Feature Selection Techniques for Effective Model Building and Estimation on Twitter Data to Understand the Political Scenario in Latvia with Supporting Visualizations

by

Lakshmi Gayatri Niharika Bollapragada

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved May 2016 by the
Graduate Supervisory Committee:

Hasan Davulcu, Chair
Arunabha Sen
Ihan Hsiao

ARIZONA STATE UNIVERSITY

August 2016

ABSTRACT

In supervised learning, machine learning techniques can be applied to learn a model on a small set of labeled documents which can be used to classify a larger set of unknown documents. Machine learning techniques can be used to analyze a political scenario in a given society. A lot of research has been going on in this field to understand the interactions of various people in the society in response to actions taken by their organizations.

This paper talks about understanding the Russian influence on people in Latvia. This is done by building an effective model learnt on initial set of documents containing a combination of official party web-pages, important political leaders' social networking sites. Since twitter is a micro-blogging site which allows people to post their opinions on any topic, the model built is used for estimating the tweets supporting the Russian and Latvian political organizations in Latvia. All the documents collected for analysis are in Latvian and Russian languages which are rich in vocabulary resulting into huge number of features. Hence, feature selection techniques can be used to reduce the vocabulary set relevant to the classification model. This thesis provides a comparative analysis of traditional feature selection techniques and implementation of a new iterative feature selection method using EM and cross-domain training along with supportive visualization tool. This method out performed other feature selection methods by reducing the number of features up-to 50% along with good model accuracy. The results from the classification are used to interpret user behavior and their political influence patterns across organizations in Latvia using interactive dashboard with combination of powerful widgets.

To my parents, family, and friends

ACKNOWLEDGMENTS

I would like to thank my advisor Professor Dr. Hasan Davulcu for the continuous support and guidance he has provided throughout my thesis. He has always been supportive and helped me for taking important decisions during my research under him. I would also like to thank Professor Dr. Sharon Hsiao and Dr. Arunabha Sen for accepting my request to be a part of my thesis committee and supporting me for my thesis preparation.

I would also like to thank Professor Dr. Mohamed Sarwat for accepting my request to be a substitute for Dr Sharon Hsiao.

I would also like to thank Anil for being supportive friend, guide and mentor. I have developed both individually and academically with you being on my side and guiding me in the right direction always.

Above all, I express my deep gratitude to my family for understanding my interests and supporting me throughout my education by guiding me in the right path and encouraging me for doing masters with thesis.I would like to convey a special thanks to my brother for being on my side during my stressful times at ASU and all the fun we shared during our masters.

Finally, I would thank all my friends from the CIPS lab and my other friends at Tempe, most of them being my seniors at ASU for including me in all your fun activities. You people made my stay at ASU memorable.Thank you.

TABLE OF CONTENTS

Page

LIST OF TABLES vii

LIST OF FIGURES viii

CHAPTER

1 INTRODUCTION 1

1.1 Motivation 2

1.2 Political Scenario in Latvia 3

1.3 Document Outline 4

2 RELATED WORK 5

3 BACKGROUND 8

3.1 Feature Selection Methods 8

3.1.1 Term Frequency and Inverse Document Frequency 8

3.1.2 Information Gain 8

3.1.3 Mutual Information 9

3.1.4 Chi Square 10

3.1.5 Iterative Feature Selection 10

3.2 SLEP Classifier 11

3.3 Twitter Crawler 11

4 SYSTEM ARCHITECTURE 12

5 DATA COLLECTION 13

5.1 Data Crawling 14

5.1.1 News Articles/Official Websites 14

5.1.2 Twitter 15

5.1.3 Facebook 15

5.2 Data Processing 15

CHAPTER Page

5.2.1 Initial Feature Selection 16
6 Feature Selection 18
6.1 Classification 19
6.2 Iterative Feature Selection 20
7 INDEXING and Visualization 23
7.1 Data Base 23
7.2 Indexing 23
7.3 Visualization 24
7.3.1 Volume Chart 24
7.3.2 Chord Widget 25
7.3.3 Heat Map 25
7.3.4 Event Time-line 25
7.3.5 Network Flow Chart 26
7.3.6 Tweet Table 26
8 RESULTS 27
8.1 Training Set Documents 27
8.2 Performance for different rho Values for SLEP 27
8.3 Performance Evaluation of Various Feature Selection Methods 28
8.4 Iterative Feature Selection Method 29
8.5 Visualization Tool 33
8.5.1 Volume Timeline Chart 33
8.5.2 Chord Diagram 34
8.5.3 Heat Map 34
8.5.4 Event Timeline 35

CHAPTER Page

8.5.5 Network Diagram 36
8.5.6 Tweet Table 37
9 CONCLUSION AND FUTURE WORK 40
REFERENCES 43
APPENDIX
A INITIAL TRAINING DOCUMENTS 44
B STOP WORDS FOR LATVIAN LANGUAGE 47
C STOP WORDS FOR RUSSIAN LANGUAGE 49

LIST OF TABLES

Table Page

8.1 Twitter and Website Documents Counts 27

8.2 Accuracies When Only Twitter Features are Considered for Training . . 30

8.3 Accuracies When Both Twitter and Facebook Features are Considered for Training 31

8.4 Iterative Feature Selection Algorithm Results for Pro-Russian vs Separatists Classification. The Results Show the Documents and Features Added in Each Iteration 32

8.5 Average Cross-validation Accuracies in Traditional Feature Selection and Iterative Method 32

8.6 Accuracies of the Iterative Feature Selection Techniques on the 20% Remaining Tweets Along With the Features Selected from Total Features by the Used Model 33

LIST OF FIGURES

Figure Page

4.1 System Architecture for the Thesis Showing the 3 Main Processes Involved 12

5.1 Latvian Classifier 14

8.1 Observed Accuracies for Varying rho Values Across Various Classifications 28

8.2 Comparison of Feature Selection Algorithms for Increasing Number of Features 29

8.3 Chi Square Accuracies for all 4 Classifications 30

8.4 Volume Chart Showing the Count of Tweets for EU, NATO and Russia Keywords With a Sliding Window for Time Selection 34

8.5 Chord Widget Showing Organization User Counts and the Shifts of Users Within Different Organizations 35

8.6 Heat Map for Tracking Down the Location of Maximum Activity 36

8.7 Event Time-line to View all the News Events Associated With the Tweets 37

8.8 Event Time-line to View all the Trending Hashtags Associated with the Tweets 38

8.9 Network Diagram Showing the Connection between Different Users Through Their Retweets and User Mention Information. On Clicking a Node a Pop-up Window Shows Up With the Twitter Profile of the User Clicked. 38

8.10 Tweet Table Used to View all the Tweets from the Selected Organizations at a Given Time 39

A.1 List of Political parties and NGOs 45

Figure Page

A.2 List of Politicians 45
A.3 List of Provocateurs 46
B.1 List of Stop Words for Latvian Language 48
C.1 List of Stop Words for Russian Language 50
C.2 List of Stop Words for Russian Language 51
C.3 List of Stop Words for Russian Language 52
C.4 List of Stop Words for Russian Language 53

Chapter 1

INTRODUCTION

In order to understand a political situation of a society, it is important to analyze the opinions and activities of the people living in the society by learning the trending topics in that region. Generally, this is done by analyzing only the opinions of important people belonging to the political organizations, media or some NGOs in a society by following their news and official blogs and websites. To understand the actual trend, one can even track down the opinions of common people. This can be achieved by collecting the data available over the Internet which focuses on the political situations involving opinions from various sections of the society. A huge amount of data related to these situations can be obtained from various sources. Since, manually analyzing all the data is difficult, it can be solved out easily by applying machine learning techniques to build a predictive model on a small set of informative labeled data. which can be used to automatically produce information about un-labeled data. The initial data can be obtained from the web articles involving official web pages of the political organizations or the social media profiles of the political leaders of these because these are the primary sources giving main ideas and ideologies of their organization. The un-labeled data comes from the common peoples Internet activity which has to be analyzed using the model built on the initial data. Social media is one place where users have a freedom to share their opinions over the Internet regarding any issue happening around. Since Twitter is a micro-blogging tool used for users to give a short and quick response to any change in social activity, it is used as the data source for collecting common peoples opinions, on which a classification is performed to categorize any user into an organization.

1.1 Motivation

Since, the data is collected from various sources, it involves various kinds of documents in terms of size, language and the quality of text. The micro-blogging sources are usually very small sized documents with about 140 150 words restriction. Also, coming to the quality of text, documents from social networking sites involve a lot of emoticons involves while the news articles are more organized and semi-structured. Since, local news articles are the main sources we are using as part of the project and hence the information mainly comes in local languages. Also, people mostly prefer posting on social media in their own local languages. A lot of standard language processing tools like stemmer, POS tagger, Lemmatization are available for some of the most common languages like English, Arabic, Chinese, and German etc. With the other languages, these standard language processing tools are not yet available which restricts the scope of this work. Also, the difficulty increases if these languages are very rich in the sense that the total vocabulary for these languages is about 200k. The use of the language processing tools is to recognize the most important words in a language that are to be considered as features for further model building and classification by eliminating the less important or common words in the language which doesn't give much information. So, with the limited available language processing tools, the main aim here is to come up with proper mechanism to select the most relevant features as the corpus. The statistical results of the tools can be well interpreted when they come up in the form of powerful visualization. So, the final goal of the thesis is to build a real-time dashboard which shows the interactions of the twitter users and their activities throughout a particular time period. This kind of visualization tool also helps the analysts with political interest for easy analysis of the activities happening in Latvia.

1.2 Political Scenario in Latvia

Latvia is a democratic country in the northern part of Europe. It was under the foreign rule from about 13th to 20th century and despite that, it could keep up with its culture and identity. Latvians were the indigenous people of Latvia. It was occupied and forcibly incorporated into the Soviet Union for a long period of time. As a result of this, it has been a home to a large number of Russians. According to the Latvian Ministry of Foreign Affairs of Latvian Republic, the ethnic groups of Latvia has 61.6% Latvian Speakers, 25.8% Russian speakers and the rest are from other groups MFARL (2015). It has been a part of a large number of organizations like North Atlantic Treaty Organization (NATO), European Union (EU) , CBSS, the IMF, NB8, NIB, OSCE and WTO. Amongst these organizations, the ones of interest are mainly NATO, EU and how the Latvians and Russians share their opinions for these two organizations. NATOBartl (2013) is a large intergovernmental military alliance which includes 22 countries as part of its peace program. Latvia has been a part of the soviet military for a long period. In the late 90s Latvia has started integrating its military with the NATO membership. It was in 2004 that Latvia has completely integrated into the NATO military. This integration has made its army one of the most modernized and strong military base in the Europe. Soon after Latvia joined NATO, it has also become a member of the European Union in 2004. EU is a politico-economic organization which includes 28 organizations under it and mainly operates on few institutions like he European Parliament, the European Council, the Council of the European Union, the European Commission, the Court of Justice of the European Union, the European Central Bank, and the Court of Auditors and it goes by decisions made by the member states.BBCNews (2016)

1.3 Document Outline

The rest of the document is organized as explained below:

Chapter 2 talks about related work

Chapter 3 talks about Background terms and definitions

Chapter 4 discusses the data collection procedure applied

Chapter 5 discusses the data collection procedure applied

Chapter 6 discusses the feature selection techniques used

Chapter 7 discusses the indexing and visualization techniques

Chapter 8 discusses all the results

Chapter 9 discusses the Conclusion

Chapter 2

RELATED WORK

Looking glass is a near real time visualization tool. It was applied in different projects to analyze the social activities for different scenarios of interest. The Indonesia Looking GlassKim *et al.* (2013) was used to analyze the hot social debated topics in online social media mainly focused around Indonesia. The main source of data is various web articles and some of the famous persons' social media profiles. This information was used to analyze the twitter streaming data mainly from that particular location. The visualization showed the orientation of the general people active on social media towards each topic of interest and the shift in their orientation for each month. Besides this, it also shows information about the most used keywords, hash-tags, and the famous users and their networks.

The UK Looking glass Kim, Nyunsu and Tikves, Sukru and Wang, Zheng and Githens-Mazer, Jonathan and Davulcu, Hasan (2013) had the same analysis as the Indonesia Looking glass. In this project a comparative analysis of SLEP, SVM and the Random forest algorithm was also done and it was observed that random forest and SVM gave high accuracies with the given training set. This project also made use of a precision matrix to find out the highly co-related discriminative keywords between the radicals and non-radicals amongst the Islamic organizations in UK. Finally a ranking system was used to rank these organizations based on their socio-activities which is decided based on the analysis of the web corpus. Finally, a real-time interactive dashboard was developed to visualize the interactions amongst the radical and non-radical organizations at a higher level and the activities of the 26 sub-organizations are also analyzed on the chord widget, volume charts with breaking hashtags, network

flow chart, heat map etc. The Looking glass for Centcom ProjectLab (2015) was used to find out the key ISIS persons who are active on social media like twitter. In addition to the usual classification, it also made use of a clustering technique which used LDA to detect various co-related topics in the set of documents with additional information about network re-tweets and user-mention information to identify these clusters. Finally a visualization tool was developed which shows a Sankey widget that is used to analyze the behavior of these clusters within a time period. Finally, the tweeters were classified as ISIS and non-ISIS users and their networks were tracked down by using their user-mention and re-tweets information. It also displays the top you-tube videos and most used urls with their domains specified that were mostly shared by a group of some interest. These tweets didn't have any location filter,so the heat map was basically used to observe all the main areas which had key social activity related to ISIS going on. This information was finally used to take down the accounts of all the key ISIS users.

While in most of the previous work, SLEP or random forest algorithms were used to classify the tweeters into their respective organizations, showing high accuracies using the 10 fold cross validation, the accuracies with the actual tweets were never verified. The training data consists of a mixture of domains which mainly includes well-structured news articles, Facebook user posts and group posts and the tweets of famous organization official pages and the test set is only on the tweets collected using the discriminative keywords generated from the training model. Since the sources vary, each of the source type follows a different distribution. The social networking site like Facebook generally involve all friendly information which could be through conversations and comments or sharing of some others posts while all the news articles have rich and serious information available. Also, most of the news articles and official websites maintain the standards in the language. There wouldn't

be a mixture of languages while in Facebook posts there would be a great mixture of languages mostly in a single posts. Similarly, twitter is different from Facebook and news articles with the restriction of the number of characters for each tweet, the content of each tweet might not be as comprehensive as in a news article. This brings in the main difference with different sources of data. Some research is happening in the area of domain adaptation and better feature selection techniques along with the focus on getting a higher accuracy which involves applying various feature selection techniques on the set of collected documents and using the feature selection method that works best on the data. Once the features are selected, a predictive model is built on this data using some appropriate classification technique that works best on the data that can be used for classification of test sets.

Also, a lot of research has been happening around to come up with a proper representation of the data on to a visualization tool. The looking glass project has been evolving from time to time with various intuitive additions into the visualizations. Though all the previous versions of looking glass had widgets to represent the timeline showing the tweet volume for a given timeline window and also other widgets showing the collection of tweets, hashtags, keywords used separately, there is no proper place which represents all these information together integrated with a timeline that supports sliding through the time period to visualize interesting events happening around during each time period with supporting information in the form of a tweet or image or the news content or videos. The addition of this information would provide great information into the visualization tool.

Chapter 3

BACKGROUND

3.1 Feature Selection Methods

3.1.1 Term Frequency and Inverse Document Frequency

Term frequency (T.F) is the count of a term that occurs in a given document and Inverse Document Frequency (I.D.F) is a value computed for each term taken to be inversely proportional to the total occurrence of the term in the complete document set Yang and Pedersen (1997). So, the inverse document frequency for a term will be unique across all the documents. The importance of a term in a document is computed by taking the product of T.F and I.D.F values of the term. For feature selection, the average T.F values for all terms are computed across all the documents and their product with their I.D.F values are computed and these terms are sorted in decreasing order of their final computed value.

3.1.2 Information Gain

Information Gain is used to know the amount of information obtained in terms of number of bits with the inclusion or exclusion of a given feature. It is used to know the goodness of the term in a given set of documents. As mentioned in Yang and Pedersen (1997) if c_i where i=1 to m are set of m classes, the information gain of a given term is equal to

$$InfoGain(t) = -\sum_{i=1}^{m} Pr(c_i) log Pr(c_i)$$
$$+Pr(t)\sum_{i=1}^{m} Pr(c_i|t) log Pr(c_i|t)$$
$$+Pr(\ t)\sum_{i=1}^{m} Pr(c_i|\ t) log Pr(c_i|\ t)$$

This is a more generic representation for a multi-class classification problem and it can be used to find the info gain for a binary class classification problem by substituting i = 1,2 in the above formula as described by Lewis and Ringuette (1994).

3.1.3 Mutual Information

Mutual information is almost similar to information gain where-in both calculate the entropy of a given term but information gain makes a decision by checking how much information each feature provides on its selection at a given stage while mutual information provides the total information a feature contributes when compared to selection or removal of other features. Consider a term t and class c. IT measures all the possible combinations of term t and class c occurrence by taking the probability of number of document occurrence where c and t co-occur, probability where c occurs without t and probability where t occurs without c amongst all documents and computes the total mutual information as probability of t co-occurring with c by the total occurrences of term t and class c individually [Church and Hanks (1990) and Fano and Wintringham (1961)].

3.1.4 Chi Square

Chi square is generally used to measure the level of independence between any two variables. For feature selection it is considered as the level to which a term and class are independent. This is measured once again by considering all the possible combinations of a term t and class cs' occurrence as calculated in mutual information. Also, an expected values of the co-occurrences are also measured and the features are ranked based on the amount of difference between actual observation and the expected values.The more is the dependence the more important the feature is considered with Moh'd A Mesleh (2007).

3.1.5 Iterative Feature Selection

The iterative feature selection methodology is a kind of domain adaptation problem where in domain adaptation the small size in-domain documents are used to train the out domain larger set training documents iteratively which works by adding more documents from outer domain into the training set by first testing the model from small in-domain documents. In this process as new documents are considered and if their predictions comes out to be true, then they are also included into the new training set. In this process, all the new set of features are also included into the actual training set with some weights assigned according to the score the documents received through the testing process as described in Peddinti and Chintalapoodi (2011).

3.2 SLEP Classifier

Sparse Learning for Efficient Projections is a classification library available to classify mainly the sparse data. There are several algorithms implemented in this package of Linear and Logistic regression by considering various parameters into consideration such as regularization , normalization , least square loss and logistic loss etc. The main classifier used in this thesis is the Logistic regression by considering the logistic loss with Regularisation. A cross-validation approach is used to mainly check the accuracy of the algorithm to build the model. This is chosen because the model built gives a score for each feature which can further be used for obtaining the discriminative features for the test set tweet collection using the crawlers.

3.3 Twitter Crawler

From the result of the model building on the training documents obtained using the SLEP classifier, the most important discriminative keywords are picked out and these keywords are used as a filter query to restrict the tweets to be collected using the Twitter 4j crawler. Along with the keywords a filter is also used to restrict the language and the location from which the tweets have to be collected. With these parameters set, the tweets are collected on a daily basis starting till the present date. The model obtained from the training is finally applied over the tweets collected which are finally classified into their respective organizations.All the classified tweets are stored back into the postgres database.

Chapter 4

SYSTEM ARCHITECTURE

The following diagram in Fig 4.1 explains the system architecture followed for this thesis. The entire thesis is broadly divided into three sections.

1. Data Collection and Pre-Processing
2. Feature Selection and Model building
3. Indexing and Visualization

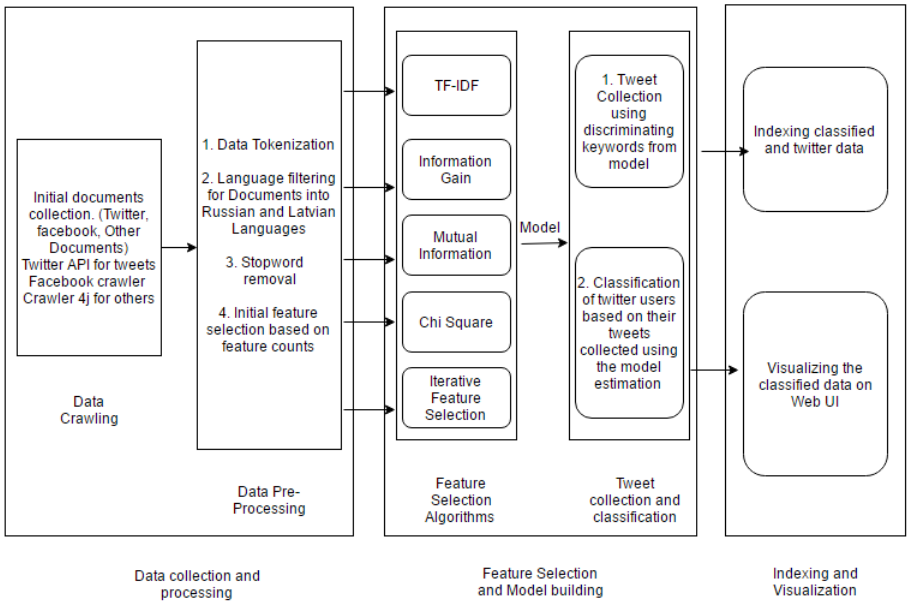

Figure 4.1: System Architecture for the Thesis Showing the 3 Main Processes Involved

The next three chapters describe the three sections in detail.

Chapter 5

DATA COLLECTION

In order to collect data, the first important step is to understand all the various organizations present in Latvia.A document containing the information about various organizations in Latvian and Russian organizations interested in Latvia are collected. Thus, all the organization official websites, the top politicians, journalists , NGO's involved in the organizations, their official websites and social networking profiles are being collected for these list of organizations. The details about all the organizations are provided further in Appendix A. The organizations for which the data is collected are first broadly categorized into two types based on language. They are the Russian speaking organizations and the Latvian Speaking organizations. The Latvian organizations are further classified into Left-Wing , Center Right and Right Wing organizations. Similarly the Russian organizations are classified as Democratic Socialists, Pro Russians and Separatists.

Figure 5.1 gives a detailed illustration of how the classification is done in two levels. The first level is based on the language for the complete set of documents. The next levels are within each language set. For Latvian documents the second level of classification is done for Left Wing vs the Right Wing Organizations. The third level is within the right wing between the Nationalists and the Radical Right Organizations. In the Russian speaking documents, the second level of classification is doe between Socialists vs Pro-Russian and Separatists organizations. The next level of classification is done amongst Pro-Russian and Separatists organizations.

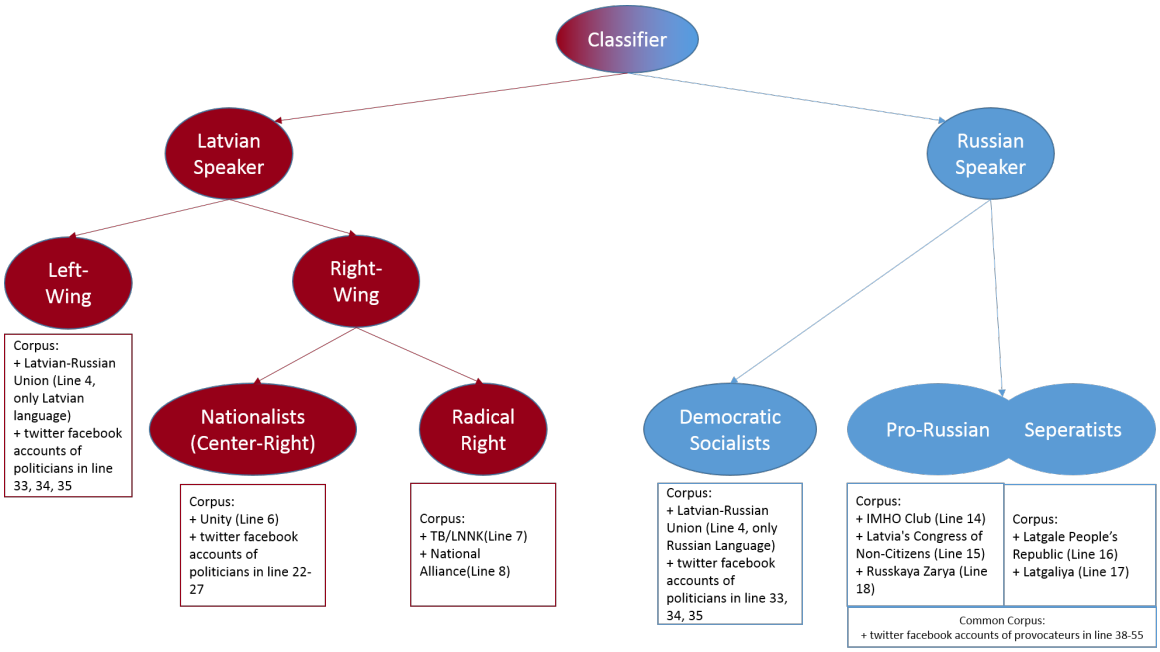

Figure 5.1: Latvian Classifier

5.1 Data Crawling

Using the initial information collected for the organization and users', the documents were mainly classified into three types. Normal documents/ News articles, Twitter profiles,User's Facebook pages. Each of these have a different method for crawling implementation. The main goal of the thesis is to analyze the documents only belonging to Latvian or Russian languages. So, a filter is made to restrict documents belonging to only these two languages in the process of crawling.

5.1.1 News Articles/Official Websites

For collecting the news articles,a library called Crawler4j is applied to initially collect all the URLs available from the source website or the news articles to a particular level of crawling depending on the correctness of the URLs collected and these URLs are further scraped to collect the data. The data collection from the URLs is

done using jsoup library which initially collects all the HTML code available for the news articles and selects just the text elements and the headlines from the articles.

5.1.2 Twitter

Twitter crawling is done using the twitter crawler search API which allows collecting all the tweets made by a user given their user-name. It also has a provision to detect the language and hence by doing this only the Latvian and Russian tweets are only selected.

5.1.3 Facebook

Facebook doesn't have a direct API provision for crawling a user's public posts without obtaining the permissions of the user. So, all the user's profiles were first saved locally and from all the documents only the text from the tags which involves information about the posts are selected and scraped.

5.2 Data Processing

All the collected documents have to be processed at this stage to obtained the dictionary and the final document-term matrix. The first step here is to The Facebook and news articles are first filtered into Latvian and Russian languages so that only using a Java's language detection library. Once the documents in both the are available, the next step is to classify the documents into two groups based on languages as Russian and Latvian documents.For each group, the next stage is to collect the vocabulary from the entire document set. For this step, tokenization has to be done which identifies each individual word based on spaces, full stops and other punctuation marks. The next step is to apply language detection over all these tokenized words. Since,the Language detection library assigns a particular language to

the complete document based on the percentage of the content of the documents, it is required to run the language detection over the entire tokenized words at this stage. Once the vocabulary is cleaned for obtaining a homogeneous language, the next step is to remove the stop words. Stop words are the most commonly occurring words in a given language which don't contribute any information for classification. These also involve removal of numbers, emoticons, URLs etc.The complete algorithm for the data processing is provided below

Algorithm 1 Processing the initial documents

```
procedure PROCESSING_DOCUMENTS(Documents)
    Find out the language of the entire document
    Filter out documents belonging to one Latvian and Russian Languages
    Tokenize all the documents on space and punctuation marks.
end procedure
procedure PROCESSING_TOKENS(Token)
    Filter tokens belonging to only Russian and Latvian language
    Remove stop words
    Remove URLs and emoticons
return token
end procedure
```

5.2.1 Initial Feature Selection

The initial vocabulary set that is obtained is very huge consisting around 200k terms. Out of all these terms a large number of terms are the ones which only have single occurrence across all the documents. All these less important documents have to be removed initially as they are considered to be the rare words. With the cleaned

and processed vocabulary set available, the next step is to obtain the the features that are to be considered. The classification technique used here is a binary classification. Thus for each classification level, the documents belonging to the two classes are considered. The vocabulary set for only these documents are also considered. For each term in the vocabulary set, the average count of each class is obtained along with the difference between the average counts. The difference between the averages helps in analyzing to what class each words belongs and the frequency helps in realizing how important the word is. Through this procedure all the single occurring words are removed first. Then the top 30k words from each group are selected from each group and combined to form the final list of corpus.The algorithm for the final vocabulary selection is follows

Algorithm 2 Initial selection of features

1: **procedure** FEATURE_REDUCTION(Vocabulary_Set , Class1 , Class2)
2: Filter Vocabulary Set belonging to Class1 and Class2 only.
3: For each term in the vocabulary set, find out the average count per document for each class
4: Find out the difference between the average counts to find their weights for each class
5: Eliminate the words which have single occurrences
6: Sort the vocab sets in each group in decreasing order of their frequency
7: Select the top 30k terms from each group and add it into the Vocab_Set
8: **return** Vocab_Set
9: **end procedure**

Chapter 6

FEATURE SELECTION

Since we are dealing with Latvian and Russian languages, the vocabulary set for these languages is very huge. Feature Selection is an important step to reduce the number of features. There are different feature selection techniques available and every technique provides different set of features based on their importance within different features and for different classes. Also, with increasing number of features though the accuracy for cross validation might be high, when the classification is actually done on the final set of the documents, the accuracy might not be the same. This could be as result of over-fitting for the initial set of features for the training documents. At the same time features shouldn't be too few that they become insufficient for classification leading to the case of more variance and biased results. So, an exact number of features have to be selected for a proper classification. Different feature selection algorithms work best on different types of data sets. So, in-order to choose appropriate features, a comparative analysis of all feature selection algorithms have to be done.

Since the test set documents are only from twitter data sets, it is necessary that, the training documents should have sufficient number of twitter documents for classification. Otherwise,if both the the sets come from different domains, they follow different feature distributions and hence the model built on one domain articles cannot be applied on other domain.This situation is very similar to the domain adaptation problem. In this case, I have come up with an algorithm that computes the model for the tweet data set based on its learning from the training set consisting of very few tweets.

For training sets containing considerable number of twitter documents, different feature selection algorithms are applied for comparison to get the best feature selection algorithm. The feature selection algorithms used for comparison on the training set are Term-Inverse document frequency, Mutual information and Information Gain algorithms.For each of these algorithms, first the original corpus is selected with the complete list of the features. The document vs features matrix is formed for these algorithms. On this matrix, each of the algorithm is applied separately. Once the values for the term-inverse document frequencies , mutual information and information gain values are obtained, the features are arranged in the decreasing order of their values computed. For classification, different sets of features are collected based on the number of features as 5k , 10k , 15k , 20k , 25k and 30k.The Chi-square algorithm used helps in picking the total number of features that are to be considered for classification. So, exclusive selection of features is not done in this process.

6.1 Classification

The SLEP classifier is used to build the model for all the new set of features. The SLEP classifier implements the Logistic regression algorithm. So, it requires deciding rho parameter which is used to decide the step size required for the gradient descent algorithm implemented by the logistic regression. The rho value has to be decided very carefully. It shouldn't be too small so that the algorithm takes very long time to reach the final conclusive point. It shouldn't be too high in which case the solution is never reached and the algorithm takes forever trying to optimize the weights equation. So, the algorithm performance is varied with different rho values and for each feature set of each feature selection algorithm, a 10 fold cross validation is applied.The validation is done by referring to the class variables of the training set and checking out how many documents turned out to be classified correctly. Since, its

a 10 fold cross validation, an average over all the 10 iterations' accuracies are taken and set as the final accuracy. This helps in coming up with the best model that can be used for further classification of the test set tweets. The model build is a vector containing the scores provided for each feature in the complete feature set. This model is used for two purposes. One is to find out the discriminative keywords which can be used for the test set tweet collection. This is done by considering the features that have high positive and negative scores.Since it is a binary classification,all the positives scores are used to collect test set tweets for the positive class/organization and all the negatives scores are used to collect the test set documents for the negative class/organization.

6.2 Iterative Feature Selection

The training set is a combination of tweets, facebook posts and news articles or official party web pages. The percentage of tweets when compared to all the other documents are very less. Hence, I have come up with an iterative feature selection method that helps in selecting the best weighted features which are more close to the twitter distribution so that the model works well for the test set. The idea follows the domain adaptation method as described by Peddinti and Chintalapoodi (2011). The algorithm is as follows: If the training set doesn't have sufficient tweet documents, around 1000 documents are manually collected and labeled by translating them into English using Bing Translator API. Once the initial training set of tweets are collected, the tweets are divided into 700 test sets and 300 training sets.A sparse matrix is formed for all the documents where the presence of a feature is represented by placing 1. Initially a model is built using the training set tweets and this model is applied on the normal documents. Once the normal documents are classified, the number of correctly classified documents are considered for the next iteration of

training with partial counts for their features. In this process, the extra features that are obtained from these new documents are added into the feature set. This new training set is used to re-estimate the classification for the rest of the documents. The advantage of using sparse matrix comes in here as the sparse only considers the matrix dimensions where the values exist. So, even if the corpus includes features from all the documents, the sparse matrix for the initial tweet documents is only built for the features that belong to the twitter documents. As new documents keep adding up those features will be represented in the final sparse matrix.

Algorithm 3 Iterative Feature Section method

```
1: procedure TWITTER_CORPUS_SELECTION(tweetSparse, docSparse)
2:     tweetSparse, docSparseformed
3:     faetureList ← features from tweetSparse
4:     Train using SLEP on tweetSparse
5:     predictdocSparse ← modelresultsforthedocSparsedocuments
6:     tweetSparse ← (predictdocSparse == docSparseClassLabels)
7:     for doc ← predictdocSparse do
8:         weight = exp(1/abs(predictedScore − meanclassScore))
9:         weight = weight/1 + weight
10:        doc = weight ∗ doc
11:        Twitter_Corpus_selection(tweetSparse, docSparse)
12:        if then predictdocSparse is null
13:            break
14:        end if
15:    end for
16: end procedure
```

Thus out of all the feature selection algorithms, the ones which provide best accuracies with meaningful number of features are selected. Using these features, the final estimation on the crawled tweets is done.

Chapter 7

INDEXING AND VISUALIZATION

7.1 Data Base

The tweets are collected into the postgres database along with other information about the tweets like the user's information, the retweets for each tweet, the hashtags withing each tweet, the user mentions. Each of these information is stored separately in different tables in the postgres database. The classified results are also stored as result table in the final database. Each of the tables are mainly linked with the tweet id as a primary key which helps in easy join of all the tables when their association is required.

7.2 Indexing

All the data that has been stored in the database when combined with the classification results give more meaningful information about various interactions of the users. The data in the database is very huge and hence handling such huge data would be very time consuming. In order to be able to use the data from the database in an efficient way, data indexing is required. Apache SOLR is a open source project which implements Apache Lucene in the background which helps in faster indexing and searching actions. It has a REST API support that allows usage of Apache SOLR from any platform. In this thesis indexing is important because the amount of data that is being collected is very huge and this data has to be queried in-order to represent it finally on the visualization tool. The indexing of the fields is done according to the requirement of the visualization. For indexing, three main files have

to be handled on the SOLR. First, the SOLR configuration file has to be filled with information about the search handler, request handler to set the facets and data folder path has to be set which helps in holding the indexed data.

7.3 Visualization

Once the indexing for all the required fields is done, the next stage is to represent all the information in an understandable way. There are different tools available to represent all the data and its interactions. Javascript is a powerful tool for rendering any sort of information on the Internet through a webpage URL. Javascript has powerful libraries that support such rendering of information. The libraries used for visualization in this thesis include d3js and google visualization charts. This allows dynamic rendering of data with interactions and is designed to handle huge amount of data. In this thesis the d3 widgets that are used are chord, network flow chart. Along with this google visualization tools are also used for representing annotated charts, heat map and tables. And finally Timeline.js is used for representing event timeline. The data represented through each of these widgets is explained below.

7.3.1 Volume Chart

The volume chart is basically a timeline which shows the volume of total tweets at any given time along with the total time-period for which the tweets have been collected. It has a window slider which is used to select the range of dates for observing the trends. This widget is used to mainly drive all the other widgets using the time range set by the window slider.

7.3.2 Chord Widget

The chord widget is used to represents users between two time periods along with their change in positions between the two time periods. It is a powerful widget which handles the data behind it through a matrix showing the previous organizations as the matrix rows and the current organizations as the matrix columns and the value being the total flow between the two organizations. The interesting scenarios using this widget is to observe the shift of a large number of users belonging to one organization into another organization over a given period of time. This further interacts with other widgets by selecting a particular organization or a particular path which basically represent all the users either belonging to an organization or the set of users who have moved from one organization to the other organization.

7.3.3 Heat Map

The heat map is a google visualization tool which is used to visualize the heat of tweets based on their locations on the map. It has a provision for selecting a polygon over a zoomed area which helps in restricting the tweets from the users belonging to only the area covered within that region. This helps in interacting with other widgets by selecting a polygon which includes all the users belonging to it to observe their activities on the other widgets.

7.3.4 Event Time-line

The event time-line is used to show the list of all events that occur during a given period from the tweets that are being collected. Thus these events can be of type hash-tags which represent the entire tweet information as an event, a news article which is picked from an URL obtained from the tweet. This event is also supported

with a small snapshot of the image related with news article. The URLs and hash-tags selected are the top 10 events that are mostly shared or talked about amongst all the tweets on a given data.

7.3.5 Network Flow Chart

The network flow chart is used to represent all the connections between the users based on the information from the user-mention in a given tweet. This shows all the users as nodes of the networks with the edges being the connections between the users. Each user is represented by the color of their organization as represented in the chord widget. Even for the networks diagram, if total number of users are to be considered, the graph would overflow and wouldn't be informative. So, the number of users have to be restricted as top 100 most influential users. This is done by using the Brandes' algorithm on the graph which computes the be-tweeness and centrality for the graph and presents the top 100 users based on their scores. The interesting scenario for the network graph would be to find out the users who have maximum degree. These will be the users who are mentioned by a large number of other users and have some influence over the others.

7.3.6 Tweet Table

The tweet table is used to represent the list of all the unique tweets that have occurred on a given day and are ordered based on their counts. Along with the tweets, this widget also represents the most frequently shared urls and also the top you-tube videos.

Chapter 8

RESULTS

8.1 Training Set Documents

The total number of Documents collected for the entire organizations are 43,020.The detailed description of the document distribution between twitter and other documents is provided below in Table 8.1

Organization	Twitter Documents	Other Documents
Left Wing	1000	5041
Nationalist	6451	3928
Radical Right	3186	2331
Socialist	834	35047
Pro Russian	1282	3782
Separatists	1411	795

Table 8.1: Twitter and Website Documents Counts

8.2 Performance for different rho Values for SLEP

The accuracies for the various feature selection techniques are plotted against increasing rho values and it is found that with the increase in rho value for all feature selection techniques applied on all classifications, the accuracies tend to be maximum for rho values around 0.3 - 0.5. The following figure 8.1 shows the variation of accuracies for increasing rho values considering 50k features for various classifications when information gain feature selection is used.

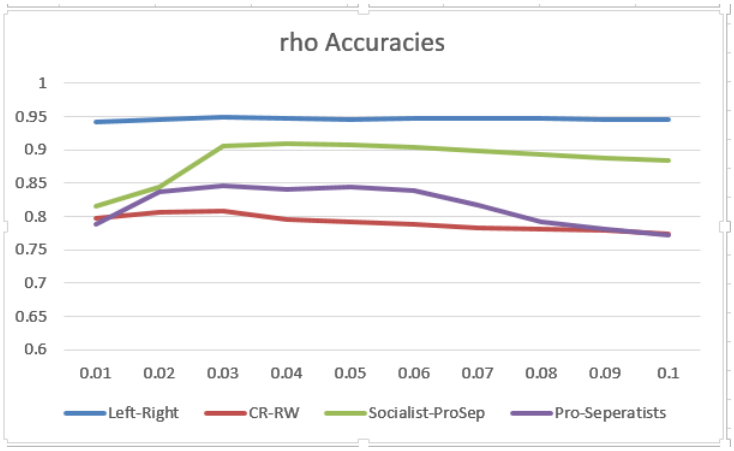

Figure 8.1: Observed Accuracies for Varying rho Values Across Various Classifications

8.3 Performance Evaluation of Various Feature Selection Methods

Experiments were conducted on the training documents initially by using different feature selections methods for varying number of feature set size and different values of rho. The different feature selection methods applied are Information Gain, Term-inverse Document Frequency , Mutual Information and Chi-Square.The following Figure 8.2 shows the performance of the first three feature selection techniques for increasing number of features. The graph has been plotted for the classification involving the Left-Wing vs Right-Wing organizations. The Right Wing organizations consists of the Nationalists and the Radical Rights. The rho value for the classification is also fixed to a value of 0.3.

From the below graph it can be observed that the performance of mutual information and information gain is almost same while the performance of tf-idf is slightly lower than that of the other two methods. Also, from this graph it can be observed that 40 -50k features are optimal for model building. If the features are increased more than 50,000 the accuracies are saturated. The trends are almost similar for the other classification techniques.

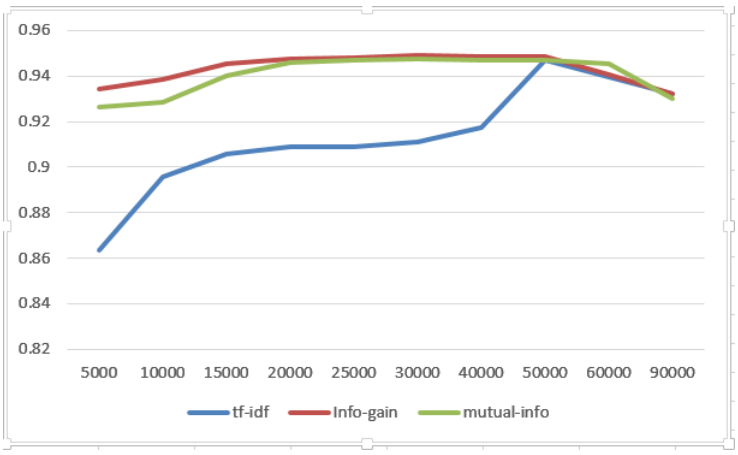

Figure 8.2: Comparison of Feature Selection Algorithms for Increasing Number of Features

Chi-square algorithm is applied over the entire features and the algorithm returns all the features with a value representing the amount of independence which is the chi score and their degree of freedom. From these results, the most important features are determined and the results obtained were less than 10k features for every classification. The accuracies obtained by running SLEP with rho value 0.3 for different classifications are shown in Figure 8.3.

8.4 Iterative Feature Selection Method

For implementing iterative feature selection, the training documents consists of both tweets and the normal documents. The tweets are collected using the twitter API from all the famous political leader profiles or organization profiles and manually labelled into their organizations. These tweets are divided into Training set and Test set in the ratio of 80:20 for all classifications.

Firstly, only the tweet documents were used to check the accuracy of the documents. Since both the domains come from the same domains, it is assumed that the model should perform well on both the domains. Table 8.2 shows the accuracies for

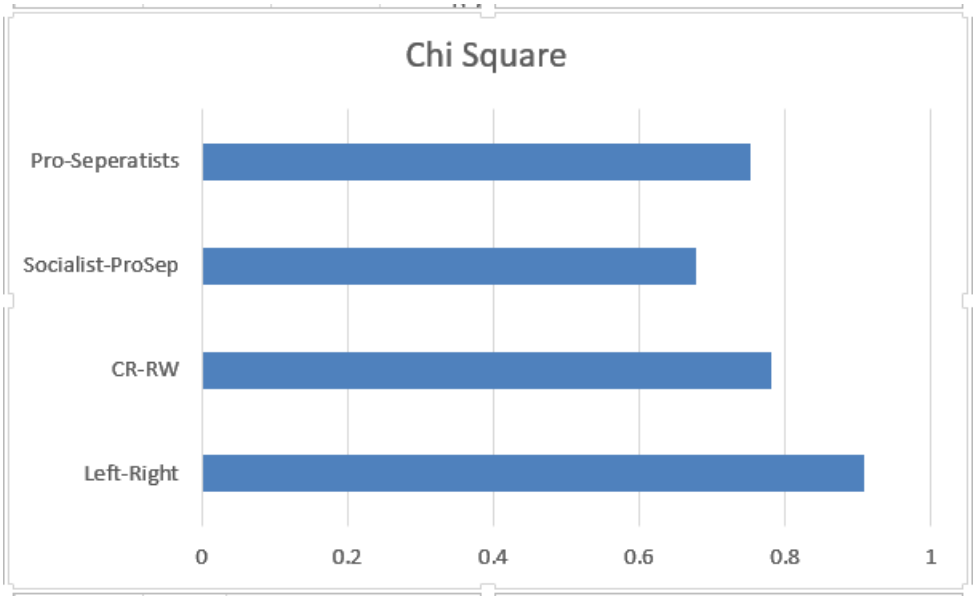

Figure 8.3: Chi Square Accuracies for all 4 Classifications

all the classifications when only the features from the initial 80% twitter documents are considered to train and estimate the model on the 20% documents. It can be observed that the accuracies obtained through this process are very low.

Classification	TwitterAccuracy
Left Wing - Right Wing	0.872
Nationalist - Radical Right	0.7199
Socialist - ProRussian/Separatists	0.8534
Pro Russian - Separatists	0.7554

Table 8.2: Accuracies When Only Twitter Features are Considered for Training

Next, I considered only the features from the social networking sites assuming that both the social networking domains follow same distribution and learning model from this combination of features should perform good on the test tweet documents. But, same as the previous observations in Table 8.2 the model didn't considerably work

well. Table 8.3 shows the performance of the model built by training on facebook and twitter documents and testing on the twitter documents. The reason is the features present in the training documents from twitter or combination of twitter and facebook are not sufficient for estimating the test documents.

Classification	Accuracies
Left Wing - Right Wing	0.896
Nationalist - Radical Right	0.74
Socialist - ProRussian/Separatists	0.8634
Pro Russian - Separatists	0.7734

Table 8.3: Accuracies When Both Twitter and Facebook Features are Considered for Training

Since, the above methods didn't work well, I have applied iterative feature selection method to learn on all training documents starting from considering only twitter documents to including all the other correctly estimated documents through iterative selection. Initially, the iterative feature selection method includes features only from tweet documents. In the following iterations, features from correctly predicted documents are added to the existing feature set. The results from iterative feature selection method for various iterations are shown in Table 8.4. The table shows results for the classification between Pro Russians and Separatist. The classification is started with 1709 training documents which included only tweets. The count of test documents or non tweet documents are 4527. The total number of features collected from the tweets are 1607. The features collected from normal documents are 51K.

The total number of features collected from this iterative feature selection method are 36790. The cross-validation accuracy for the model obtained from these features

Iteration	FeaturesAdded	Correctly_Predicted_Docs
1	21345	1154
2	9689	871
3	3582	341
4	1165	187
5	618	106
6	211	76

Table 8.4: Iterative Feature Selection Algorithm Results for Pro-Russian vs Separatists Classification. The Results Show the Documents and Features Added in Each Iteration

is 83.17 %. The comparative analysis of the accuracies obtained for cross-validation for traditional feature selection techniques vs iterative feature selection techniques are shown in the following table 8.5.

Classification	CV_Accuracy	Iterative_FS_CV
Left Wing - Right Wing	0.9482	0.965
Nationalist - Radical Right	0.8286	0.8318
Socialist - ProRussian/Separatists	0.9086	0.9356
Pro Russian - Separatists	0.8137	0.8229

Table 8.5: Average Cross-validation Accuracies in Traditional Feature Selection and Iterative Method

The increase in the accuracy for iterative feature selection method can be justified because, this technique uses more focused twitter documents initially to train model and estimate on normal documents with partial counts obtained from the scores

obtained through the estimated results. Thus, this gives a selective feature selection and every iteration these features keep getting refined and new model is estimated. It can be observed from the results that maximum number of features are selected in the first iteration and this count decreases in the following iterations. The model obtained from the iterative feature selection is estimated on the 20 % test set tweet documents and the following results in Table 8.6 are observed.

Classification	Accuracies	Features	Features Selected
Left Wing - Right Wing	0.9326	141k	65k
Nationalist - Radical Right	0.8134	60k	30k
Socialist - ProRussian/Separatists	0.9286	316k	59k
Pro Russian - Separatists	0.8027	83k	37k

Table 8.6: Accuracies of the Iterative Feature Selection Techniques on the 20% Remaining Tweets Along With the Features Selected from Total Features by the Used Model

From the above table it can be observed that the size of the feature set is reduced to approximately 50% compared to other feature selection techniques.

8.5 Visualization Tool

8.5.1 Volume Timeline Chart

The volume chart timeline is used to see the trends of a given keyword. The following volume chart is used to observe the number of tweets that discuss about EU, NATO and Russia Topics in both Latvian and Russian Languages. This has a flexibility to see the trends for a window size which allows different zoom levels.

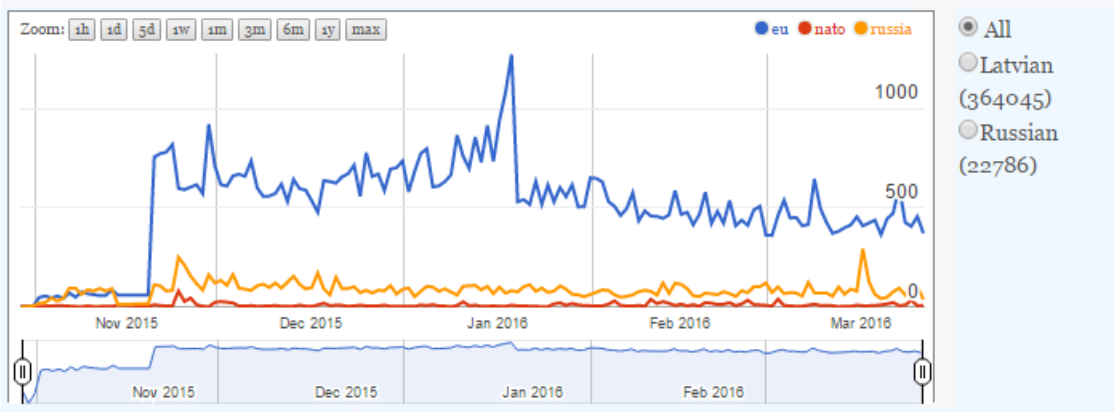

Figure 8.4: Volume Chart Showing the Count of Tweets for EU, NATO and Russia Keywords With a Sliding Window for Time Selection

8.5.2 Chord Diagram

The chord diagram is used to analyze the total users belonging to a particular organization at a given time. It also shows the users who have changed their position at any given point of time. Figure 8.5 shows Socialist, Pro-Russian, Separatist, Left-Wing, Right-Wing and Nationalist organizations and also their migration of between different organizations over time as selected in the volume chart shown in 8.5.

8.5.3 Heat Map

Heat Map is used to observe the total volume of tweets associated with a location. The information about the location is obtained from the tweets when it is collected using the twitter API. For this visualization, the main location restriction is set as Latvian region and only tweets from Latvia are mainly observed. The below heat map in Figure 8.6 shows that most of the tweets mainly come from Riga and Ogre regions.

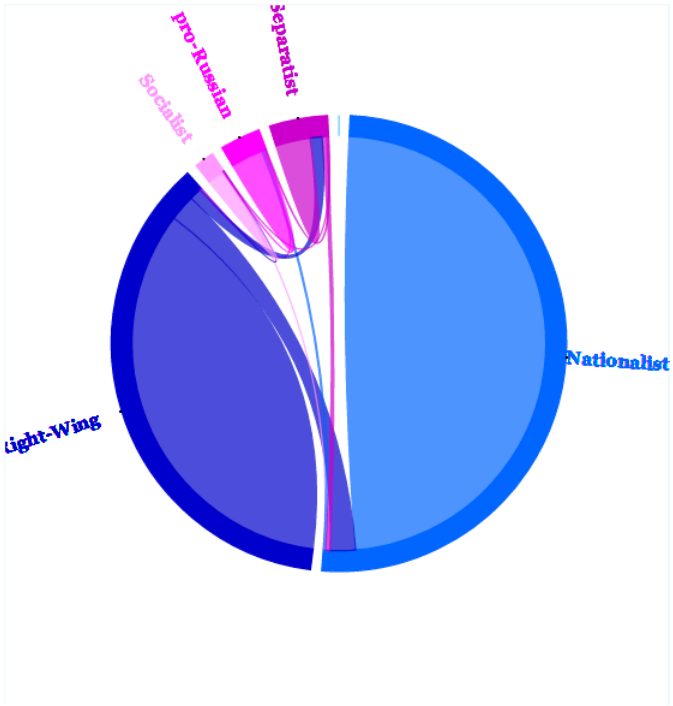

Figure 8.5: Chord Widget Showing Organization User Counts and the Shifts of Users Within Different Organizations

8.5.4 Event Timeline

The event timeline is used to represent various news articles related to the events happening around Latvia. These events are in the form of news articles and tweet hashtags. These are selected to be the top 10 most famous events and hashtags shared within all the tweets occurring each day. Figure 8.7 shows a news article from October 31st talking about NATO's military activity post Cold War in connection with Russia. The following Figure 8.8 shows a tweet with hashtag on NATO on March 11th which is among the top 10 most famous events. The event timeline is structured in the way that the slider has a time series at the bottom showing all the dates having the slider showing the list of tweets and news articles on a given date. The event which is selected from this slider is being shown in the main event area.

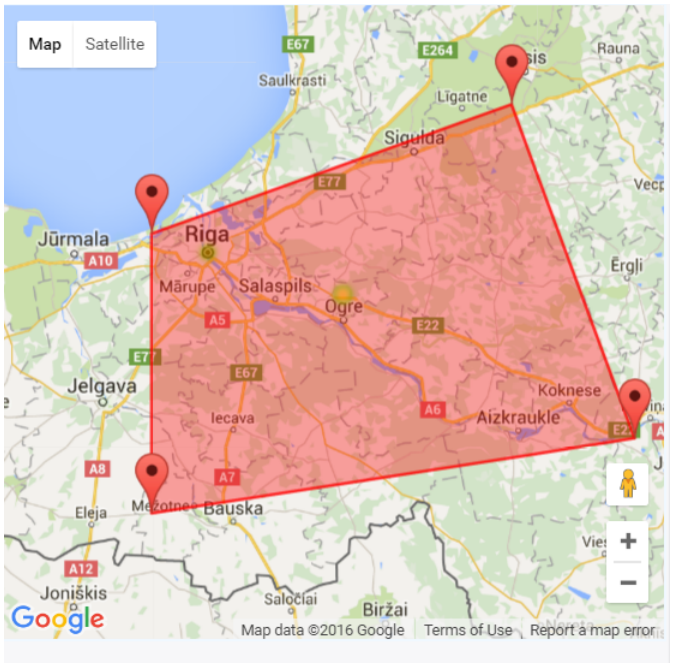

Figure 8.6: Heat Map for Tracking Down the Location of Maximum Activity

If its a hashtag, it shows the hashtag with the actual tweet containing the hashtag shown along with its date of occurrence which is available using the twitter API. If its a news articles, it shows the news title with the news content and the screen shot of the URL from an API called pagepeeker.

8.5.5 Network Diagram

The network diagram represents the information about a user and their followers which is obtained from the tweets. These followers are mainly the ones who are mentioned in a tweet or the ones whose tweets have been re-tweeted in other users tweet. Brandes algorithm is used to compute betweenness and centrality between the users and the top 100 most central users are found out for the network graph showing their followers. Each node in the network graph is clickable which on click opens a window taking to the twitter account of the user. Each user is represented with

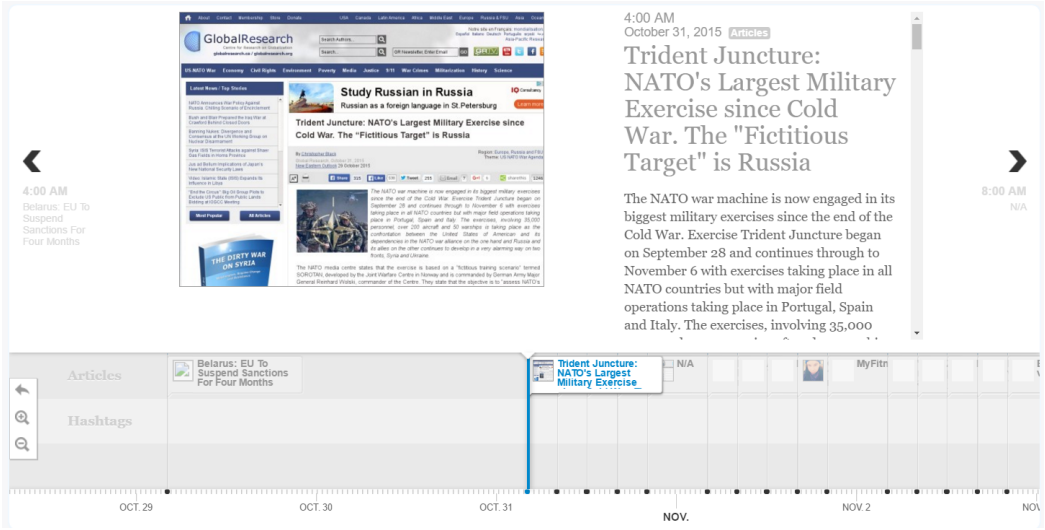

Figure 8.7: Event Time-line to View all the News Events Associated With the Tweets

a color according to their organization color from chord diagram. This also shows various clusters formed within the users for a given selection. For example in figure 8.10 on selection of Socialist organizations, a network diagram showing three clusters with screen names lauziC, belka_shi, and pimpin_is_hard are shown as central users with their followers.

8.5.6 Tweet Table

The tweet table is used to view complete list of tweets that have been talked about on a given date with all the other selections from chord, polygon for heat map. These tweets are organized in the decreasing order of their popularity.

Figure 8.8: Event Time-line to View all the Trending Hashtags Associated with the Tweets

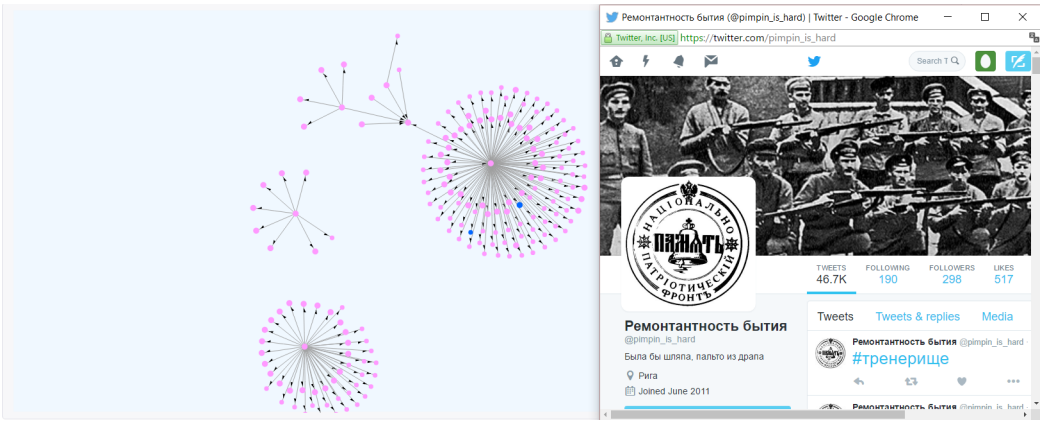

Figure 8.9: Network Diagram Showing the Connection between Different Users Through Their Retweets and User Mention Information. On Clicking a Node a Pop-up Window Shows Up With the Twitter Profile of the User Clicked.

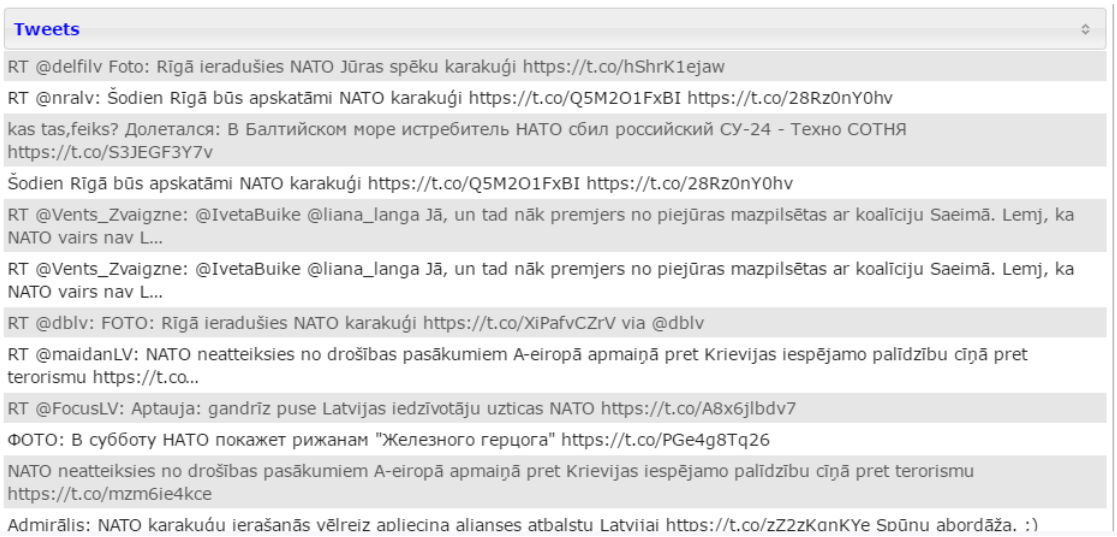

Figure 8.10: Tweet Table Used to View all the Tweets from the Selected Organizations at a Given Time

Chapter 9

CONCLUSION AND FUTURE WORK

In this thesis, I have come up with an end-to-end data modelling approach for understanding the political scenarios in Latvia. This analysis is done mainly on the social networking tweets obtained from the discriminative keywords, collected from the training model on the initial documents. The initial documents mainly comprise web articles collected from organization websites and social networking profiles of the users. Once, the estimation of model on the collected tweets is done, the predicted results are represented on an interactive dashboard which also shows other information about the users being classified from the initial tweet collection. All the previous approaches involved either Arabic or English languages and a lot a language processing tools were available to deal with the vocabulary set and reduction of the final feature list. The current thesis involves documents only restricted to Latvian and Russian languages and not much language processing tools are available to deal with these languages.

Also, the classification technique implemented is purely a binary classifier wherein from all the set of available organizations, a combination of two organizations can only decide the model for the final estimation. In the previous work, multiple binary classifiers had to be applied to analyze the difference between various organizations and the combinations that provide maximum accuracies are being selected for the final model and the estimation is done using these models. However in this scenario, the classification is well defined which starts initially with classification among Latvian and Russian groups followed by sub classifications within each group.

All the previous techniques involve applying different machine learning algorithms in-order to improve the accuracy results of the final classification. From all the previous work, it has been observed that random-forest, SVM and SLEP are the best tools available for classification. The main problem with classification involving various languages other than english is the size of the feature set. Since, the size of the features is large, the classification accuracies could be high due to overfitting of the data. In this thesis, I have compared application of various feature selection methods and iterative feature selection method that I implemented to reduce the total number of features while still maintaining the same accuracies.

Feature selection techniques like Term frequency and inverse document frequency, mutual information, chi-square and information gain have been applied. It is observed from the results that mutual information and information gain both perform well on the document set where as tf-idf has slightly lower accuracies. Another important thing to be noticed is the problem of domain adaptation. Since, the training set is a mixture of different domains while the test set is a pure collection of tweets, there will generally be a mismatch between distribution of the two domains that are being collected. I have come up with a solution to deal with this problem by implementing an iterative feature selection approach which starts with training initially on the available set of small tweet documents, Once the training is done, this model is used for estimation on the other training normal documents from which the correctly predicted documents are taken out and included into new training model. In this way iteratively, all the documents are added till no more documents can be added any further. . The performance of the iterative feature selection algorithm is also verified against the labeled tweets from the test set. It has been observed that the algorithm reduced the feature set size by 50% and gave good classification accuracies for the test set tweets.

The next stage of the thesis is to visualize the statistical results obtained from the classification. Various widgets are used to represent the activities of users over a given time period along with their interaction with other users and their locations and their organization position shift from time to time. In addition, the visualization also includes event-timeliene which is an interesting widget used to visualize all the news article url links that mainly occur in a tweet and also the most occurring hashtags from the tweets.

A lot of research is still going on in coming up with interesting data analytics solution for political scenarios. The current model doesn't support analyzing opinions of people in real-time. Dynamic model building techniques should be implemented to understand political situations in real-time. Though the iterative feature selection technique helps in selecting most relevant features from documents involving different domains, a lot of work can be done to develop better model building techniques to address the cross domain problem.

REFERENCES

Bartl, E., "Latvias integration into nato (2013)", URL http://liia.lv/en/blogs/latvias-integration-into-nato-was-the-quick-adapti/ (2013).

BBCNews, "Latvias country profile (11 february 2015)", URL http //www.bbc.com/news/worldeurope17522134 (2016).

Church, K. W. and P. Hanks, "Word association norms, mutual information, and lexicography", Computational linguistics **16**, 1, 22–29 (1990).

Fano, R. M. and W. Wintringham, "Transmission of information", Physics Today **14**, 56 (1961).

Kim, N., S. Gokalp, H. Davulcu and M. Woodward, "Looking Glass: A visual intelligence platform for tracking online social movements", IEEE pp. 1020–1027 (2013).

Kim, Nyunsu and Tikves, Sukru and Wang, Zheng and Githens-Mazer, Jonathan and Davulcu, Hasan, "MultiScale modeling of Islamic organizations in UK", pp. 13–18 (2013).

Lab, C. P. C., "Looking Glass for analyzing ISIS and Non-ISIS tweeters", URL http://129.219.60.22:8081/LG_ISIS/example/reuter/index.html (2015).

Lewis, D. D. and M. Ringuette, "A comparison of two learning algorithms for text categorization", in "Third annual symposium on document analysis and information retrieval", vol. 33, pp. 81–93 (1994).

MFARL, "Latvijas republikas rlietu ministrija. retrieved on 2 december 2011", URL http://www.mfa.gov.lv/en/policy/society-integration/integration-policy-in-latvia-a-multi-faceted-approach/ethnic-structure-and-promotion-of-national-minorities-cultural- identity (2015).

Moh'd A Mesleh, A., "Chi square feature extraction based svms arabic language text categorization system", Journal of Computer Science **3**, 6, 430–435 (2007).

Peddinti, V. M. K. and P. Chintalapoodi, "Domain adaptation in sentiment analysis of twitter", in "Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence", (2011).

RANKNL/Latvia, "Latvian stopword list", URL http://www.ranks.nl/stopwords/latvian (2016).

RANKNL/Russia, "Russian stopword list", URL http://www.ranks.nl/stopwords/russian (2016).

Yang, Y. and J. O. Pedersen, "A comparative study on feature selection in text categorization", in "ICML", vol. 97, pp. 412–420 (1997).

APPENDIX A

INITIAL TRAINING DOCUMENTS

Position	PARTY/GROUP	Political Orientation
Pro-Russian	Par Dzimto Valodu!	National Bolshevism
Pro-Russian	Saskaņa (TSP - Harmony)	Social democracy
Left-wing	Latvian Russian Union (PCTVL)	Democratic socialism
Centre	For Latvia and Ventspils	Centre
Centre-right	Unity	Centre-right
Right-wing	TB/LNNK	Conservative, Nationalist
Right-wing	National Alliance (Visu Latvijai! - VL_TBLNNK)	Conservative, Nationalist, Liberal
Position	**NGO**	**Political Orientation**
	LATO	
	Latvian Centre for Human Rights (LCHR)	
	Anti-NATO Latvia	
Pro-Russian	IMHO Club	
Pro-Russian	Latvia's Congress of Non-Citizens	Latvia's 300,000 Non-Citizens
Separatist	Latgale People's Republic	
Separatist	Latgaliya	
Pro-Russian	Russkaya Zarya	

Figure A.1: List of Political parties and NGOs

Position	POLITICIAN	Political Orientation	PARTY/GROUP	Position
	Vaira Vīķe-Freiberga		Independent	Former President
Centre-right	Edgars Rinkēvičs	liberal conservatism	Unity	Minister for Foreign Affairs
Centre-right	Sandra Kalniete	liberal conservatism	Unity	former foreign minister
Centre-right	Veiko Spolitis	liberal conservatism	Unity	Foreign Ministry adviser
Centre-right	Valdis Dombrovskis	liberal conservatism	Unity	Former prime minister
Centre-right	Artis Pabriks	liberal conservatism	Unity	Minister of Defence
Centre-right	Laimdota Straujuma	liberal conservatism	Unity	Prime Minister
	Raimonds Vējonis		Green Party	President
	Aleksandrs Bartasevics			Mayor of Rezekne
	Aivars Lembergs		For Latvia and Ventspils	Mayor of Ventspils
	Aleksandrs Mirskis		Alternative	
	Yuriy Zaytsev (Jurijs Zaicevs)			Daugavpils city council
Left-wing	Tatyana Zhdanok (Tatjana Zdanoka)	Socialist	Latvian Russian Union (PCTVL)	MEP
Left-wing	Miroslav Mitrofanov (Miroslavs Mitrofanovs)	Socialist	Latvian Russian Union (PCTVL)	Co-chairman
Left-wing	Yuriy Petropavlovskiy (Jurijs Petropavlovskis)	Socialist	Latvian Russian Union (PCTVL)	Co-chairman

Figure A.2: List of Politicians

Position	PROVOCATEUR	PARTY/GROUP	Position
Pro-Russian	Janis Kuzins	Association Against Nazism	activist
Pro-Russian	Yuriy Alekseyev (Jurijs Alekseyevs)	IMHO Club	founder
Pro-Russian	Iosif Koren (Josifs Korens)	Latvian Antifascist Committee	
Pro-Russian	Viktor Gushchin (Viktors Guscins)	Latvian Council of Civic Organisations (LCCO)	
Pro-Russian	Yelena Bachinskaya (Jelena Bacinska)	Latvia's Congress of Non-Citizens	
Pro-Russian	Aleksandr Gaponenko (Aleksandrs Gaponenko)	Latvia's Congress of Non-Citizens	co-chairman
Pro-Russian	Yelizaveta Krivtsova (Elizabete Krivcova)	Latvia's Congress of Non-Citizens	lawyer
Pro-Russian	Einars Graudins	Latvia's Congress of Non-Citizens	senior member
Pro-Russian	Vladimir Linderman (Vladimirs Lindermans)	Par Dzimto Valodu!	leader of Latvian branch Russia's National Bolshevik Party
Sep	Anatoliy Matyukovskiy	Russia's Other Russia group	separatist in Ukraine
Pro-Russian	Illarion Girs (Illarions Girs)	Russkaya Zarya	co-leader
Pro-Russian	Yevgeniy Osipov (Jevgenijs Osipovs)	Russkaya Zarya	member
Sep	Aijo Beness		separatist in Ukraine
Sep	Stanislavs Bukains		separatist in Ukraine
Sep	Grigoriy Kosnikovskiy (Grigors Kosnikovskis)		separatist in Ukraine
Sep	Valentin Milyutin		separatist in Ukraine
Sep	Dmitriy Prokopenko (Dmitrijs Prokopenko)		separatist in Ukraine
Sep	Vyacheslav Vysotskiy		separatist in Ukraine

Figure A.3: List of Provocateurs

APPENDIX B

STOP WORDS FOR LATVIAN LANGUAGE

aiz	un	nedz	esi	kluvi
ap	bet	tik	esam	kluva
ar	jo	nevis	esat	kluvam
apakš	ja	turpretim	bušu	kluvat
arpus	ka	jeb	busi	klustu
augšpus	lai	iekam	bus	klusti
bez	tomer	iekam	busim	klust
caur	tikko	iekams	busiet	klustam
del	turpreti	kolidz	tikt	klustat
gar	ari	lidzko	tiku	klušu
iekš	kaut	tiklidz	tiki	klusi
iz	gan	jebšu	tika	klus
kopš	tadel	talab	tikam	klusim
labad	ta	tapec	tikat	klusiet
lejpus	ne	neka	tieku	varet
lidz	tikvien	itin	tiec	vareju
no	vien	ja	tiek	varejam
otrpus	ka	jau	tiekam	varešu
pa	ir	jel	tiekat	varesim
par	te	ne	tikšu	var
par	vai	nezin	tiks	vareji
pec	kamer	tad	tiksim	varejat
pie	ar	tikai	tiksiet	varesi
pirms	diezin	vis	tapt	varesiet
pret	droši	tak	tapi	varat
priekš	diemžel	iekams	tapat	vareja
starp	nebut	vien	topat	vares
šaipus	ik	but	tapšu	
uz	it	biju	tapsi	
vinpus	tacu	biji	taps	
virs	nu	bija	tapsim	
virspus	pat	bijam	tapsiet	
zem	tiklab	bijat	klut	
apakšpus	iekšpus	esmu	kluvu	

Figure B.1: List of Stop Words for Latvian Language

This list is obtained from RANKNL/Latvia (2016)

APPENDIX C

STOP WORDS FOR RUSSIAN LANGUAGE

а	можно	мало	наши
е	может	надо	ничего
и	можхо	один	начала
ж	мор	одиннадцать	нередко
м	моя	одиннадцатый	несколько
о	моё	назад	обычно
на	мочь	наиболее	опять
не	над	недавно	около
ни	нее	миллионов	мы
об	оба	недалеко	ну
но	нам	между	нх
он	нем	низко	от
мне	нами	меля	отовсюду
мои	ними	нельзя	особенно
мож	мимо	нибудь	нужно
она	немного	непрерывно	очень
они	одной	наконец	отсюда
оно	одного	никогда	в
мной	менее	никуда	во
много	однажды	нас	вон
многочисленное	однако	наш	вниз
многочисленная	меня	нет	внизу
многочисленные	нему	нею	вокруг
многочисленный	меньше	неё	вот
мною	ней	них	восемнадцать
мой	наверху	мира	восемнадцатый
мог	него	наша	восемь
могут	ниже	наше	восьмой

Figure C.1: List of Stop Words for Russian Language

This list is obtained from RANKNL/Russia (2016)

вверх	всех	более	без	десять
вам	всею	должно	день	десятый
вами	всю	пожалуйста	занят	ею
важное	вся	значит	занята	её
важная	всё	иметь	занято	их
важные	всюду	больше	заняты	бы
важный	г	пока	действительно	еще
вдали	говорил	ему	давно	при
везде	говорит	имя	девятнадцать	был
ведь	года	пор	девятнадцатый	про
вас	году	пора	девять	процентов
ваш	где	потом	девятый	против
ваша	да	потому	даже	просто
ваше	ее	после	алло	бывает
ваши	за	почему	жизнь	бывь
впрочем	из	почти	далеко	если
весь	ли	посреди	близко	люди
вдруг	же	ей	здесь	была
вы	им	два	дальше	были
все	до	две	для	было
второй	по	двенадцать	лет	будем
всем	ими	двенадцатый	зато	будет
всеми	под	двадцать	даром	будете
времени	иногда	двадцатый	первый	будешь
время	довольно	двух	перед	прекрасно
всему	именно	его	затем	буду
всего	долго	дел	зачем	будь
всегда	позже	или	лишь	будто

Figure C.2: List of Stop Words for Russian Language

будут	каждое	тобой	сама
ещё	каждая	собою	сами
пятнадцать	каждые	тобою	теми
пятнадцатый	каждый	сначала	само
друго	кажется	только	рано
другое	как	уметь	самом
другой	какой	тот	самому
другие	какая	тою	самой
другая	кто	хорошо	самого
других	кроме	хотеть	семнадцать
есть	куда	хочешь	семнадцатый
пять	кругом	хоть	самим
быть	с	хотя	самими
лучше	у	свое	самих
пятый	я	свои	саму
к	та	твой	семь
ком	те	своей	чему
конечно	уж	своего	раньше
кому	со	своих	сейчас
кого	то	свою	чего
когда	том	твоя	сегодня
которой	снова	твоё	себе
которого	тому	раз	тебе
которая	совсем	уже	сеаой
которые	того	сам	человек
который	тогда	там	разве
которых	тоже	тем	теперь
кем	собой	чем	себя

Figure C.3: List of Stop Words for Russian Language

тебя	ты
седьмой	три
спасибо	эта
слишком	эти
так	что
такое	это
такой	чтоб
такие	этом
также	этому
такая	этой
сих	этого
тех	чтобы
чаще	этот
четвертый	стал
через	туда
часто	этим
шестой	этими
шестнадцать	рядом
шестнадцатый	тринадцать
шесть	тринадцатый
четыре	этих
четырнадцать	третий
четырнадцатый	тут
сколько	эту
сказал	суть
сказала	чуть
сказать	тысяч
ту	

Figure C.4: List of Stop Words for Russian Language