained on tweets of all length, the second considered only tweets containing 6 or more words and the third model considered tweets containing 11 or more words.

Model	all tweets	6 words or longer tweets	11 words or longer tweets
Training samples	11 709	10 568	8 057
Test samples	3 042	2 777	1 927

Table 5

The model evaluations were made using the *F1 score*, that is the harmonic mean of the precision and the recall. The results of the evaluation can bee seen on *tables 6, 7,* and *8*. The best performing is the model with the training data restricted to 11 words tweets shown on *table 7*. This model also managed to predict 52023 tweets with no hashtag to be connected to the party leader debate.

	precision	recall	f1-score	support
other	0.89	0.89	0.89	1650
pldebatt	0.87	0.87	0.87	1392

Out of 876 923 tweets without hashtags 107 414 were predicted to have #pldebatt.

Table 6: The model applied to all tweets

	precision	recall	f1-score	support
other	0.92	0.88	0.90	1504
pldebatt	0.87	0.91	0.89	1273
Out of 718 539 tweets without hashtags 83 916 were predicted to have #pldebatt.				

Table 7: The model applied to 6 words or longer tweets

	precision	recall	f1-score	support
other	0.91	0.89	0.90	945
pldebatt	0.90	0.91	0.90	982
Out of 427 308 tweets without hashtags 52 023 were predicted to have #pldebatt.				

Table 8: The model applied to 11 words or longer tweets

Based on this outcome the *#pldebatt context* is defined as all tweets containing the hashtag *#pldebatt* and all tweets predicted to have the hashtag *#pldebatt*. This way we calculated $27657 + 52023 = 79680$ tweets related to the party leader debate.
The semi-supervised approach used the *#pldebatt* context instead of just hashtag in order to improve the results of the semi-supervised model which was used for answering the second research question of how much are agenda topics discussed on Twitter.

Having the *#pldebatt* context we are also able to answer the first research question of how big part of the data stream does the party leader debate take. $4,694\%$ of the tweets in the data belong to the *#pldebatt* context.

6.2 Identifying Agenda Topics

In order to compare the agenda topic discussions, we calculated the amount of tweets connected to the agenda topics using the topic lists. *Figure 12* displays the number of tweets connected to each agenda topic. All tweets that contained at least one word from the topic list were considered to be connected to the corresponding agenda topic. *Figure 12* is not a distribution of tweets since a tweet can be connected to more than one agenda topic.

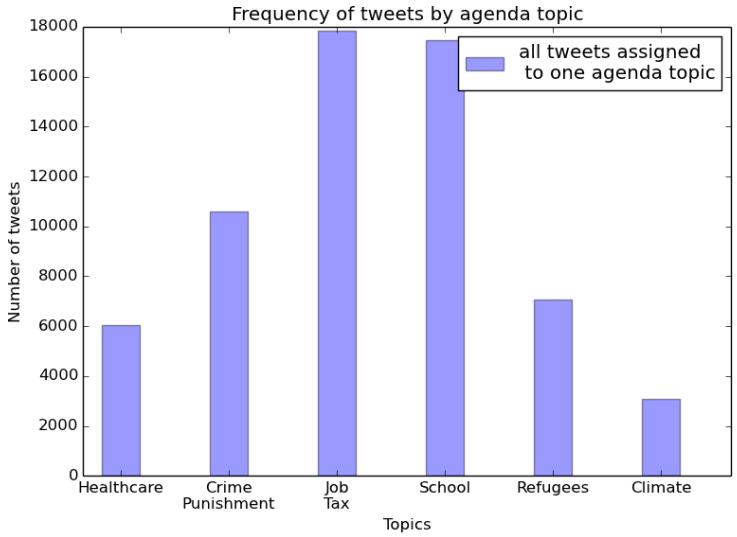

Figure 12: Tweet frequency per agenda topic for all Swedish tweets

In order to get the number of agenda topic tweets in relation to the party leader debate we measured the overlap of the agenda topic tweets with the *#pldebatt* context. *Figure 13* shows the number of tweets connected to the agenda topics which overlap with the *#pldebatt* context.

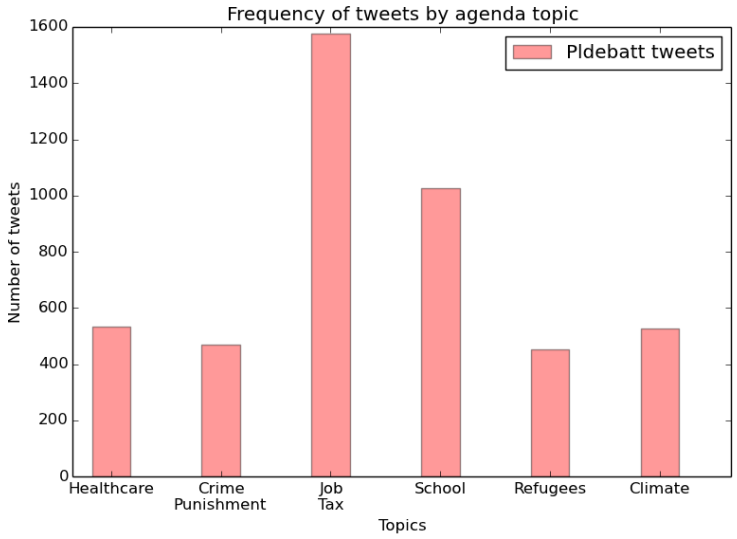

Figure 13: Tweet frequency per agenda topic for all party leader debate context tweets

Figure 12 and *figure 13* shows the effect of taking the intersection of the agenda topics, with the semi-supervised method. The proportions of the number of tweets for each

topic was mostly preserved, except for the *school* agenda topic. The reason for this could be that the time of the debate was October, one month after the school year starts. Many non-political tweets could fit to the agenda topic *school*. After we took the intersection with the *#pldebatt* context, the non-political *school tweets* were dropped out.

6.3 Finding Latent Topics

The goal of the unsupervised approach is to answer the second research question about what other materials appear in the Twitter discussion. After we applied LDA on the data, the outcome was latent topics represented as a distribution of unique words. In order to find the right model, we performed an evaluation on the available models.

While comparing the models with different document aggregation methods, the aggregation by users gave the best result. Manual evaluation was made by the domain expert of all the aggregation techniques, and user aggregation gave the most reasonable outcome. The results of the manual evaluation were not surprising since the user aggregation method also created the smallest number of LDA documents, in other words it left the least tweets not aggregated, compared to other methods. Having stand-alone tweets as document for the LDA model is not desirable since it leads to less accurate topics [16].

Another hyperparameter that was considered is the number of topics. The number of topics was also evaluated manually by the domain expert. The best fitting topic number can be retrieved when trying out different topic numbers and observing the LDA topics [54]. If the LDA topic tends to be a mixture of different topics or theme that could be interpreted as too few topics. In contrast when more than one LDA topic tends to describe one manually interpreted topic or the topics describe more objects and items, the number of topics could be interpreted as too high. The number of topics should be set to be between the two described cases.

The latent topics that the LDA outputs are distributions of unique words from the corpus. In order to visualize them we took the top 10 words of each words which have the highest probability of occurring in that topic. *Table 9* shows numbered LDA topics using user aggregation and set topic number to 10. This model was discarded as it does not describe the topics of the data well because of the following reasons: topic number 5 seems to be a mixture of *job&tax* and *healthcare* topics and topics 1, 3, 4, 9 and 10 are describing a general theme of the party leader debate instead of a topic. This model was interpreted as having too small number of topics hyper parameter value.

Topic #	Representative words for the topic
1	björklund jan reinfeldts partiledardebatten björklunds studion politiken https saknar
2	debatten debatt kvällens diagram tv attackerade vann vinnare fuizbm
3	reinfeldt prata sanningen politik partiledardebatt flexibel varandra inställning partiledarna
4	svpol agenda löfven reinfeldt fredrik miljarder alliansen svtagenda oppositionen
5	alliansen sverige val sd jobb vanligt fram sjukvård besked
6	reinfeldt hatar arbetslöshet regeringen skattesänkningar löfven pratar utanförskap statsminister
7	skolan romson björklund skola åsa betyg partiledare behövs fall
8	löfven stefan mnniskor skatter politik pengar löfvens högre valet
9	åkesson sjöstedt jimmie politik jonas twitter göran ner ord
10	lööf annie svt romson svara löfven politiker mp lyfter

Table 9: Topics found by LDA with number of topics set to 10 and user aggregation

The next value for the number of topics hyper parameter was 20 seen on *table 10*. This model performed visible better than the model described above: new topics emerged which were not present in the previous model like topic number 5 which is about the hair style of a politician named Björklund and topic number 11 about the refugees agenda topic. However, there are still problems with this model since some agenda topics still seem to share one LDA topic like topic number 6 which is about agenda topics *job&tax* and *healthcare* and a discussed agenda topic *climate* did not appear.

Topic #	Representative words for the topic
1	björklund högre debatten lön avbryter tjänar sagt håret tiden
2	agenda svpol reinfeldt alliansen fredrik oppositionen jobb sverige rödgröna
3	diagram prata partiledarna attackerade twitter ordet fuizbm själva lär
4	debatten bort talar låt pldebatten ner partierna jimmy riksdagen
5	björklund jan hägglund göran björklunds pldebatt land chockbeskatta frisyr
6	hatar miljarder arbetslöshet politik löfven sjukvården skattesänkningar jobb sjukvård
7	lööf annie romson åsa pratar hägglund självmål jävla plakat
8	alliansen vanligt skatter sänkt fram hittills vilket valet partiledare
9	politik pratar fram siffror skatt klarar lila snälla slips
10	reinfeldt löfven val statsminister sanningen flexibel fredrik inställning hört
11	sverige sd svenska mindre behövs flyktingar europa asyl söka
12	frågan partiledare varandra politiker ord egna prata utbildning fp
13	löfven stefan besked valfläsk skatten köper tydligt kvällens vinnare
14	reinfeldt reinfeldts parti https partier lova rösta politiska resten
15	svt kvällens partiledardebatten tv debatten studion partiledardebatt plats paus
16	svpol svtagenda vann kd mp dn bort frågar debatten
17	åkesson sjöstedt jimmie jonas människor slut affär snacka åkessons
18	debatt skolan svensk saknar politiken problem fall illa klokt
19	mp alliansen löfvn lyfter svara annie landsbygden centerpartiet debatten
20	skolan skola lärare betyg människor läxhjälp fas sveriges tänk

Table 10: Topics found by LDA with number of topics set to 20 and user aggregation

The models using the number of topics hyperparameter value 50 and 100 were considered next. The model with the topic number 50 is shown in the *appendix E*. It does not describe the data reasonably and is discarded because of the following reasons:

many agenda topics are distributed over numerous LDA topics *job&tax* appears in topics 2,11,12 and 14, the school agenda topics appears in topics 4, 17, 20 and 22, climate agenda topic appears in topics 19, 28 and 50. Finally the discussion about Björklunds funny hair surfaces in topics 16 and 9. This model is considered as having too high value for the number of topics hyperparameter. The model having the number of topics hyperparameter value set to 100 was also evaluated with similar results as the model with number of topics set to 50.

The model that was finally chosen to be the best LDA model of the topics in the data has the number of topics hyperparameter value set to 30 and is shown on *table 11*. This model still has some unwanted characteristics: the school agenda topic is distributed over LDA topics 13, 16, 19 and some topics which define agenda topics school and *job&tax* like topic number 20. However, the domain expert considered it as a middle ground between models having number of topics 20 and 50.

Topic #	Representative words for the topic
1	reinfeldts studion https pldebatten lär löfvens håret talade eepwalkl
2	löfven stefan köper valfläsk monopolpengar kvällens koll använder statsministern
3	reinfeldt val sanningen flexibel inställning rösta synd ljuger partier
4	partiledare vann twitter debatt egna bort politiker rödgröna omröstningen
5	politik skattesänkningar politiska resten arbetslösheten dålig väljer hånleende satsar
6	reinfeldt löfven skriker avbryta vågar programledare statsministern sverige hänt
7	hägglund sjöstedt politik göran regeringens jonas pratar leder persson
8	tv fram twitter ner sämst parti dessa klara roll
9	sverige pldebatt människor vanligt unga hit klimat välfärden kärnkraft
10	diagram attackerade ordet partiledarna fuizbm sagt gårdagens utrymme minst
11	reinfeldt alliansen löfven oppositionen besked fredrik bort tydligt debatt
12	alliansen mp ansvar klart miljöpartiet siffror tydligt läsa långt
13	björklund jan själva chockbeskatta rör politiker flumskolan ståflumskola
14	romson åsa jämställdhet heja snyggt frågan behövs prata samtycke
15	svpol svtagenda sverige budget miljarder mp taggad vinnare ökat
16	skolan skola lärare betyg svensk hört val resultat härskartekniker
17	björklunds plakat frisyr tänk politisk saknar pratar vart vilde
18	hatar arbetslöshet löfven utanförskap regering arbetslösa politiskt fas retorik
19	svpol agenda fredrik media kd expressen pol synd innehåll
20	jobb skatter skolan pengar människor skatt sänkt jobben jobba
21	prata miljarder regeringen sjukvården vården pratar frågan alliansens fram
22	sverige sjukvård flyktingar eu fall europa asyl fp mp
23	sjöstedt högre jonas lön självmål sjuksköterskor råd vvwyq yzgg
24	debatten debatt varandra brott straff ord utbildning paus ökar
25	björklund sjöstedt svenska läxhjälp skolan partiledardebatt jävla mindre låt
26	åkesson jimmie sd dn invandring åkessons svälja invandrare invandringen
27	reinfeldt debatten lova klarar hägglund talar bron moderaterna debatter
28	lööf annie lyfter centerpartiet romson landsbygden pratar vindkraft ner
29	svt kvällens partiledardebatt partiledardebatten plats debatten analys tv spännande
30	reinfeldt vanligt statsminister löfvn svara valet partiledardebatten snälla avbryter

Table 11: Topics found by LDA with number of topics set to 30 and user aggregation

For interpretations of the topics found by the 30 topic model see *table 12*.

- Topic number 4 is a discussion about who, Twitter users think had most the successful discussion among the party leaders in the debate. According to the top ten representative words for that topic from *table 11* no specific party leader comes up as a winner.
- Topic number 5 seems to discuss the relation of tax reduction and unemployment. Due to the correlation between the tax reduction and unemployment, the discussion of weather unemployment will decrease or increase when relevant taxes are reduced.
- Topic number 9 is connected to the issue of Sweden's use on nuclear power plants. Compared to western countries like Germany and Denmark, Sweden is highly dependent on nuclear energy. The effect of the nuclear waste on the climate is a hot topic in Sweden.

Topic #	Topic description	Noise level
1	-	high
2	-	high
3	-	high
4	Who won the party leader debate?	medium
5	Tax reduction and unemployment	**low**
6	-	high
7	-	high
8	-	high
9	Climate and the use of nuclear power plants	**low**
10	Attack diagram	**low**
11	-	high
12	-	high
13	-	high
14	Lets talk about equality! - The Green Party	**low**
15	-	high
16	Decreasing school results	medium
17	Björklunds hairstyle	**low**
18	Unemployment	**low**
19	-	high
20	Job and tax	**low**
21	-	high
22	The effect of refugees on Europe	medium
23	Higher salaries in the healthcare	medium
24	The effect of education on crime	medium
25	School and homework	medium
26	Immigration	**low**
27	TV series Bron	medium
28	Wind turbines	**low**
29	-	high
30	-	high

Table 12: Topic interpretation for 30 topics

With the help of the domain expert, Linn Sandberg, we evaluated and interpreted some of the less noisy topics of the model.

- Topic number 10 is about an attack diagram, *figure 14*, which was created by a political scientist Anders Sundell [61] who was calculating during the party leader debate how many times to party leaders criticizing each other. After the debate Sundell drew a diagram and posted it on Twitter. Besides being entertaining this diagram shows insight about the party leaders debate strategy and their relation to each other in the aspect of responses to each others words during the party leader debate. The attack diagram went viral among Swedish Tweets and therefore came up as the least noisy topic in our model.

- Topic number 14 brings up the question of equality even though it was not part of the debate. It was brought up by the Green Party as seen on *table 11* under topic number 14.

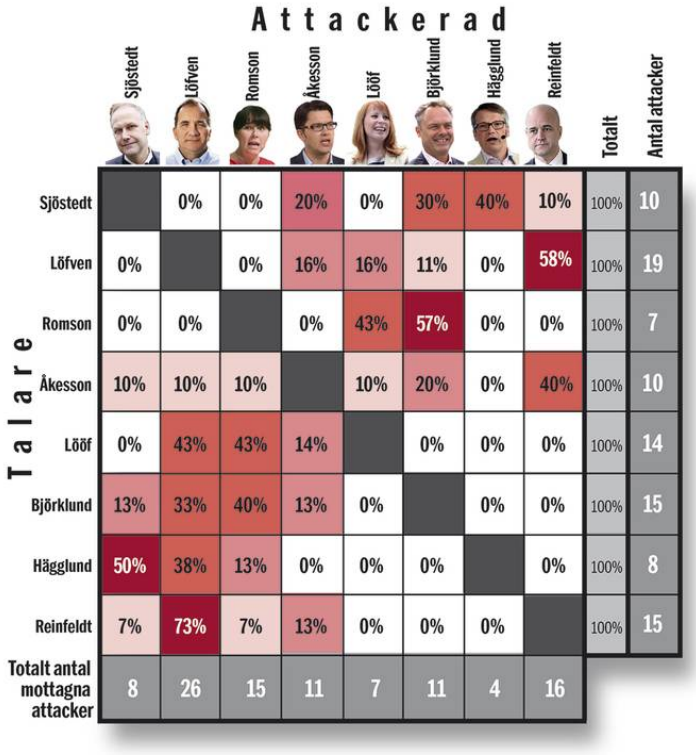

Figure 14: Attack diagram created by Anders Sundgren

- Topic number 16 discusses the problem of decreasing school results. The outcome of the PISA test [32] showed that Swedish pupils do worse in schools than some years ago. There is ongoing discussion about what approach should be taken by the politicians in order to tackle the issue.
- Topic number 17 is about the hairstyle of FP party leader Jan Björklund which was considered to be funny by the Twitter users who were following the debate. Even other politicians tweeted about it.
- Topic number 18 is about unemployment. Unemployment does not appear as a agenda topic itself but its closely related to *job&tax*, because *job&tax* discusses creation of new jobs.
- Topic number 20 corresponds to the agenda topic *job&tax*. This is the only agenda topic which came up as an LDA topic.
- Topic number 22 is about the influence of refugees on Europe. The war in Syria was going on at the time, the number of refugees has increased from previous years in EU countries like Italy and Greece and there were also fatal accidents with the refugee ships capsizing [33].

- Topic number 23 is the only topic that came up connected to the healthcare in the model and it seems to be specifically about the issue of increasing salaries for healthcare workers.

- Topic number 24 could be interpreted as a mixture of education from the *school* topic and *crime&punishment* topic. Based on the broadcasted party leader debate the topic is more probably about the roll of education in crime and punishment, and this would be seen as the agenda *crime&punishment* topic.

- Topic number 25 is about assisting pupils with homework as part of the school agenda topic. Discussion is going on about how much homework should be given to pupils and also how it should be distributed.

- Topic number 26 is about immigration that is related to the refugees agenda topic. Instead of focusing on the international issue of refugees this topic discusses immigration into Sweden. This has been a sensitive topic in Sweden.

- Topic number 27 is a discussion about a popular TV series *Bron* which was broadcasted in the same time as the party leader debate on a different channel. Twitter users were discussing whether to switch to watching *Bron* or the party leader debate.

- And finally topic number 28 is about wind turbines and part of the agenda topic *climate*. There are ongoing discussions in Sweden about switching to clean energy and wind turbine are playing a large role in this idea.

This model is expected to answer the research questions about what other party leader debate related discussion come up on Twitter, which are not the agenda topics. Based on the topics the unsupervised model provides it can be said that in two days range around the October party leader debate in Sweden the additional political issues, which were not part of the agenda topics, Twitter users were discussing were the lack of equality discussion represented as topic number 14 and immigration topic number 26 seen on table *table 11*. The other non-agenda party leader debate discussions which appeared on Twitter were the attach diagram topic number 10, Björklunds hairstyle topic number 17 and TV series *Bron* topic number 27 seen on table *table 11*.

7 Discussions

Unexpected topics came up in the result when we ran LDA topic modeling on the data. We presumed that LDA would find other pressing political themes which were not present at the party leader debate. Some of these topics are feminism, equality and racism. To our surprise the strongest topics were not related to politics, in fact they were related to the TV broadcast of the debate. The attack diagram, seen on *figure 14* is the most consistent topic that was found by the LDA. It was the result of models having various topic numbers and it was not as noisy as the other topics. The next strong non-political topic was Björklunds hair. Beside these discussions, some political non-agenda topics did appeared in the LDA result with lower topic quality: equality and immigration.

Agenda topics were not represented in the LDA results, besides *job&tax*. The reason for the lack of the other agenda topics could be that they were created by the Agenda program, in contrast LDA topics were formed based on how Twitter users use words in their tweets. Based on how LDA models topics, users do not seem to use agenda topics the same way. Words representing the agenda topics did appear in the LDA outcome, but they did not form a coherent topics. In order to test this assumption we were looking for the agenda topics in LDA results while changing the number of topics hyperparameter. If the agenda topics would form a LDA topic, they might emerge when we were modeling a fine grained topic setup. However this did not happen. With a large number of topics some agenda sub-topics were surfacing; the large the number of topics hyperparameter was, the finer sub-topics showed up. We concluded that LDA topic models cannot capture well topics defined by people.

We concluded, that LDA cannot capture well the topics that Agenda defined, but it did capture the topics that the people discussed and found interesting. It could be that people do not care about the whole topics of *school*, but they have interests in specific subtopics like help with homework. Therefore, that is also what we see in the results.

The resulting amount of tweets related to the agenda topic was confirmed by the domain expert, Linn Sandberg. After performing sample tests, the domain expert concluded that the agenda topic *job&tax* does seem to be the largest agenda topic on Twitter, followed by *school*. Further interpretation of the results is not part of this work.

The hashtag classification results that we achieved are highly comparable with similar works. A study has trained a classifier on hashtags related to broad topics like *celebrity*, *games*, *music* and *political* [69]. The accuracy of the classifier in the mentioned study was up to 80%. In different circumstances our classifier achieved more then 90% accuracy. However the classifier we use is a step in the data process pipeline. As errors accumulate in data techniques connected in serial, it is hard to say how big impact does the 10%

error rate have on the final result of the semi-supervised approach. In this research we use the classifier to gather more data, so a classifier with even higher accuracy would be valued.

8 Conclusion

Social media is taking politics to a digitalized age. Discussions between voters and politicians, new political campaigning possibilities and information network of political parties are hosted on social media platforms. Analyses of political discussions online are of a great interest to politicians as well as political scientists. However, due to the number of users the data from social media is too large for conventional analysis methods. Machine learning, statistics and natural language processing methods provide suitable methods for the data, which can be used for performing further analysis directed at answering political research questions.

In this work, we presented the use of machine learning and natural language processing techniques in order to model political topics on Twitter. The models targeted Swedish party leader debate and was focusing on providing quantitative values for answering the research questions. The methods we use in this work measure the party leader debate discussion amount in the Twitter flow, in order to determine the extent of the Twitter discussions. The topics discussed on the party leader debate by the politicians on the TV broadcast are identified and measured on Twitter. These topics are also called agenda topics and the measurements of their magnitude are compared to the attention these agenda topics got on the TV broadcast and from the politicians, in the ongoing political study.

In this work, we concluded that $4,694\%$ of the Swedish Twitter stream data is connected to the party leader debate. The Twitter discussions about the party leader debate do take a noticeable space in the Twitter user stream. The proportion of the magnitude of the agenda topics discussed on Twitter was proportional to the amount of tweets the domain expert classified as agenda topic related. Beside the agenda topics, other topics emerged in the Twitter discussions which are reactions to the party leader debate TV program and other political issues. We therefore accept the Null hypothesis stated in *section 1.2*.

8.1 Limitation and Future Work

This section will describe the known limitation of this work that include limitations on the techniques as well as evaluation. The future work or improvements of this study will be made after this thesis is presented.

8.1.1 Identifying Agenda Topics

In the first step of the semi-supervised approach, all tweets that contain at least one word from the initial topic list are selected as tweets connected to the agenda topic corresponding to the initial topic list. In order to do a more accurate matching, the lemmas of the words in the tweets are used for the matching to the words on the initial list which are mostly in their lemma form. However, this method does not handle out-of-vocabulary (OOV) terms. The observation of the domain expert was that OOVs in the Swedish party leader debate context tend to be concatenations of in vocabulary (IV) words. To exploit this observation, the prefix and suffix of the OOV could be used for creating the lemma of the OOV. This lemma could then be used to match against the words on the initial topic lists. This is not always simple since the suffix and prefix of the word might be connected with a stand alone letter which will lead to incorrect lexical splitting of the prefix and suffix. For example the word *sjukvårdslandstingsråd* meaning medical councelor is not in the Karp lexicon, however separately the words *sjukvård* and *landstingsråd* are in the lexicon.

The evaluation of this method does not include a comparison of this method to other semi-supervised approaches. LSI is one of the widely used methods for topic modeling with supervised steering of the parameter weights. In future LSI should be run on the data and its result should be evaluated and compared to the semi-supervised method used in this study.

After sorting the terms based on their *tf-idf* in the documents created by the initial list, the top 10 related terms are manually selected and added to the initial list. This selection could be automated if evaluation of the *tf-idf* list shows high accuracy of relevant terms appearing on the top of the list.

Hashtag Classification The classification was trained by selecting positive training data and negative training data. The positive training data was selected using *#pldebatt* hashtag: all tweets that contained *#pldebatt* were selected for positive training samples. Negative training samples were tweets that had no hashtag and did not contain any of the words from the politics list (*appendix D*). Selecting negative training data can be improved by adding terms to the politics list and choosing words from tweets based on their lemmas rather than on the words themselves.

8.1.2 Finding Latent Topics

For creating LDA documents, other aggregation methods can be considered as well. For example, the process that aggregates Tweets a by time bucket might outperform the

used aggregation methods. This method would not have the problem of leaving stand alone tweets as documents since for all tweets connected to the party leader debate it is very improbably that only one or zero tweets are contained in the time bucket. If that were the case, the time bucket could be increased.

As seen on *table 11* topic 23 contains 2 words which were considered to be noise since they are not meaningful Swedish words: `vvwyq` and `yzgg`. When the data was pre-processing no filtering on meaningful words was performed and therefore it is not surprising that such words are left in the data. However at first glance it seems unusual to have these words as one of top 10 words which represent a topic in the model. Because of this, these two words were part of a popular tweet which got retweeted many times and therefore they co-occurred with more meaningful words many times and got to be a high probability word of an LDA topic. To tackle this issue, future work should consider removing retweets. Removing retweets would lead to some important tweet being downgraded in importance for the model. For topic modeling, this is an undesirable behavior since words contained in a tweet that is being retweeted should have influence on all the documents which contain the retweets based on the aggregation process. A middle ground between removing the retweets and keeping them should be to curb the number of retweets down to the logarithm of the number of retweets. This would lead to partial removal of the retweets which will allow retweets to change the topics but if there is a tweet with very high number of retweets it will not form its own topic. The dilemma only remains about which retweets should be removed. If a random function would be used for removing retweets, its impact on the models output topics should be considered. For this study we kept all retweets as part of the data.

During evaluation, the models having number of topics 10, 20, 30 and 50 were evaluated. In order to check for a fine scaling of the number of topics parameter models with the number of topics of 25 and 35 were also manually evaluated, however the model outcome did not yield noticeable difference compared to the chosen model with the number of topic parameter set to 30. With more accurate evaluation method like having more domain experts of automatic evaluation, the hyperparameter number of topics value can be set more accurately.

In order to find latent topics, the LDA algorithm checks for word co-occurrence in Tweets. Words in different lexical forms will be treated as different words in LDA. The model can use the provided lemma of each word in the data and have less noise in the LDA topics. However, it is not necessary to use the lemmas of the words. Different lexical forms will end up in the same topic since they cooccure in similar LDA documents. The advantage of not using the lemma form of the words is that different lexical forms can be used in the algorithm. This way the topic outcome of the model is more descriptive.

As seen on *table 11* the LDA topics are represented with 10 words, which have the highest probability of occurrence in the corresponding topics. The models were also evaluated with 15 words, but they did not performe better than the models with 10 words. However deciding on how many words are representative for a topic is a common question in topic modeling. Beside deciding on a constant number of words with highest occurrence probability for displaying the topic, the number of top keywords representative of the topic can also be found with plotting the probabilities of words occurring in the topics in descending order.

In future work unsupervised models need to be evaluated automatically, and the results should be analyzed and compared to manual evaluation. The semi supervised model needs to be evaluated automatically and manually. Evaluation results need to be analyzed and compared.

References

[1] BrightPlantet *Twitter Firehouse vs. Twitter API: Whats the difference and why should you care?*, Retrieved February 27, 2014.

[2] Twitter FAQ *How are rate limits determined on the Streaming API?*, Retrieved February 27, 2014.

[3] Twitter Blog *Filtered stream - is it filtering out of 1% sample stream?*, Retrieved February 28, 2014.

[4] Ingrid Lunden *Twitter May Have 500M+ Users But Only 170M Are Active, 75% On Twitter?s Own Clients*, Retrieved February 27, 2014. http://spraakbanken.gu.se

[5] Språkbanken *Official website*, Retrieved February 28, 2014.

[6] Lotan, Gilad, et al. *The Arab Spring: the revolutions were tweeted: Information flows during the 2011 Tunisian and Egyptian revolutions.* International Journal of Communication 5 (2011): 31.

[7] Wired: Ben Austen *Public Enemies: Social Media Is Fueling Gang Wars in Chicago* Retrieved February 27, 2014.

[8] Tumasjan, Andranik, et al. *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* ICWSM 10 (2010): 178-185.

[9] Cheong, C. S. Lee, et al. *A microblogging-based approach to terrorism informatics: Exploration and chronicling civilian sentiment and response to terrorism events via Twitter* Information Systems Frontiers, 2011, Volume 13, Issue 1: 45-59.

[10] Pak, Alexander, and Patrick Paroubek. *Twitter as a Corpus for Sentiment Analysis and Opinion Mining.* LREC. 2010.

[11] Kumar, Akshi, and Teeja Mary Sebastian. *Sentiment Analysis on Twitter.* International Journal of Computer Science Issues (IJCSI) 9.4, 2012.

[12] Wijksgatan, Olof, and Lenz Furrer. *GU-MLT-LT: Sentiment Analysis of Short Messages using Linguistic Features and Stochastic Gradient Descent.* Atlanta, Georgia, USA (2013): 328.

[13] A. Smalla. *WHAT THE HASHTAG? A content analysis of Canadian politics on Twitter* Atlanta, Information, Communication & Society, Volume 14, Issue 6, 2011: 872-895

[14] Blei, David M., and John D. Lafferty. *Topic models.* Text mining: classification, clustering, and applications 10 (2009): 71.

[15] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. *Latent dirichlet allocation.* J. Mach. Learn. Res. 3 (March 2003), 993-1022.

[16] Liangjie Hong and Brian D. Davison. *Empirical study of topic modeling on Twitter.* In Proceedings of the First Workshop on Social Media Analytics (SOMA '10). ACM, New York, NY, USA, 80-88.

[17] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. *Comparing twitter and traditional media using topic models.* In Proceedings of the 33rd European conference on Advances in information retrieval (ECIR'11), Paul Clough, Colum Foley, Cathal Gurrin, Hyowon Lee, and Gareth J. F. Jones (Eds.). Springer-Verlag, Berlin, Heidelberg, 338-349.

[18] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. *TwitterRank: finding topic-sensitive influential twitterers.* In Proceedings of the third ACM international conference on Web search and data mining (WSDM '10). ACM, New York, NY, USA, 261-270.

[19] Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. *Characterizing Microblogs with Topic Models.* ICWSM. 2010.

[20] Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. *Labeled LDA: a supervised topic model for credit attribution in multi-labeled corpora.* In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1 (EMNLP '09), Vol. 1. Association for Computational Linguistics, Stroudsburg, PA, USA, 248-256.

[21] Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. *Learning to classify short and sparse text & web with hidden topics from large-scale data collections.* In Proceedings of the 17th international conference on World Wide Web (WWW '08). ACM, New York, NY, USA, 91-100.

[22] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. *The author-topic model for authors and documents.* In Proceedings of the 20th conference on Uncertainty in artificial intelligence (UAI '04). AUAI Press, Arlington, Virginia, United States, 487-494.

[23] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. *Improving LDA topic models for microblogs via tweet pooling and automatic labeling.* In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR '13). ACM, New York, NY, USA, 889-892.

[24] Thomas Hofmann. *Probabilistic latent semantic indexing.* In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '99). ACM, New York, NY, USA, 50-57.

[25] Letsche, Todd A., and Michael W. Berry. *Large-scale information retrieval with latent semantic indexing.* Information sciences 100.1 (1997): 105-137.

[26] John Roesler. *LaTeX-Topic-Model-Menagerie*, Retrieved February 27, 2014.

[27] Deerwester, Scott C., et al. *Indexing by latent semantic analysis.* JASIS 41.6 (1990): 391-407.

[28] Carlberger, Johan, et al. *Improving precision in information retrieval for Swedish using stemming.* the Proceedings of NODALIDA. 2001.

[29] M.F. Porter *Snowball: A language for stemming algorithms*, Retrieved February 28, 2014.

[30] Rajaraman, Anand, and Jeffrey David Ullman. *Mining of massive datasets.* Cambridge University Press, 2012.

[31] Tf-idf *A Single-Page Tutorial*, Retrieved June 9, 2014.

[32] OECD *PISA 2012 Results*, Retrieved May 20, 2014.

[33] Europa Portalen *Europe and the Refugee Disaster*, Retrieved May 20, 2014.

[34] Riksdagen *SåFunkar Riksdagen*, Retrieved May 20, 2014.

[35] lagen.nu *Beslut om Ny Regeringsform*, Retrieved May 20, 2014.

[36] Erich Owens *Text Classification and Feature Hashing: Sparse Matrix-Vector Multiplication with Cython*, Retrieved May 20, 2014.

[37] Anthony Goldbloom *Kaggle*, Retrieved May 20, 2014.

[38] Gurevitch, Michael, Stephen Coleman, and Jay G. Blumler. *Political communication?Old and new media relationships.* The ANNALS of the American Academy of Political and Social Science 625.1 (2009): 164-181.

[39] Van Biezen, Ingrid, Peter Mair, and Thomas Poguntke. *Going, going,... gone? The decline of party membership in contemporary Europe.* European Journal of Political Research 51.1 (2012): 24-56.

[40] Larsson, Anders Olof, and Hallvard Moe. *Studying political microblogging: Twitter users in the 2010 Swedish election campaign.* New Media & Society 14.5 (2012): 729-747.

[41] Tumasjan, Andranik, et al. *Election forecasts with Twitter how 140 characters reflect the political landscape.* Social Science Computer Review 29.4 (2011): 402-418.

[42] Borondo, J., et al. *Characterizing and modeling an electoral campaign in the context of twitter: 2011 spanish presidential election as a case study.* Chaos: an interdisciplinary journal of nonlinear science 22.2 (2012): 023138.

[43] Hong, Sounman, and Daniel Nadler. *Which candidates do the public discuss online in an election campaign?: The use of social media by 2012 presidential candidates and its impact on candidate salience.* Government Information Quarterly 29.4 (2012): 455-461.

[44] Hsu, Chien-leng, and Han Woo Park. *Mapping online social networks of Korean politicians.* Government Information Quarterly 29.2 (2012): 169-181.

[45] Golbeck, Jennifer, Justin M. Grimes, and Anthony Rogers. *Twitter use by the US Congress.* Journal of the American Society for Information Science and Technology 61.8 (2010): 1612-1621.

[46] Kim, Minjeong, and Han Woo Park. *Measuring Twitter-based political participation and deliberation in the South Korean context by using social network and Triple Helix indicators.* Scientometrics 90.1 (2012): 121-140.

[47] Auer, Matthew R. *The policy sciences of social media.* Policy Studies Journal 39.4 (2011): 709-736.

[48] Moe, Hallvard. *Who participates and how? Twitter as an arena for public debate about the data retention directive in Norway.* International Journal of Communication 6 (2012): 23.

[49] Ampofo, Lawrence, Nick Anstead, and Ben O'Loughlin. *Trust, confidence, and credibility: Citizen responses on twitter to opinion polls during the 2010 UK general election.* Information, Communication & Society 14.6 (2011): 850-871.

[50] Bruns, Axel. *JOURNALISTS AND TWITTER: HOW AUSTRALIAN NEWS ORGANISATIONS ADAPT TO A NEW MEDIUM.* Media International Australia (8/1/07-current) 144 (2012).

[51] Lee, Eun-Ju, and Soo Yun Shin. *Are they talking to me? Cognitive and affective effects of interactivity in politicians' Twitter communication.* Cyberpsychology, Behavior, and Social Networking 15.10 (2012): 515-520.

[52] Lee, Eun-Ju, and Soo Youn Oh. *To personalize or depersonalize? When and how politicians' personalized tweets affect the public's reactions.* Journal of Communication 62.6 (2012): 932-949.

[53] Conover, Michael, et al. *Political polarization on twitter.* ICWSM. 2011.

http://www.matthewjockers.net

[54] Matthew L. Jockers *"Secret" Recipe for Topic Modeling Themes*, Retrieved June 1, 2014.

[55] Ng, Andrew *CS229 Lecture notes - SVM.* CS229 Lecture notes 1.1 (2000): Chapter 1.

[56] Ng, Andrew *CS229 Lecture notes - Supervised Learning.* CS229 Lecture notes 1.1 (2000): Chapter 3.

[57] Scikit Learn *Stochastic Gradient Descent Documentation*, Retrieved June 3, 2014.

[58] Kalsnes, Bente, Arne H. Krumsvik, and Tanja Storsul. *Social media as a political backchannel: Twitter use during televised election debates in Norway.* Aslib Proceedings. Vol. 66. No. 3. Emerald Group Publishing Limited, 2014.

[59] Williams, Shirley A., Melissa M. Terras, and Claire Warwick. *What do people study when they study Twitter? Classifying Twitter related academic papers.* Journal of Documentation 69.3 (2013): 384-410.

[60] Enli, Gunn Sara, and Eli Skogerb. *PERSONALIZED CAMPAIGNS IN PARTY-CENTRED POLITICS: Twitter and Facebook as arenas for political communication.* Information, Communication & Society 16.5 (2013): 757-774.

[61] Anders Sundell *PhD Student Information*, Retrieved June 3, 2014.

[62] Poblete, Barbara, et al. *Do all birds tweet the same?: characterizing twitter around the world.* Proceedings of the 20th ACM international conference on Information and knowledge management. ACM, 2011.

[63] Bergsma, Shane, et al. *Language identification for creating language-specific Twitter collections.* Proceedings of the Second Workshop on Language in Social Media. Association for Computational Linguistics, 2012.

[64] Carter, Simon, Wouter Weerkamp, and Manos Tsagkias. *Microblog language identification: Overcoming the limitations of short, unedited and idiomatic text.* Language Resources and Evaluation 47.1 (2013): 195-215.

[65] Preotiuc-Pietro, Daniel, et al. *Trendminer: An architecture for real time analysis of social media text.* Sixth International AAAI Conference on Weblogs and Social Media. 2012.

[66] Binkley, David, et al. *To camelcase or under_score.* Program Comprehension, 2009. ICPC'09. IEEE 17th International Conference on. IEEE, 2009.

[67] Stephan Richter *lxml - XML and HTML with Python*, Retrieved June 6, 2014.

[68] Blei, David M. *Probabilistic topic models.* Communications of the ACM 55.4 (2012): 77-84.

[69] Posch, Lisa, et al. *Meaning as collective use: predicting semantic hashtag categories on twitter.* Proceedings of the 22nd international conference on World Wide Web companion. International World Wide Web Conferences Steering Committee, 2013.

Appendix A JSON

JSON is built on two structures:
A collection of name/value pairs. In various languages, this is realized as an object, record, struct, dictionary, hash table, keyed list, or associative array. An ordered list of values. In most languages, this is realized as an array, vector, list, or sequence. These are universal data structures. Virtually all modern programming languages support them in one form or another. It makes sense that a data format that is interchangeable with programming languages also be based on these structures.

An object is an unordered set of name/value pairs. An object begins with (left brace) and ends with (right brace). Each name is followed by : (colon) and the name/value pairs are separated by , (comma). An array is an ordered collection of values. An array begins with [(left bracket) and ends with] (right bracket). Values are separated by , (comma). A value can be a string in double quotes, or a number, or true or false or null, or an object or an array. These structures can be nested.

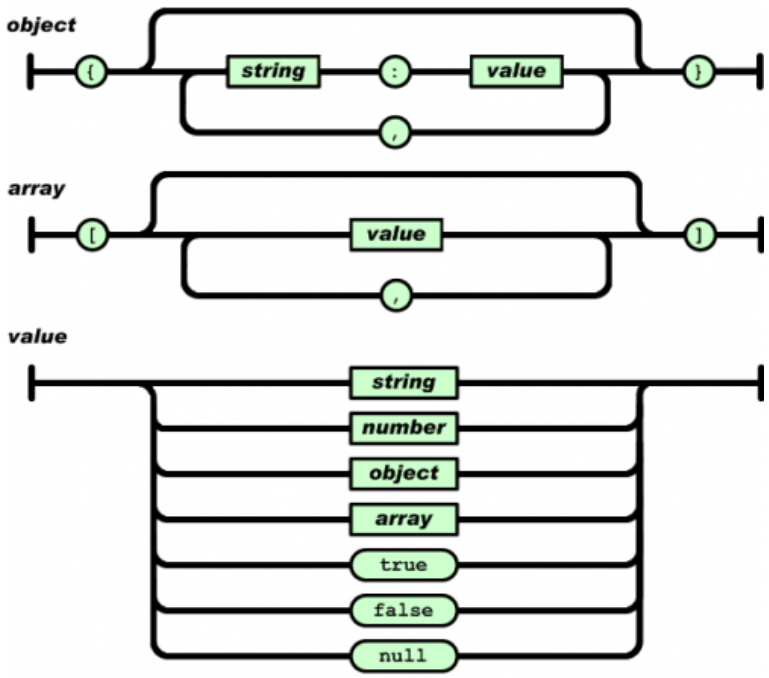

Figure 15: Diagram of JSON object array and value structure

A string is a sequence of zero or more Unicode characters, wrapped in double quotes, using backslash escapes. A character is represented as a single character string. A string is very much like a C or Java string

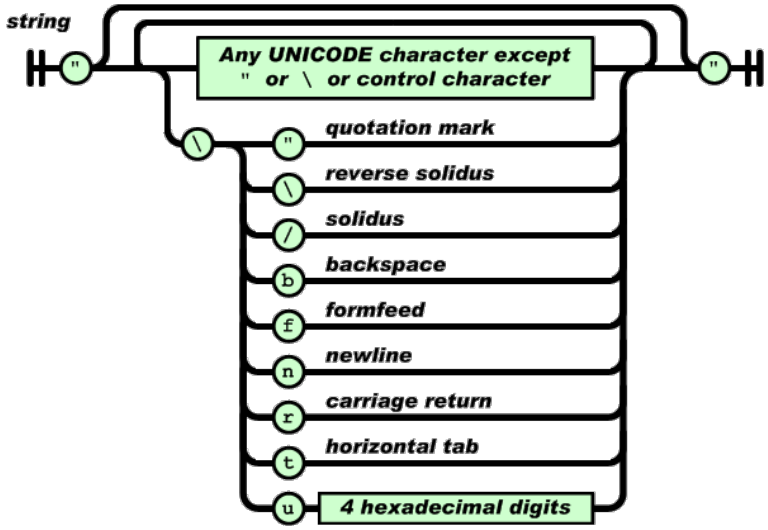

Figure 16: Diagram of JSON string structure

A number is very much like a C or Java number, except that the octal and hexadecimal formats are not used.

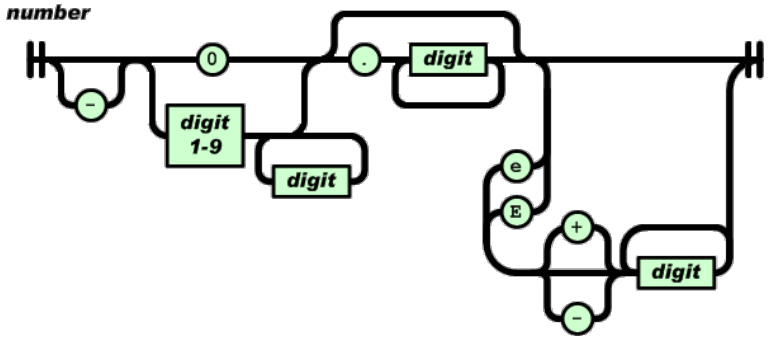

Figure 17: Diagram of JSON number structure

Appendix B Initial topic lists

skola	skolförvaltning	kunskap	komvux
folkhögskola	skolinspektör	betyg	lärarlegitimation
friskola	skolinspektionen	lärarutbildning	skolorganisation
grundskola	skolk	resultat	elevhälsa
gymnasium	skollov	klasser	skolmat
högskola	matematik	lärarlön	plugg
högstadium	skollag	läxhjälpsavdrag	kunskapsnation
internat	skolledare	läxhjälp	pedagogik
klass	skolledning	flumskola	betygssteg
läroverk	skollunch	flumskolan	skolk
privatskola	skolnämnd	flumskolans	yrkesutbildning
mellanstadie	skolpersonal	flumskolor	alliansskolan
mellanstadium	skolprov	flumskolors	baskunskap
lågstadie	skolsituation	flumskolorna	lärartjänst
lågstadium	skolundervisning	flumskolornas	utbildning
högstadie	skolungdom	mellanstadie	utbildningsminister
skolpeng	skoluppsats	utbildningssystem	skoldebatt
skolform	skolutbildning	disciplin	lärarförbundet
skolgång	skolväsende	internatskola	speciallärarutbildning
skolmässig	skolämne	skolarbete	privatundervisning
speciallärare	skolår	skolavslutning	student
särskola	skolövning	skolplikt	klassrum
yrkesskola	lärare	skolschema	skolfrågan
skolreform	lärarfortbildning	skolväsen	högskoleutbildning
läroplan	lärarhögskola	skolpolitik	lärarlöner
läsår	lärarkandidat	läxa	friskolereform
rektor	lärarkollegium	katederundervisning	högskoleplats
rektorsområde	lärarkår	lärlingssystem	undervisning
skolansvarig	lärarledd	lärlingsutbildning	undervisningstid
skolavgift	lärarlegitimation	spetsutbildning	pedagog
skolbarn	lärarlinje	yrkesinriktad	undervisa
skolbetyg	lärarroll	elevdemokrati	skolresultat
skolbibliotek	lärarrum	lärarstil	skolminister
skolbänk	lärartäthet	återförstatliga	elev
skoldag	lärarutbildning	återförstatligande	skolverket
skolelev	provår	kunskapsskola	utbildningsbakgrund

Table 13: Agenda topic *school*

skolsystem
kunskapsklyftor
kunskapsklyfta
gymnasiebehörighet
studieresultat
studietradition
auktoritet
akademisk
studieprestation
idealskola
klassundervisning
resultatförsämring
specialskola
examen
yrkeshögskola
pisa
behörig
läraryrke
basämne
lärarfack
likvärdig

Table 14: Continuation: Agenda topic *school*

sjukvård	avdelningssköterska	landstingspolitiker
akutvård	nattsköterska	vårdskillnad
intensivvård	sjukskötare	specialistvård
landsting	sjuksyster	vårdföretag
långvård	undersköterska	landstingsdriven
primärvård	allmänpraktiker	specialistsjuksköterska
sjukvårdande	avdelningssköterska	inlandssjukvård
sjukvårdare	barnläkare	patientsäker
sjukvårdsapparat	chefskirurg	kö, köer
specialistvård	distriktsläkare	närsjukvård
vårdgivare	doktor	akutsjukhus
vårdpaket	farmaceut	sjukhusdebatt
patientplats	kirurg	patientval
sjukvårdsbiträde	klinik	carema
sjukvårdsförsäkring	kvinnoläkare	vårdföretagarna
sjukvårdsnämnd	logoped	caremaskandal
sjukvårdspersonal	nattsköterska	vårdförbundet
sjukvårdspolitik	privatläkare	vårdfrågan
sjukvårdsreform	privatpraktik	vårdguiden
sjukvårdsutbildning	receptarie	sjukvårdupplysning
vårdanställd	sjukgymnast	smittskydd
vårdgaranti	underläkare	medicinteknisk
vårdinsats	överläkare	cancermedicin
vårdpersonal	förlossning	hottledsoperation
vårdplats	förlossningsarbete	landsringsråd
vårdsökande	förlossningsklinik	vårdvalssystem
vårdval	förlossningsskada	öppenvård
vård	vårdsektor	journal
vårdbehov	vårdkö	patientlagstiftning
vårdbehövande	specialistsjukvård	tandvård
vårdcentral	lifescience	tandvårdsreform
vårdkrävande	sjukvårdssystem	patientsäkerhet
skillnad	läkarupprop	ehälsa
kötid	läkaruppropet	sjukhus
arbetsmiljö	vårdförbund	överbeläggning
platsbrist	hemtjänst	operation
utförare	akuten	läkare
åldrande	ambulans	ambulanssjuksköterska

Table 15: Agenda topic *healthcare*

landstingspolitiker
strokevård
barnsjukhus
ambulanspersonal
patientavgift
vårdtillfälle
palliativ
allmändtjenstgöring
alternativmedicin
anhörigvård
kirurgi
diagnos
dosering
dospatient
dosrecept
apotek
apoteksombud
apotekare
förskrivning
generiska
högkostnadsskydd
licensläkmedel
lekmedelsförmån
ordination
medicinsk
patientjournal
receptbelagda
remiss
överbeläggning
akutmottagning
vårdtid
vårdtider
sjukvårdslandstingsråd
läkarlinje
äldrevård
äldreomsorg
demenspatient
barnmorska
sjuksköterska
sjukhussäng
patient
sjukvårdskris

Table 16: Continuation: Agenda topic *healthcare*

brottslighet	häktning	överfallsvåldtäkt	fängelseförhållande
brott	brottsanalys	rättsväsende	fängelsegaller
våldsbrott	hittelön	sexuallagstiftning	plit
överfall	fängelsestraff	bevisbörda	häkte
lagöverträdelse	bov	kvinnojour	fångvård
upphovsrättsbrott	brottslig	kvinnojourer	skjuta
olaglighet	brottsling	tjejjour	skjutning
regelbrott	brottslighet	tjejjourer	skjutvapen
sexbrott	kriminalitet	brottsoffer	utpressning
sexualbrott	kriminell	uppklara	misshandel
svindel	kriminal	rättsväsende	oskyldig
mutbrott	kriminalvård	hovrätt	barnmisshandel
domstol	påföljd	tingsrätt	dödsmisshandel
narkotikabrott	olaglig	påföljd	familijevåld
rättsfall	gängbråk	böter	förgripa
nätbedrägeri	gängvåld	fängelse	hot
ogärning	lagstridig	stölder	bombhot
kvinnofridskränkning	gärningsman	bilstöld	ungdomsbrottsling
kortbedrägeri	förbrytare	cykelstöld	brå
jaktbrott	korrupt	kontostöld	förföljd
inbrott	korruption	kortstöld	överfall
försäkringsbedrägeri	polis	tillgrepp	gruppvåldtäkt
hatbrott	polisman	tjuveri	brottsförebyggande
fortkörning	snut	tjuv	barnahus
bokföringsbrott	polisdistrikt	riksåklagaren	åtal
angivelsebrott	polismakten	åklagare	åtalas
arbetsmiljöbrott	polisiär	frihetsberövande	brottsofferparagraf
barnpornografi	poliskår	smuggling	brottsofferbegrepp
bedrägeri	polismakt	smuggla	misstänka
straff	polisväsende	heroinsmuggling	misstänkt
straffa	kriminalpolis	knarksmuggling	vite
straffföreläggande	polisutbildning	människosmuggling	laglydig
straffskala	våldtäkt	tensta	rättskunskap
straffri	våldta	rån	rättskedjan
straffregister	våldtagana	råna	skyddstillsyn
straffpåföljd	massvåldtäkt	rånare	samtyckeslagstiftning
straffnivå	sexutnyttjande	centralanstalt	våldtäktslagstiftning
straffbarhet	sexuelltutnyttjande	fängelsechef	sexualbrottslagstiftning
påföljd	våldföra	fängelsedirektör	
livstidsstraff	våldtäktsman	samtycke	

Table 17: Agenda topic *crime&punishment*

flykting	fattigdom	statslösa
fly	uppehållstillstånd	flyktinginvandring
flykt	förföljelse	flyktingbarn
kvotflykting	libanon	flyktingfamilj
kvotflyktingar	libanes	flyktingläger
immigration	papperslösa	flyktingmottagande
immigrera	krig	flyktingmottagning
asyl	krigsdrabbande	migrationsdomstolarna
asylrätt	lampedusa	humanitärinvandrare
asylskäl	båtflykting	tredjelandsmedborgare
asylsökande	båtflyktingar	migranter
invandring	viseringskrav	identitetshandling
anhöriginvandring	viseringskravet	asylfall
arbetskraftsinvandring	jordanien	nyanlända
massinvandring	visumregler	utvandring
migrationsverket	visum	utvandrare
invandringskritisk	tortyr	massflykt
invandringskvot	unhcr	Genevekonventionen
invandringspolitik	rodakorset	humanitär
invandringsvåg	amnesty	familjeanknytning
flyktingkatastrof	amnesti	anpassning
flyktingsmuggling	asylombud	integration
syrien	flyktingbostad	sfi
syriska	flyktingbostäder	multikultur
asylregler	familjeåterförening	svenskhet
beskickning	anhöriginvandring	språkundervisning
ambassad	sydsudan	migrationsminister
visumbrott	integrationsverket	asylrätt
flyktingkonvention	invandrare	socialbidrag
flyktingstatus	migrationspolitik	human
internflyktingar	flyktinghjälpen	rasist
inbördeskrig	flyktinghjälp	rasism
flyktingläger	flyktinsmuggling	härkomst
flyktingströmmar	flyktinsmuggla	asylinvandring
flyktingström	afghanistan	antirasist
hemland	pakistan	antirasistisk
mottagarland	somalia	antirasistiskt
asylprocess	irak	antirasism
ensamkommande	sudan	fascism

Table 18: Agenda topic *refugees*

främlingsfientlighet	asylsökningsfråga
främlingsfientliga	flyktingskatastrof
främlingsfientlig	islamofobi
hederskultur	segregation
expo	mångkultur
islamisering	
medborgarskap	
integrationsprocess	
migrationstryck	
utlänningslag	
dublinförordning	
asylprocedur	
skyddsbehov	
asylsystem	
ansökningsförfarande	
arbetstillstånd	
återvändande	
ursprungsland	
ursprung	
bevilja	
skyddsskäl	
konventionsflykting	
flyktingproblem	
flyktingorganisation	
flytingrörelse	
flytingvåg	
visering	
avslag	
etnicitet	
folkgrupp	
etnisk	
assimilation	
assimilering	
nationalism	
krisdrabbad	
flyktingpolitik	
syrier	
asylpolitik	
flyktingfrågan	

Table 19: Continuation: Agenda topic *refugees*

jobb	skattekil	bidrag
jobba	skattekrona	barnbidrag
heltidsjobb	skattemedel	familjebidrag
instegsjobb	skattemässig	socialbidrag
svartjobb	symbolskatt	statsbidrag
jobbprognos	trängselskatt	studiebidrag
jobbskapande	vinstdelningsskatt	subvention
jobbskatteavdrag	återbäring	tandvårdsbidrag
anställning	arbetslöshet	utbildningsbidrag
arbeta	massarbetslöshet	vårdbidrag
deltidsarbete	strukturarbetslöshet	vårdnadsbidrag
extraarbete	ungdomsarbetslöshet	bidragsberoende
skatt	företag	bidragsgrundande
alkoholskatt	affärerna	bidragssystem
arvsskatt	bolag	förtidspension
avdrag	familjeföretag	förtidspensionera
beskatta	firma	förtidspensionerande
engångsskatt	företagsverk	förtidspensionering
förmögenhetsskatt	koncern	ungdomsavgifter
genomsnittsskatt	moderföretag	näringspolitik
grundskatt	småföretag	näringspolitisk
inkomstskatt	småföretagare	innovation
inkomsttaxering	storföretag	utgifter
intäktschablon	kris	budget
kommunalskatt	eurokris	skuggbudget
konsumtionsskatt	fastighetskris	statsbudget
lyxskatt	finanskris	underbalans
marginalskatt	resurskris	utgiftsbudget
mervärdesskatt	skuldkris	åtstramningsbudget
nolltaxerare	välfärdskris	överbalanserad
punktskatt	krisfond	budgetalternativ
skatta	krishantering	budgetarbete
skattebas	krispaket	budgetförslag
skattebelopp	krispolitik	budgetmotion
skattebetalande	krisstöd	budgetproposition
skattebetalare	kristid	budgetunderskott
skattebetalning	tillväxt	budgetöverskott
skattefri	krisåtgärd	fas3

Table 20: Agenda topic *climate*

inkomstskatt	arbetslinjen
inkomstskattesänkning	jobbfrågan
lågkonjunktur	arbetsgivare
avdrag	sysselsättning
löneavdrag	skattepolitik
rotavdrag	skattepeng
rutavdrag	långtidsarbetslöshet
skatteavdrag	låginkomsttagare
beskatta	
beskattande	
beskattning	
dubbelbeskatta	
obeskattad	
straffbeskatta	
särbeskatta	
chockbeskatta	
åtgärdspolitik	
jobbpolitik	
chockskatt	
underskott	
arbetslösa	
skattesänkning	
utanförskap	
skattehöjning	
arbetsgivaravgift	
närningsliv	
skatteintäkt	
arbetsmarknad	
utförsäkra	
företagande	
pension	
restaurangmomsen	
skattechock	
finanser	
skattesubvention	
finansminister	
entreprenörskap	
reformutrymme	

Table 21: Continuation: Agenda topic *climate*

Appendix C Stochastic Gradient Descent

The following description is based on the Scikit Learn documentation [57]. The Scikit Python library has a class `SGDClassifier` which implements stochastic gradient descent (SGD) classifier. The `SGDClassifier` trains a linear SVM and it fits it to the training data using stochastic gradient descent algorithm. The loss functions and regulators can be chosen as parameters. The input to the SGD classifier is a two dimensional numpy array `X` with the rows being the training samples and the columns being the features and an array `y` representing the labels of the training samples. Here is an example of creating and training a trivial a SGD classifier with hinge-loss and L2 regularizer:

```
>>> from sklearn.linear_model import SGDClassifier
>>> X = [[0., 0.], [1., 1.]]
>>> y = [0, 1]
>>> clf = SGDClassifier(loss="hinge", penalty="l2")
>>> clf.fit(X, y)
SGDClassifier(alpha=0.0001, class_weight=None, epsilon=0.1, eta0=0.0,
       fit_intercept=True, l1_ratio=0.15, learning_rate='optimal',
       loss='hinge', n_iter=5, n_jobs=1, penalty='l2', power_t=0.5,
       random_state=None, rho=None, shuffle=False, verbose=0,
       warm_start=False)
```

In order to predict labels to some test samples the model works as following:

```
>>> clf.predict([[2., 2.]])
array([1])
```

To get the learned parameters of the linear model the class member `coef_` should be called:

```
>>> clf.coef_
array([[ 9.91080278,  9.91080278]])
```

The bias or intercept of the classifier can be obtained using the `intercept_` member:

```
>>> clf.intercept_
array([-9.990...])
```

The `SGDClassifier` can use various loss functions when learning the classification parameters:

```
loss="hinge": (soft-margin) linear Support Vector Machine,
loss="modified_huber": smoothed hinge loss,
loss="log": logistic regression,
```

`hinge` and `modified_huber` loss functions are implemented in a lazy way: the classifier parameters are modified only if a training sample is within the margin. That is the reason for the classifier training with these two loss functions to be more time efficient. `log` and `modified_huber` have the class method `predict_proba` which tells the probability of the predicted labels given the training data for each training sample: $P(y|x)$.

```
>>> clf = SGDClassifier(loss="log").fit(X, y)
>>> clf.predict_proba([[1., 1.]])
array([[ 0.0000005,  0.9999995]])
```

The penalties used by the classifier can be set. The following penalties can be chosen:

```
penalty="l2": L2 norm penalty on coef_.
penalty="l1": L1 norm penalty on coef_.
penalty="elasticnet": Convex combination of L2 and L1;
                      (1 - l1_ratio) * L2 + l1_ratio * L1.
```

When choosing the L1 penalty small parameters will become zeros which will lead to a sparse outcome. This is valuable of the number of features is much larger than the number of training samples. In contrary if there are many more training samples than features L2 regularization will do better since it is directly related to minimizing the Vapnik-Chervonenkis dimension of the learned classifier.

Appendix D Negative Training Sample List

The following words were used for selecting non-party leader debate related tweets:

politik
politisk
politiskt
debatt
debattera
debatten
svpol
pldebatt
pldebatten
politiskdebatt
parti
partiet
reinfeldt
reinfeldts
partie
partien
partieledar
partieledare
partieledaren
partieledardebatt

Appendix E LDA Result With 50 topics

Topic #	Representative words for the topic
1	reinfeldt regeringen politiska resten hånleende nöjd därför bära värv
2	löfven reinfeldt skattesänkningar skriker arbetslösheten hjälper jobben hänt partiledardebatter
3	reinfeldt löfven alliansen oppositionen fredrik vinnare debatten debatt taggad
4	pratar låt skola mindre människor politiken regeringen punkt tjänar
5	diagram attackerade fuizbm ordet minst följa utrymme analys tweets
6	debatten brott straff avbryter polisen tiden statsminister talat ärligt
7	löfven stefan fredrik pressad människor svarar tydligt otroligt tyst
8	frågan partiledarna retorik sämst svara partier parti fall långt
9	björklund jan björklunds plakat frisyr politisk max vilde pratade
10	pldebatten slips lila bron jimmy titta ton gamla betydligt
11	pengar ökat alliansen välfärden vanligt resurser skatten antalet sänka
12	jobb skatter skatt sänkt högre anställa betala företag skapa
13	alliansen debatt valet kvällens någonsin programledarna regeringsalternativ sitta alliansens
14	reinfeldt hatar arbetslöshet utanförskap fas arbetslösa pratar utanförskapet hata
15	partiledardebatten partiledardebatt lär missat uttrycken tyda viktigt vv korrekt
16	björklund råd håret lärare arbete rufsat fridolin löven väljer
17	sverige skolan skola lärare länder klimat betyg ansvar hit debatten
18	vanligt argument paus synd dessa fakta särskilt ljuger förklara
19	sverige statsminister invandring världens väljarna frågar sossarnas miljö bort
20	politik svensk regeringens dålig läxhjälp fokus läxor nivångt
21	reinfeldt sanningen flexibel inställning fredrik prata lova partiledarna klarar
22	björklund skolan hört val härskartekniker flumskolan leda sjöstedt framstår
23	största regering politiskt leder märkbart unga spel upprörd yatzy
24	svt kvällens tv gårdagens partiledardebatt spännande plats klokt partiledarna
25	romson åsa prata behövs heja ordet debatt lööf samtycke
26	mp åkesson centerpartiet lyfter försörjning dö livslång roll syrierna
27	twitter tittar hand rödgröna låg vinner resultat ggr lågt
28	lööf annie löfven valfläsk köper monopolpengar landsbygden vindkraft stefan
29	svtagenda skolan mp brottsligheten migpol minskat själva landet fungerar
30	löfven besked mp överens alliansen tydligt löfvens regeringsfrågan rödgröna
31	löfvn siffror svara läsa drar sagt imponerad lyckas vart
32	partiledare politiker rösta parti partiledarna fortsätter partier demokratisk bug
33	miljarder sjukvården vården skolan skattesänkningar satsar sjukvård hägglund regeringen
34	sjöstedt jonas vänsterpartiet nämen usa reinis dbctklaei gått slipsar
35	sverige högre flyktingar europa lön asyl fp eu sjuksköterskor
36	debatten saknar utbildning partierna riksdagen forskning perspektiv jämställdhet vilket
37	debatten vann ståpartiledare chockbeskatta rör egna miljöpartiet ider
38	hägglund göran sjöstedt åkesson ökar totalt romson straffen slut
39	åkesson jimmie sd svälja invandrare dn åkessons snacka skärp
40	pldebatt sd land människor problem kristina satsa oskar karl
41	björklund hittills vanligt mp snälla satsat tala assåpassar
42	skolan själva betyg halvlek tyckte pausen tänka flumskola ngn
43	reinfeldts studion självmål https talade eepwalkl läckt fusklapp poäng
44	agenda svpol fredrik märks samarbete invandringen tyst svtplay heja
45	val talar kd hägglund regering exempel väljare regera vinner
46	varandra ner ord twitter lyssna dåligt klart lägg asylpolitiken
47	lööf annie pratar politik tänk prata opposition attackerar sämre t
48	reinfeldt svenska fram jävla moderaterna sjukt majoritetsregering folket hålla
49	svpol väljarna redo slår sämst mån omröstningen kompetens full
50	alliansen bort fram kvällens klimatet kärnkraft ledde granskarna debatten

Table 22: Topics found by LDA with number of topics set to 50 and user aggregation

Appendix F Words Filtering Swedish Language Tweets

på, är, och, det, jag, inte, med, att, för, har, till, du, som, av, om, så, nu, ett, vi, kan, ska, var, från, ju, när, lite, bara, får, bra, den, min, vara, eller, blir, mig, idag, vill, kommer, också, vad, han, hur, efter, finns, då, bli, skulle, mer, gör, än, här, ja, kanske, hade, mot, alla, få, mycket, dig, vet, måste, se, ni, göra, väl, över, nog, tror, helt, tack, går, tycker, vid, två, dag, Stockholm, där, varit, din, kul, under, själv, blev, fick, igen, säger, aldrig, hela, nya, bästa, be, ändå, precis, hon, sen, mitt, år, alltså, snart, ikväll, många, ny, borde, alltid, även, Sverige, detta, mina, dem, samma, första, hos, allt, dock, något, bättre, gärna, svpol, ingen, verkligen, sig, några, redan, sin, vår, nästa, gå, börjar, behöver, ser, fått, gillar, verkar, igår, innan, oss, någon, nej, typ, ta, varför, känns, riktigt, imorgon, faktiskt, väldigt, säga, denna, mest, dom, folk, kör, va, Malmö, hej, andra, låter, jobbet, dagen, utan, lika, hoppas, inför, vecka, dagens, mellan, tar, åt, gjort, kunna, fortfarande, skriver, ut, sett, tre, sitter, sedan, länge, grattis, plats, upp, kom, läsa, kl, köpa, ligger, tid, barn, menar, hemma, kolla, sina, sista, står, varje, funkar, dags, tänker, fler, ens, nästan, dagar, honom, saker, ej, ditt, rätt, ännu, jobb, veckan, Göteborg, nytt, gjorde, sjukt, kunde, jobbar, timmar, genom, såg, först, tyvärr, nyheter, senaste, komma, skriva, liten, vilken, vem, gick, sa, höra, tips, istället, haft, ute, svårt, spännande, blivit, bild, stor, håller, kaffe, jobba, olika, börja, enda, ur, gäller, helgen, direkt, minuter, förstår, er, glad, älskar, egentligen, sätt, bilder, fin, heller, våra, heter, mat, iaf, tillbaka, åh, söker, människor, åka, inget, läs, deras, ganska, hans, dina, blogg, Sveriges, sitt, fredag, känner, kväll, både, igång, händer, tänkte, sova, tydligen, alls, pratar, spelar, undrar, kvar, längre, par, ger, brukar, here, hitta, egen, ibland, nån, nåt, fel, hittar, see, annars, inom, inga, hem, äta, emot, roligt, bok, vore, skall, fråga, världen, gång, jo, pengar, stan, årets, fint, prata, veta, svar, gott, sant, säkert, frågan, lyssna, åker, ve, förra, utanför, liv, ofta, trodde, försöker, trött, flera, ord, massa, köra, läser, jävla, väntar, följa, annat, använder, handlar, enligt, musik, nä, använda, världens, spela, pga, tidigare, trevligt, sån, sak, middag, tur, å, visst, vårt, medier, skönt, frågor, Sthlm, gånger, vänner, lördag, henne, intressant, äntligen, sluta, kr, twittboll, ge, måndag, skolan, bör, långt, ses, gammal, följer, möte, kring, start, hört

Appendix G Data Storage

For this work, we obtained data in the form of a file from Språkbanken. This construction of the data was not adequate for processing, so, therefore, we decided to store the data in a more convenient format. We parsed the data file we received, and stored it in a database in a more suitable format.

Språkbanken gathered the tweets, lexically annotated it and saved it to an XML data file. We received the data in this format. XML stands for Extendable Markup Language. It is used as a format of data that is readable by humans and also easily parsed automatically. *Figure 14* shows a small part of the data file from Språkbanken. The XML data is not easily readable for humans due to the extensive markup compared to the content. The root element is describing the dataset, in our case it is named `corpus` and it has an attribute `id="twitter-pldebatt"`. The main data units are the users, the corresponding XML elements are named `user`. Tweets are grouped by the users who tweeted them. This is not the same data structure that was obtained from Twitter Public Stream API: Språkbanken. By adding the lexical information and changing the main data units from tweets to users, Språkbanken changed the structure of the data. The result of this was that existing methods for processing Twitter data were not applicable. We needed to create custom methods for parsing the data file and storing it in a database.

```
<corpus id="twitter-pldebatt">
    <user created="2012-12-09" description="~ never look back ♡"
    followers="82" following="42" id="1000007083" name="Felicia
    Nanasi ♡" tweets="270" username="felicianansi">
        <text datefrom="20130617" datetime="2013-06-17 09:25:09"
        dateto="20130617" hashtags="|" id="346544029774778368"
        mentions="|" replies="|" retweets="0" weekday="Mon">
            <sentence id="96045aff71-96079773d5">
                <w dephead="03" deprel="TA" lemma="|för|för övrigt|"
                lex="|för..pp.1|för_övrigt..abm.1|" msd="PP" pos="PP"
                 prefix="|" ref="01" saldo="|för..1|för..5|för..6|för
                ..7|för_övrigt..1|" suffix="|">För</w>
                <w dephead="01" deprel="HD" lemma="|övrig|för
                övrigt:01|" lex="|övrig..pn.1|för_övrigt..abm.1:01|"
                msd="AB.POS" pos="AB" prefix="|" ref="02" saldo="
                |övrig..1|för_övrigt..1:01|" suffix="|">övrigt</w>
            </sentence>
        </text>
    </user>
</corpus>
```

Figure 18: Partial data in XML format as received from Språkbanken

Our first step in storing the, data was to parse the XML file we received. We could

not use available methods for parsing, since no known method could handle an embedded XML structured data of 6.6 GBs in size. After we tried the online available open source XML parser implementations we did not find any of them suitable to our data structure and size. Therefore, we implemented an event-based sequential access parser called SAX using ElementTree API by lxml [67], this parser didn't run into memory or performance issued and we were able to parse the data. SAX parsers have remarkably better computational performance since they report each parsing event as they happen. The information once reported doesn't get processed more times and the parser keeps track of unclosed elements in a stack. The memory required for the SAX parser is proportional to the maximum depth of the XML file and the size of the data involved in a single XML event, taking attributes of the element into account.

We converted each XML element, which was parsed, to a Python dictionary. Dictionaries are built-in Python data types. In other languages associative arrays represent dictionary like types. Dictionaries are indexed by string, number or tuple typed immutable keys. Sequences differ from dictionaries since they are indexed by a range of numbers. Dictionary can be visualized as an unordered set of key-value pairs, where the keys are unique. A pair of curly braces {} creates an empty dictionary. Placing a comma-separated list of key-value pairs within the braces adds initial key-value pairs.

Given the XML file shown on *figure 18* the `user` elements can be seen as dictionaries having the elements attributes as key-value pairs. The embedded XML element called `text` will be converted an embedded value of the user dictionary, which will have a key `text` with a value of the `text` dictionary. In cases like the `w` embedded elements, the value of the sentence dictionary's `w` key will be list of dictionaries converted from the `w` XML elements.

In most programming languages hash tables require a lot of hand coding and handling of pointer references. The advantage of Python dictionaries, beside the flexible indexing, is that they are available as a built-in data type. Furthermore Python dictionaries allow direct access to data via named references: no need to perform a search on the whole dictionary. The final reason for transferring XML elements to Python dictionaries is that they are structured in a similar way as JSON objects and the data will be stored as JSON object in the final phase of data storing.