Social Media Mining of the Icelandic Blogosphere

Sindri Freyr Sigursteinsson

Kongens Lyngby
IMM-M.Sc.-2011-68

ABSTRACT

This thesis presents experiments related to the Icelandic Twitter network. It is based on an analysis made from two perspectives, a general analysis of the overall network, including its users, followed by a more detailed sentiment analysis. The main objective of the general analysis was to identify features of the Icelandic network as a whole and characteristics of its Icelandic users. The growth rate of the Icelandic Twitter network was studied as well as tweeter activity among Icelandic Twitter users. Additionally, the relationship between different variables was viewed, for example between activity level and the numbers of followers, which has been used in prior work for identifying influential tweeters. The purpose of the sentiment analysis was to detect and extract the emotional content of Icelandic tweets, whether they are positive, neutral or negative.

ACKNOWLEDGEMENTS

First I would like to thank my supervisor, Finn Årup Nielsen, for all of our meetings, his opinions, and ideas and how quickly he always replied to my emails throughout the thesis period. My special thanks go to my lovely, almost nine months pregnant, and fiancée for being so supportive, encouraging and helpful during this time and for tolerating me when the work was not coming around as expected. At last, my two beautiful daughters, Svava Marín Sindradóttir and Jenný Ísabel Sindradóttir, also have to be mentioned because I have not been able to spend as much time with them lately as I would have wanted, but that is about to change.

Table of Contents

Table of Figures

1 Introduction

1.1 Twitter

Microblogging is a form of communication that allows users to share information about their activities or opinions in short posts. Twitter, which was launched in 2006, has become a popular microblogging platform in which the ever increasing usage of microblogging has been attributed to. Generally known as tweets, microblog posts are immensely brief compared to regular blog posts, with a maximum length of 140 characters. This simple form of communication enables users to broadcast information to millions of people around the world through short text updates on great variety of topics (Bollen, Mao and Pepe, 2009), ranging from daily activities to events, news stories or current affairs and other interests (Java, Song, Finin and Tseng, 2007). Considering this, a description of Twitter as a forum for various usages is quite sensible.

One of the above mentioned usages of Twitter involves conversations, which reflects users directing tweets to other specific users, namely their so-called *followers* (Java et al., 2007) who receive all tweets posted by those they follow (Kwak, Lee, Park and Moon, 2010). Twitter is different from other social network services in that it provides a social-networking model which enables users to choose who they want to follow without seeking any permission. Conversely, users may also be followed by others without the requirement of seeking any permission beforehand (Weng, Lim, Jiang and He, 2010). This relationship of following and being followed by others does not require reciprocation either. This means that a tweeter can follow another tweeter without him or her needing to replicate by following back (Kwak et al., 2010). It differs, amongst followers, who they decide to follow and also how many they decide to follow. Some follow only personal friends while others follow people they do not know personally, but find interesting for some reason, for example celebrities and politicians. Some follow thousands of tweeters, while others follow only few of them (Boyd, Golder and Lotan, 2010). This interplay between tweeters and their followers explains why Twitter is not only described as a microblog, but as a social network site as well (Thelwall, Buckley and Paltoglou, 2011). Studies have revealed that Twitter users also exploit it as a means of communication and social networking (Java et al., 2007).

A unique characteristic of Twitter involves *re-tweeting*, when a user reposts a tweet that already has been written by another user. This is generally done in the purpose of spreading information to the poster's followers (Boyd, Golder and Lotan, 2010), but it could also be done for other reasons, including making it easier for them to find older posts (Thelwall, Buckley and Paltoglou, 2011). Re-tweeting has the effect of bringing new users into a thread, inviting them to engage without addressing them directly. Thus, spreading tweets serves as to engage with others as well as to get various messages out to new users (Boyd, Golder and Lotan, 2010), which can happen fast since a re-tweeted tweet is expected to be able to reach on average of 1000 users (Kwak et al., 2010).

Social networks and microblogs such as Twitter have without a doubt become a popular tool of choice for dissemination of information, communication and networking. With a global reach and increasing amount of adopters, Twitter can be used to broadcast information efficiently and at a fast rate. Different groups of people have become interested in adopting this new platform and for different reasons. It has reached the attention of young and old, from politicians to business people and for reasons ranging from staying close to friends and family to the use in citizen journalism. This gives a relatively clear idea of the great amount of information that can be harnessed from sites such as Twitter, which makes it an interesting field for research of different kinds since the information can be used for various different purposes. Twitter data can for example shed light upon epidemic behavior as it can be used in relation to economic analysis, decision support or policymaking (Cheong and Lee, 2009).

1.2 Structure of the thesis

The thesis is organized as follows. In chapter two, literature review will be given. Here, the main objective is to reflect on previous studies related to Twitter network analysis and sentiment analysis. In chapter three, data collection and preprocessing will be introduced, where there is focus on the Twitter API as well as structure and data storage. Under chapter four, methodology will be discussed in details. Here, the main focus is on Twitter data used to identify Icelandic tweeters from their properties and to detect emotion in

Icelandic tweets. Chapter five provides results from studying both the Icelandic Twitter network in general and from the perspective of sentiment analysis. In chapter six, discussion and conclusions are provided. Here, the main results will be iterated and discussed from various angles. Factors affecting performance, possible challenges and suggestion for future work will also be discussed.

2 Literature review

2.1 Twitter network analysis

The rising popularity of online social networking services has evoked an interest among researchers in studying their attributes and activities. Twitter is one of those services that has attracted much attention since its launch and thereupon, stimulated an interest for carrying out various researches at different levels (Kwak et al., 2010). Prior work of Java et al. (2007) on Twitter as a microblogging platform emphasized the Twitter user spread in terms of geographic location, social networks a user belongs to and the intentions of a user when microblogging. Krishnamurthy, Gill and Arlitt (2008) focused on identifying properties of distinct classes of Twitter users and their behaviors as well as on looking into the growth of the Twitter user network. They classified users by follower/following counts, means and mechanisms of their engagement and volume of use, i.e. the number of tweets per time period. Kwak et al. (2010) also studied Twitter, but with a focus on the entire Twittersphere.

The relationship between tweeters and their followers has also gained an attention from the research community, where topological and geographical properties of the social network formed by Twitter users have been studied (Java et al., 2007). There can for example be found prior work that aims at identifying influential tweeters by using the number of followers they have as in indication of influence. This calls upon the assumption that the more followers a tweeter has, the more impact he or she makes in the Twitter context because of the increased popularity generated from having many followers. Another metric that is similar to this one uses a ratio between the number of a tweeter's followers and the number of friends, as a following relationship can be characterized by the so-called friend, who is the tweeter whose updates are being followed, and the follower, or the one who is following. Yet another metric uses a ratio of the attention a tweeter gets to published tweets. It could for example be in the form of re-tweets or comments on relevant tweets (Weng et al., 2010).

In order to identify influential tweeters the most common method used in prior work involves an application of the PageRank algorithm, which measures tweeters' influence

with only link structure of the network taken into consideration (Brin and Page, 1998). Kwak et al. (2010) are among researchers who have focused on finding influential tweeters with the use of PageRank algorithm were they were ranked by the number of followers.

2.2 Sentiment analysis

Sentiment analysis, also known as opinion mining, is an area of computational studies that addresses opinion-oriented natural language processing. It has been described as the extraction of opinions from text at various levels, such as document, sentence or phrase levels (Pang and Lee, 2008). The most common one is the document level, where for example positive reviews are distinguished from negative ones, but there has also been focus on sentiment analysis related to the sentence and phrase levels (Wilson, Wiebe and Hoffmann, 2005). The research field of sentiment analysis, or sentiment classification, has been gaining an attention lately and a range of topics have been studied from this perspective, such as movie reviews (Pang, Lee and Vaithyanathan, 2002), product reviews (Na, Sui, Khoo, Chan and Zhou, 2004) and news and blogs (Bautin, Vijayarenu and Skiena, 2008). Common approaches of previous work have included focusing on either the subjective nature of text, i.e. determining whether it is subjective or objective, or the identification of polarities (Pang and Lee, 2008). This might include word sentiment scoring, where the aim is to identify the sentiment scores of single words, or sentiment amplification and negation, where sentiment strength on amplifying words are modified and sentiment scores on negated words are reversed (Heerschop, van Iterson, Hogenboom, Frasincar and Kaymak, 2011).

According to Boyi, Hens, Deschacht and Moens (2007) sentiment classification mainly predicate upon two techniques, machine learning techniques and symbolic techniques. Similarly, Thelwall, Buckley and Paltoglou (2011) place emphasis on full-text machine learning and lexicon-based methods as common sentiment analysis methods but add linguistic analysis as being among the three most common ones. This project is based on the machine learning and the lexicon-based approaches and therefore those two will be described in further detail below.

2.2.1 Machine learning approach

Machine learning approach involves constructing a model from a training corpus, which basically is an electronically stored set of texts (Boyi et al., 2007). In order to train an algorithm to identify features that associate with positive, negative and neutral categories, such set of texts, annotated for polarity by human coders, are used (Pang, Lee and Vaithyanathan, 2002).

An important part of classification of documents involves making a decision regarding the choice of the feature set (Boyi et al., 2007). It is typical that the text features used are sets of all words, word pairs and word triples (Pang, Lee and Vaithyanathan, 2002). The approach to feature selection that has been regarded as the most classic one involves the use of unigrams, which places emphasis on single words. This approach can be best described as a representation of documents as a feature vector, where the elements designate the presence or frequency of a particular word. Thus, the document in question is represented by its keywords. If, however, the features in a given document representation are for example pairs (bigrams) or triples (trigrams) instead of single words the use of n-grams apply (Boyi et al., 2007). When a decision about the feature set has been made, the trained algorithm can search for the same features in new texts in order to predict their polarity (Pang, Lee and Vaithyanathan, 2002).

Classic supervised learning techniques such as Naïve Bayes, Support Vectore Machines (SVM) and Maximum Entropy (MaxEnt) can be used to train a classifier for sentiment recognition in texts. Applying such algorithms means the use of labeled training corpus in order to learn a certain classification function. In this project experiments were done by using one of the above mentioned classifier, the Naïve Bayes. This is a classifier that constructs a model by fitting a distribution of frequencies of each feature for all the documents in question (Boyi et al., 2007). It has been considered as a rather simple model, but despite its simplicity it has performed well in prior work on text categorization (Manning and Schuetze, 1999) and more specifically in applications related to opinion mining (Pang, Lee and Vaithyanathan, 2002).

2.2.2 Lexicon-based approach

Lexicon-based approach involves the use of manually crafted rules and lexicon. It is sometimes referred to as the bag-of-words approach and has been described as the simplest representation of a text. It identifies a document as a list of single words, or lexicon, with no regard for whether or not there can be found relations between them. The sentiment of every single word is determined and the outcome can then be compared with various aggregation functions, such as average or sum. So basically, the lexicon approach involves creating a lexicon and scoring each word for valence in order to combine text with the list of words (Boyi et al., 2007). Word lists for this approach can be created manually (Tong, 2001) or automatically (Hatzivassiloglou and McKeown, 1997).

SentiStrength is a lexicon-based algorithm designed by Thelwall, Buckley, Paltoglou and Cai (2009) and used to classify for positive and negative sentiment strength in short informal English text. Similar to the lexicon approach, the core of this algorithm is the usage of a list of sentiment words. The work of Thelwall et al. (2009) focused on testing SentiStrength on a set of MySpace comments where the aim was to identify the strength of sentiment on a scale from 1, meaning no sentiment, to 5, meaning very strong positive or negative sentiment. The results where then compared to machine-learning approaches in order to see which performed better in the case of short informal text. Originally developed for MySpace comments, Thelwall, Buckley and Paltoglou (2011) also tested the SentiStrength algorithm in their prior work on Twitter statuses. They used it to classify the sentiment strength of different tweets and, as in the case of the studies on MySpace comments, it turned out to perform well. This does, however, not come as a surprise since the algorithm is tailored to the use on short informal texts and to consider particular characteristics of a text with a length limit, such as slang and abbreviations.

3 Data collection and preprocessing

3. 1 Twitter API

As this project relies primarily on data from Twitter, their web based API, an interface that offers methods used to collect information, was used. To be able to use the interface, a registered Twitter account is needed and an application has to be registered at Twitter. The API is a powerful tool, even though it has its limitations regarding responses. The number of requests to the service is limited to 350 requests per hour and therefore collection of data happens at a much slower pace than it could be if it were not for this particular limitation.

For simplicity and convenience for developers using Twitter a number of wrappers around the Twitter API have been written. As this project is implemented with the programming language Python a focus was put on searching for python wrappers around the interface. Following an informal testing of different wrappers, one named tweepy, a Twitter API library for the Python programming language, was chosen as it satisfied certain needs, for example for quickness and efficiency of responses. The API offers a number of methods, for example in relation to the creation and deleting of tweets. However, many of those were left unused and a focus was put on four functions that provide data regarded as particularly important for this project. Those involve the collecting of tweeter data, the relations between tweeters and their tweets. Following is a brief description of the usage of the four functions:

- For each Icelandic tweeter the function *get_user* was used. It returns a dataset containing core information about the tweeter requested and stores information in the database. This function was used at the beginning of the data collection process, when gathering information for the base tweeters and it was also used to update the information about the Icelandic tweeters.
- Information about each tweeter's followers and friends was downloaded by using the tweepy functions *followers* and *friends.* These functions return a list of other tweeters that are stored in the database as well as the relations between them and the one that is being queried. Both of these functions return at most 100 tweeters at a time so if there are more tweeters following or being followed they have to be invoked a number of times for each tweeter.

- The tweepy function *user_timeline* is used to download statuses for each Icelandic tweeter. It is a function that returns at most 200 statuses at a time but by iteration it is possible to get the 3200 newest tweets.

3.2 Structure

Among various batches in the overall programming structure of this project there are two that use the above mentioned API functions for collecting data. One is to collect information about tweeters and their relations to each other. A part of this collection process is to filter the Icelandic tweeters from other tweeters, which will be described further in chapter 4.1. The other batch downloads tweets, detects whether they are written in Icelandic or English and analyzes them sentimentally. This part will be described in chapter 4.2.

Figure 1 reflects a high-level overview of the overall structure of programs and services that were implemented to solve various different tasks that all relate to this project. For the purpose of explaining the functionality of the structure, it can be split vertically into the three following sub-structures:

- The search part is responsible for searching Twitter for new Icelandic tweeters, their friends and followers, the relations between them and their tweets. It is also responsible for analyzing each downloaded tweet sentimentally. The functionality of the sentiment analyzer will be described in more details later.
- The batch part has a number of batches that run on a regular basis where each has a different purpose. The batches are collecting scoring information from external web services other than Twitter, calculating some scores among the Icelandic tweeters and mining the most common topics mentioned in the tweets.
- The web part contains functionality for displaying different results in a graphical manner on a web page. The web is only used to display information without any ability to interact.

All these sub-structures rely on the same database and often on the same data. For simplicity, a base class was created in which the data base classes inherited their data base

connection from. A helper class, which contains a number of functions that are used by more than one class and have common functionalities, was also implemented. In addition, a number of data classes were created and used at various levels in the structure. Each of the sub models can also be split horizontally into the three layers of classes described below, where each class has a different purpose:

- The database level has the SQL commands and queries for interacting with the database and keeps their execution separated from other parts of the model. When the functions in the database classes have a return value, it is always a tuple of database records.
- The core functionality is placed in the classes on the business level. The functional execution goes through these classes and they combine all the classes in each sub module. They get invoked by function calls from classes on the interface level, they use the database classes to retrieve, insert and update data, and finally they use data classes for new data type constructions as well as working with the data. Additionally, they implement all external function calls.
- Interface classes are the ones that are supposed to be invoked by the number of batches that are running for collection and calculation of data and externally by the web page.

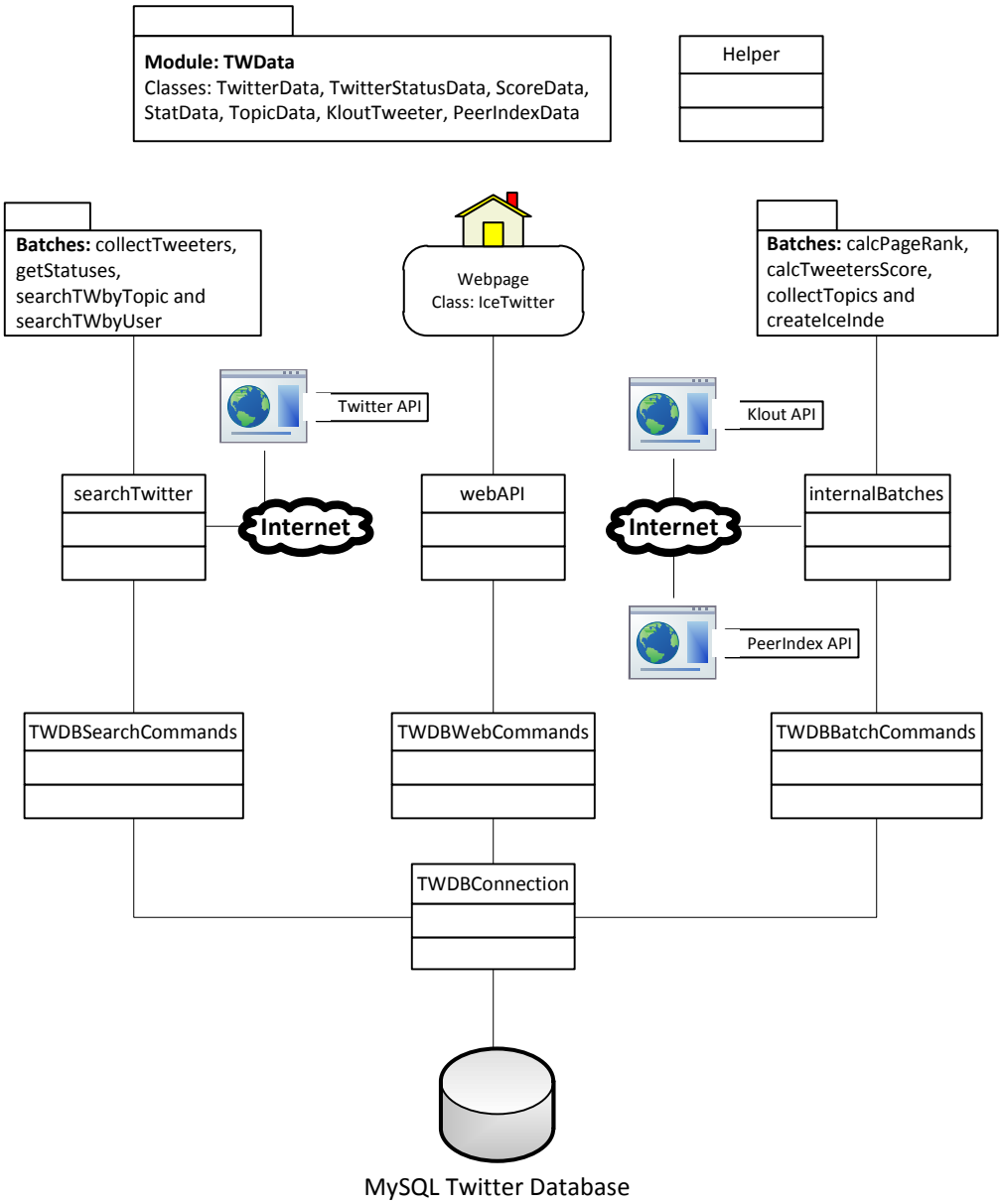

Figure 1 - Overview of the Structure

3.3 Data storage

The nature of this project calls upon a storage of a large amount of data and MySQL database server was used for that purpose. Following is a brief description of the tables created and used for this project in addition to a database schema shown in **Figure 2**.

- *Twitter, TwitterStatus* and *TwitterFriends* contain information about tweeters, their tweets and their relations. All those three tables store data that is downloaded from Twitter.
- *ScoringMethod* and *scoring* store the sentiment score calculated for each Icelandic tweet.
- *PageRank* and *PageRankRun* store daily calculation of the page rank among the Icelandic tweeters.
- *KloutTweeters* and *KloutInfluence* store scores and other information from the klout service.
- *PeerindexTweeters* stores scores and other information from the peer index service.
- *ManualScore* stores information about the tweets that have been manually labeled and used as training data.
- *TopicRun* and *Topics* store topic mining information characterized as positive, negative and neutral.

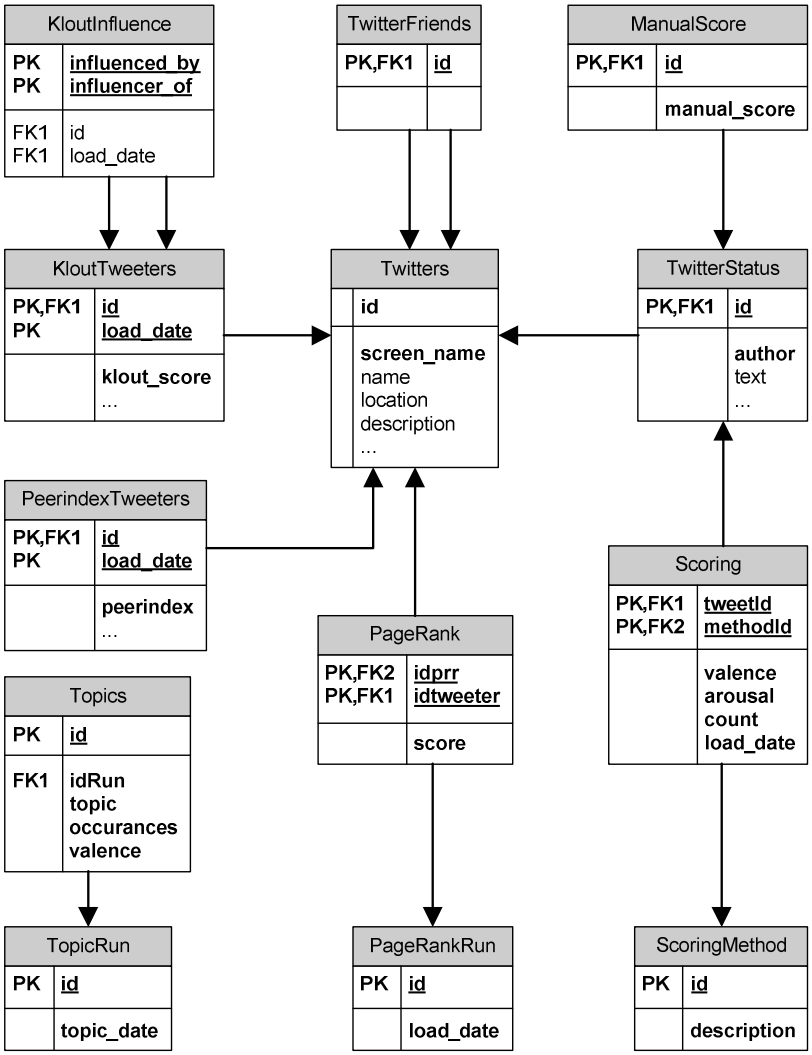

Figure 2 - Database schema

4 Methodology

4.1 Filter for Icelandic tweeters

The process that collects tweeters gets core information about a single tweeter, lists of all his followers and friends and works with the data and stores it. A list of 483 manually identified Icelandic tweeters, defined as the base Icelandic tweeters, was used for the purpose of collecting more Icelandic tweeters. Before the data related to a new tweeter was stored the tweeter's profile was run through an Icelandic tweeter filter which will be described in details below. The process was run continuously for all tweeters that were assumed to be Icelandic. During this process the set grows if new tweeters are assumed to be Icelandic.

To differentiate Icelandic tweeters from other tweeters a number of properties from the tweeter's dataset, retrieved from the Twitter API, were used to determine whether or not they were Icelandic. A functionality to score all tweeters that passed through the process was created where four properties were scanned for certain patterns. They are; *location, name, description* and each tweeter's newest *tweet*. To verify whether the properties pass the Icelandic filter three Icelandic files were created for the properties to be compared to. In addition, one English file was used to check whether the tweets were written in English. Then the score from each comparison was stored separately. Following is a description of the four files:

- *icelandic_locations*; a file that contains a list of common locations in Iceland. Names of locations that may exist in other countries were left out.
- *icelandic_names*; following a number of experiments this file ended up only containing the most common female surname's ending in Icelandic, (dóttir/dottir). This ending is also used in the Faroese language but it is not as common as it is in the Icelandic language. The Faroese are also substantially fewer than Icelanders.
- *responsible_Icelandic*; a file containing pairs of a word and an integer between minus 5 and plus 5. As this is a file to detect Icelandic, all Icelandic words were given a positive score and non Icelandic words were given a negative score. Words that occur both in Icelandic and other languages were left out. This file was also used later to determine whether each Icelandic tweet was written in Icelandic or not.

- *responsible_English*; a file similar to the *responsible_Icelandic* file except that English words have a positive score and Icelandic words have a negative score. All other pairs keep their negative score unchanged. This file was also used later when tweets were collected.

The already introduced location property was compared to the content of the *Icelandic_location* file and similarly the name property was compared to the content of the *Icelandic_name* file. If a match between the properties and the files was found, a positive score for each comparison was stored. The description text was split up and compared to the content of two files, the *responsible_Icelandic* file and the *responsible_English* file. The value of the comparisons was summarized and stored for both English and Icelandic. The tweeter's dataset also contained the tweeter's newest tweet which was compared to the *responsible_Icelandic* file. If any of the Icelandic scores were positive after this filtering the tweeter was assumed to be an Icelandic one.

The property *url* is also among properties that might give information about the nationality of a tweeter because of the country code. However, that particular property seemed to give unreliable results so it was removed from the filter. The other four, which were not removed from the filter, are of course not completely reliable either. To give examples, it is not uncommon that Icelanders use nicknames instead of their real names, leave the location property empty or write their profiles description and tweets in English. Therefore, to get more Icelanders into the set of Icelandic tweeters, it called upon some manual interruption where they were added by hand. If one or more of the scores were positive then the process assumed that the tweeter was an Icelandic one. Tweeters, who are not Icelanders, can also state that they live in Iceland and therefore they would pass the filter, which would also give unreliable results. This also called for some interruption where they were removed manually.

When the filtering of Icelandic tweeters had been implemented it gave various kinds of information about the Icelandic Twitter network that was for example used to identify possible relationships between different variables. What was used for this purpose included three variables provided by Twitter, i.e. followers count, friends count and numbers of tweets, and other three, calculated after the filtering process, i.e. the number

of Icelandic followers that Icelandic tweeters have, the ratio between international and Icelandic followers and page rank calculations. In addition to those six, two more scores, klout and peerindex, were collected from two different external scoring services that use those variables to identify how influential a tweeter is. Results from those calculations can be seen in chapter 5.1.

4.2 Sentiment analysis

Every single downloaded tweet was analyzed sentimentally as long as it was assumed to be written in Icelandic or in English. The two methods used for this analysis were a machine learning approach, where the focus was on the use of the Naïve Bayes algorithm, and lexicon-based approach, similar to SentiStrength. The former approach applied for tweets as long as they were written in Icelandic but the latter one applied for tweets written both in Icelandic and English.

When analyzing each tweet six different scores were stored. Five of them apply to the lexicon-based approach, where valence and arousal value were stored, and one applies to the Naïve Bayes classifier, where only the class of the analysis was stored. These scores will be described later but since they were calculated by using combinations of different files, each file will be described briefly first:

- AFINN-111 (Nielsen, 2011); a file that contains a list of English words and integers separated by tabs. The integer represents the valence rating for the words, between minus 5 and plus 5, where negative numbers represent negative emotion and positive numbers represent positive emotion.
- ice_emo; a file that contains a list of Icelandic emotional words or a beginning of a word with the symbol '*' added to it. The '*' is used because of the nature of Icelandic language being able to have various different endings to a single word and because there cannot be found a stemmer for it. Thus, the usage of the '*' symbol prevents that the same word is written many times, with different endings. For every word in the list which has the '*' symbol as an ending, the text was analyzed with the purpose of finding out if the part of the word that becomes before the '*' symbol, appears in the text as a beginning of a word or a complete word. The above mentioned list is built up of various pairs of a word and an

integer, separated by tabs. The integer, which is between minus 5, meaning negative, and plus 5, meaning positive, reflects the rating for valence. The main source of words came from the file AFINN-111, with an addition of Icelandic slang and other common emotional words that the author added.

The frequency distribution of valence scores in the emotion file can be seen in **Figure 3**, which reflects the scores of both words and beginning of words. The file has a total of 540 positively scored words and a total of 1110 negatively scored words. Thus, more than 2/3 of the emotional words are scored as negative, which means that it has a negative bias. As can be seen, most of the negative words were scored with minus 2 and most of the positive words were scored with plus 2.

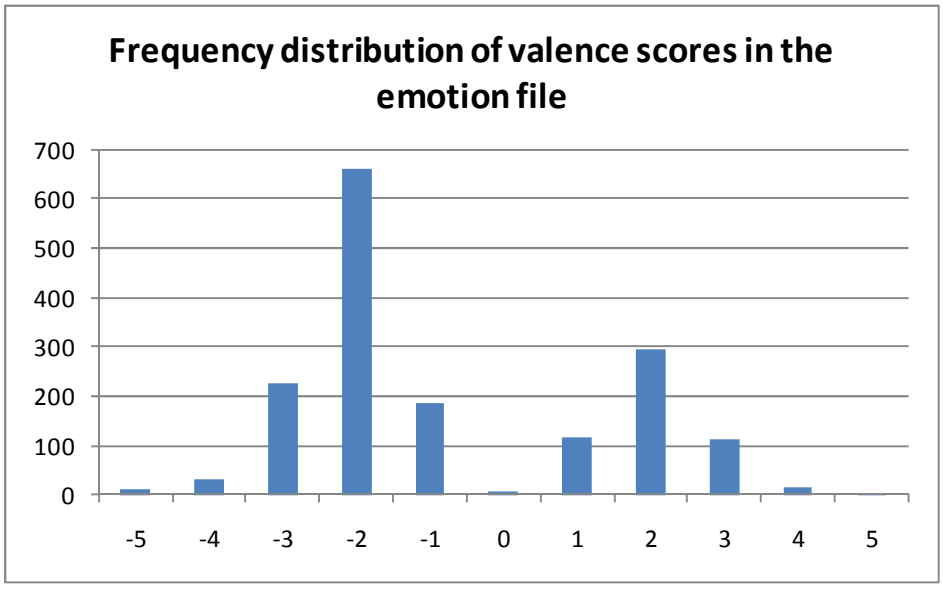

Figure 3 - Valence score distribution

- ice_neg; a file that contains a list of negation words in Icelandic. It is used to change the valence score of every two words after a negation word in a text. Thus, a word with the valence of plus 3 changes to the valence of minus 3. This reversing of polarity is sometimes referred to as switch negation (Saurí, 2008).
- ice_booster; a list of Icelandic booster words which are used to increase the valence of the word that follows the booster word. It should be kept in mind that in Icelandic language positive booster words are sometimes used with negative

emotional words and vice versa. This has the effect of increasing the valence score in emotional text.

- Emoticons; a list of the most common emoticons, both positive and negative. The list is built up of various pairs of emoticons and integers, separated by tabs. The integer, which can be minus 1 or plus 1, reflects whether a particular emoticon is positive or negative. Since emoticons are not language based they have exactly the same meaning in Icelandic as in other languages.

Now, the five different scores that apply to the lexicon based approach and that were found by using combinations of these files will be described. The first one, the emoticon score, was found by using the emoticon file to detect whether emoticons appeared in the text or not. The valence score for each emoticon was either plus 1 or minus 1. The occurrences of emoticons were added together and that score was stored. When the emoticon score had been found the text was cleared by removing every single symbol, besides alphabetical letters, links to web-pages and mentions, i.e. @, from the text. In addition to that, the text was formed into lower case letters. After this clearing, the second score, the emotional score, was found by comparing the text to the emotional file. Then the sums of all the valence and arousal scores from the text were stored as the emotional score.

The third score, the negation score, involves usage of the negation file. To be able to find this score the text was compared to the negation file to detect if negation occurred in the text. In cases where it did not occur, the emotional score for the text was exactly the same as the score for the text where only the emotional file was used. If negation did occur, however, the valence scores for the two words following the negation word were multiplied by minus 1. That score was summarized and stored separately. The fourth score, the booster score, was found by using the same approach as used when finding the negation score. The only difference was that another file was used, i.e. the booster word file, and instead of multiplying the valence score with minus 1 it was multiplied by 2. This had the effect of increasing the polarity of the words in the text. The fifth score, the booster negation score, was calculated by using the three files described above, the emotional file, the booster word file and the negation file.

As mentioned before, the sixth score was different from the five just described because it was calculated by using the Naïve Bayes classifier which applies to the machine learning approach instead of the lexicon based approach. In order to get a score from the Naïve Bayes, a trained classifier is needed. The training and test data for the classifier used consists of 1142 manually labeled tweets that were randomly selected from the base of already downloaded Icelandic tweets. Those 1142 tweets were collected in two steps where a total of 200 tweets were collected at first and then, about a month later, a total of 942 tweets were collected. All of these tweets were scored on the scale between minus 5, meaning very negative, and plus 5, meaning very positive. When the tweets are used as training data they most likely contain words that are common but do neither have a positive or negative bias. To remove those words, an Icelandic stop word list was needed and it had to be created since it is not available in the nltk package. That was done by translating the English nltk stop word list and by the end of that process a list of 316 Icelandic words and inflections was constructed. The results for the effect of the stop words removal can be seen in chapter 5.2.2. The manually labeled tweets mentioned above was also used to calculate accuracy in the bag of words approach and the results for that can be seen in chapter 5.2.

5 Results

5.1 The Icelandic Twitter network

This section describes some of the main characteristic properties of the Icelandic Twitter network and its Icelandic users.

5.1.1 Icelandic Filter

Following the process of collecting tweeters the filter identified a total of 17.462 tweeters as Icelandic. Those results were, however, not completely reliable. Among possible reasons involve the fact that a tweeter could easily claim living in Iceland or there are female non-Icelandic tweeters with the Icelandic ending -dottir (e.daughter) in their surname. Also, non-Icelanders living in Iceland could be included in the total, which, for the purpose of this project, they should not. This called upon a comprehensive manual work in order to remove those who were not Icelanders and to collect more Icelandic tweeters instead. In this context, relations between tweeters in the set of Icelanders already made were regarded. If, for example, tweeters that were identified as being Icelandic in the beginning had about 5000 followers but only two to five of them were Icelanders it would imply that they were falsely identified as Icelandic and they were therefore removed from the set. The same was considered in relation to collecting more Icelandic tweeter, i.e. if the proportion of Icelandic followers were high then the ones being followed were assumed to be Icelandic and therefore added to the set of Icelandic tweeters. The results of this manual work were a total of 2834 Icelandic tweeters being added and a total of 2175 tweeters being removed, which means that a total of 18.121 were, at the end, identified as Icelandic tweeters. There should, however, be noted that the manual work does not guarantee that there are not non-Icelanders still included in the total because this is impossible to know for sure. **Figure 4** shows detailed information of how the filter performed based on different scores.

Scores:	Numb:	Correct:	Wrong:	Correctness:
LDNT	19	19	0	100%
LDN	25	25	0	100%
LDT	433	433	0	100%
LNT	202	202	0	100%
DNT	7	7	0	100%
LD	338	329	9	97%
LN	1022	985	37	96%
LT	1668	1668	0	100%
DN	14	13	1	93%
DT	262	234	28	89%
NT	251	251	0	100%
L	7207	6782	425	94%
D	787	304	483	39%
N	2021	1879	142	93%
T	3206	2156	1050	67%
Sum	17462	15287	2175	88%
M	2834	2834	0	100%
SUM	**20296**	**18121**	**2175**	**89%**

Figure 4 - Filter Performance

As already explained in chapter 4.1 the following four scores were used to evaluate whether a tweeter was Icelandic or not; location score (L), description score (D), name score (N) and tweet score (T). LDNT means that all the four scores were positive, which means that the tweeters this applies to are most likely Icelandic. LDN means that all of the properties were positive except for the tweet score etc. As can be seen in **Figure 4**, reliability is highest when in relation with the name and location scores and also when two or more scores are positive. When the name score is the only one positive, it gives 93% reliability. What prevents this from being even higher is explained by referring to the Icelandic female surname ending, -dottir, which is also used in the Faroe Islands. Thus, Faroese tweeters are assumed to explain the difference. If the location score is the only one positive, there is 94% reliability. This could be explained by referring to non-Icelanders living in Iceland and to the number of spam bots claimed to be located in Iceland. The description score gives much lower reliability than the two already mentioned, or only 39%. The main reason involves some kind of a programming error that was discovered late in the process. In addition, there are words in the Icelandic word list that are a part of different languages as well. In this case, it particularly applied to the Swedish and the Faroese languages and therefore a large proportion of those tweeters who were wrongly identified as Icelanders were in fact Swedes and Faroese. The tweeter

score, with 67% reliability, can also be explained by referring to Faroese and Swedish tweeters, like in the case of the description score. The difference, though, is that Faroese on Twitter are considerable fewer than Swedish tweeters but the similarity, however, is that there are many words in the Icelandic language that are written in the same way in the Faroese language. The manual score (M), shows the total of tweeters that were added following the manual work already described.

5.1.2 General analysis

Figure 5 shows the growth rate for Icelandic tweeters for a period of almost four years. For each month the maximum value for the user identifier, as provided by the Twitter API, is indicated. As this data was recently collected it does not feature information about Icelandic tweeters who had disabled their Twitter accounts before the collection of data. Thus, the figure reflects only open Twitter accounts of Icelandic users, but does not consider tweeters that have registered as well as closed their accounts. What the figure does not show is that the growth rate for Icelandic tweeters from the launch of Twitter was very slow, with only few opening an account from July 2006 to January 2008. The growth rate continued to be relatively slow and constant until the beginning of 2009 where it suddenly grew tremendously, indicating that becoming a tweeter was starting to be a trend for Icelanders. After this period, the rate at which new users were joining the network deteriorates but what is interesting is another sudden growth of new Icelandic tweeters around March 2011. A possible reason for this might include that Icelandic football is played during the summer and the pre-season starts in March. Tweeting is a trend among football players in Iceland and since football is the most popular topic among Icelandic tweeters this reason is highly expectable. It is also interesting to see the high proportion of users that are inactive, referring to those who have never posted a tweet or have no followers.

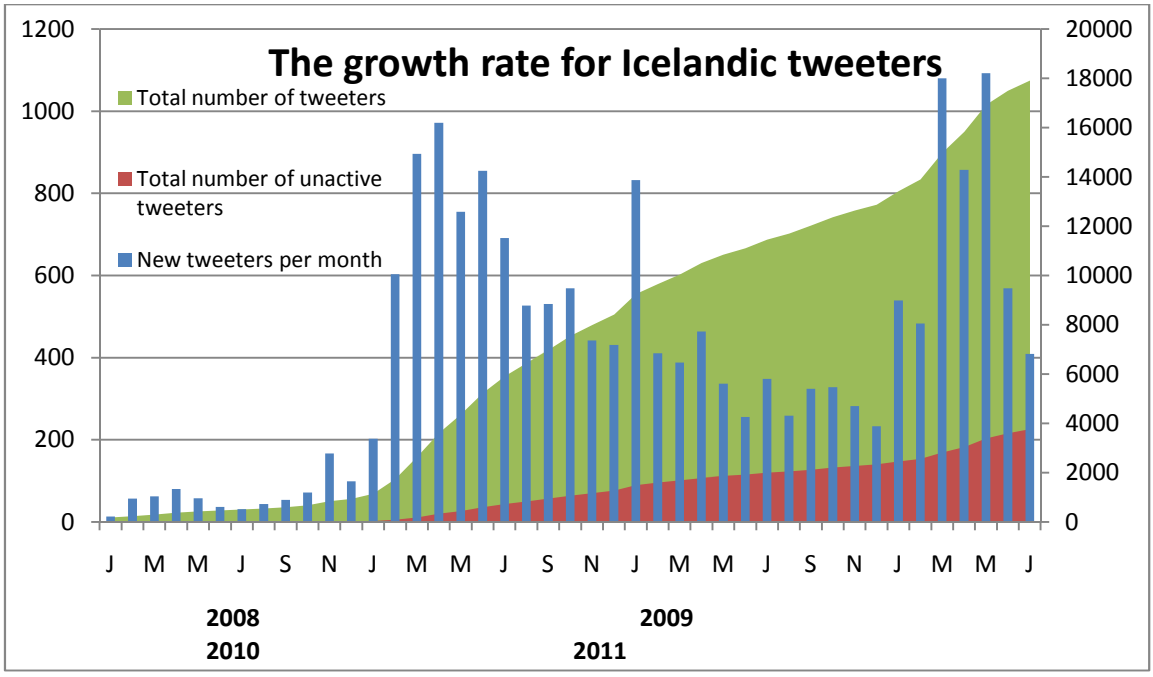

Figure 5 -The growth rate for Icelandic tweeters

In addition to looking at the growth rate for Icelandic tweeters, the growth rate for Icelandic tweets was considered as well. **Figure 6** reflects the maximum value for each month for the post identifier as provided by the Twitter API. What the figure does not show is that the first tweet in Icelandic was posted in October 2006, with a slow but steady continuing growth onwards. In relation to **Figure 5** this might be confusing since the first Icelandic tweeters opened accounts from July 2006. Even though they were very few in the beginning it might seem strange that they opened an account without posting a single tweet. However, this has reasons that can in a way be related to the fact that this data cannot be completely relied upon. This is because the data was collected not long ago and since the Twitter API only enables one to collect the 3200 newest tweets it is highly likely that this data is missing tweets from active users. For example, it would not take highly active tweeters a long time to get up to this limit of 3200 tweets. Thus, there are tweets that are not involved in **Figure 6** since the users posting those tweets have either closed their accounts before the collection of data or posted more than those 3200 tweets.

There is an obvious similarity between the growth rate for Icelandic tweet and the growth rate for Icelandic tweeters in relation to the sudden growth around March 2011. This does not, however, come as a surprise since it is considered to be based on the same grounds as

before, i.e. the beginning of the Icelandic football season. As can be seen it is popular to post tweets written in English but writing in Icelandic is also common. In addition to this there are tweets that are difficult to identify as either Icelandic or English. This could be because of very short tweets, consisting of few words or even emoticons alone. Another common factor that could affect this involves Icelandic tweets written by using the English alphabet and therefore the indigenous Icelandic letters cannot be used to detect that this is an Icelandic tweet. Last but not least, unidentified tweet could be those who reflect use of words that do not emerge in the list used to identify whether tweets are in Icelandic or not.

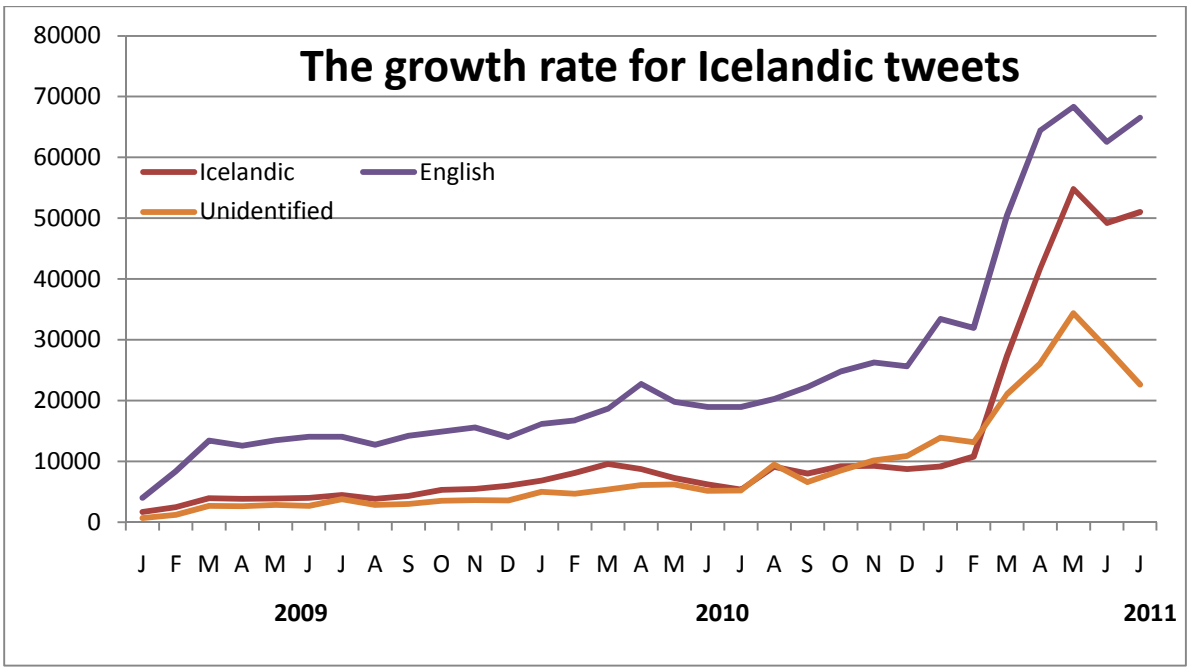

Figure 6 - The growth rate for Icelandic tweets

Now, when the Icelandic Twitter network has been discussed from the perspective of growth rate for both Icelandic tweeters and Icelandic tweets, it motivates for taking an even closer look at the network. The growth rate has shown that since the launch of Twitter it has become a trend among Icelanders to tweet and therefore it might be interesting to take a look at some of the network characteristics related to activity level. This could for example be done from the perspective of time. **Figures 7 and 8** reflect this by showing tweet activity of Icelandic tweeters by day of the week and hour of the day. Both of these pictures are based on data from the period from January till July 2011.

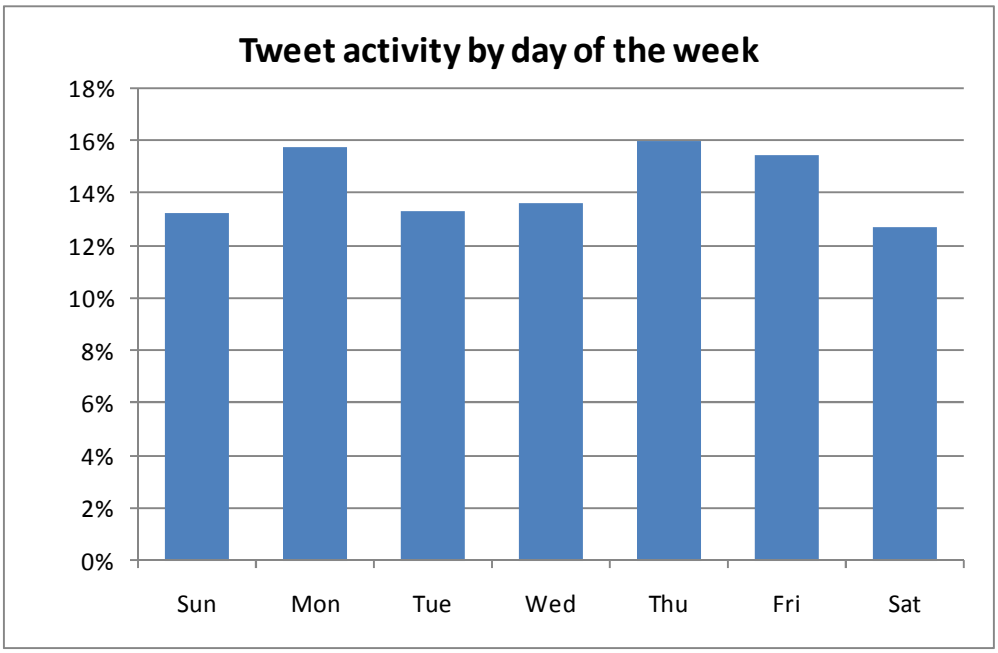

Figure 7 - Tweet activity by day

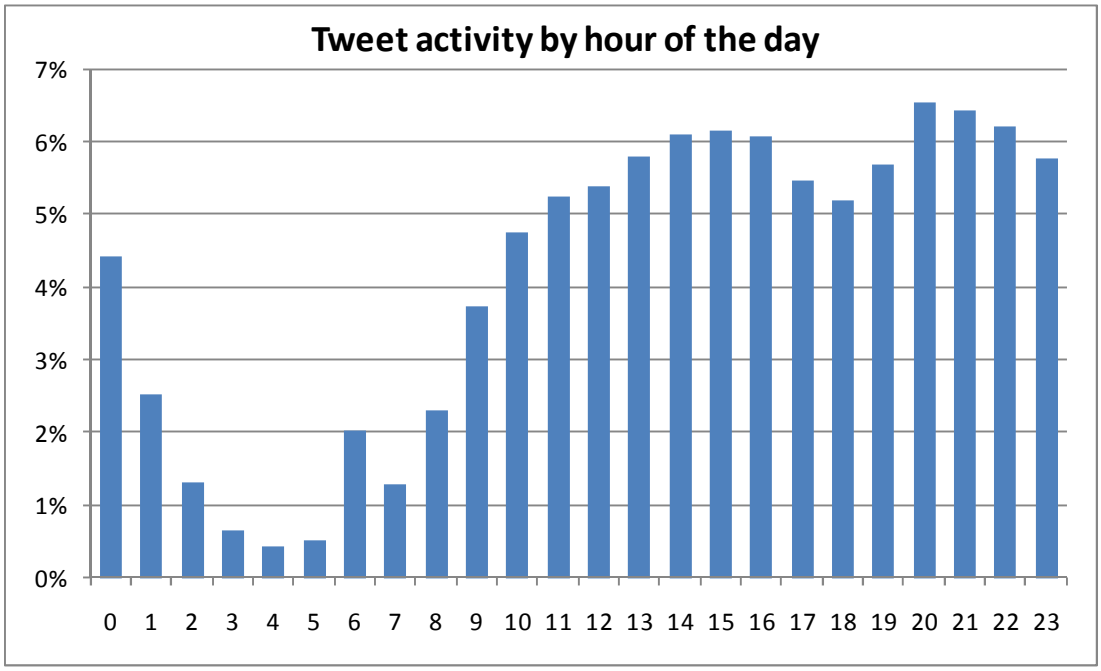

Figure 8 - Tweet activity by hour

In addition to viewing the Icelandic Twitter network distinctively a possible relationship between different variables was also studied. **Figure 9** shows the Spearman's correlation coefficient calculated for the following variables; followers count (FO), friends count

(FR), number of tweets (ST), page rank score (PR), klout score (KL), peer index (PI), Icelandic followers (IFO) and the ratio between Icelandic and international (II).

Spearman's correlation matrix								
	FO	FR	ST	PR	KL	PI	IFO	II
FO		0,695	0,667	0,768	0,605	0,447	0,781	-0,042
FR			0,559	0,470	0,480	0,345	0,049	-0,158
ST				0,476	0,589	0,302	0,453	-0,178
PR					0,482	0,365	0,935	0,380
KL						0,292	0,478	-0,005
PI							0,038	-0,016
IFO								0,047
II								

Figure 9 - Spearman's Correlation matrix

As can be seen there was a positive correlation between all the variables, except for the Icelandic/international index, or the rate between Icelandic and international followers. The strongest positive correlation was between Icelandic followers and page rank. This is understandable because the ranking score is dependent on the number of followers a tweeter has, i.e. the tweeter scores higher on the page rank list as the number of followers increases. Otherwise, the correlation is quite uniform between the variables but it is interesting to compare the results for the variables from the two different external scoring services, klout score (KL) and peer index (PI). Correlation between them and the other variables are very different, which implies that they seem to focus on different factors when identifying how influential a tweeter is. **Figure 10** can be connected with this since it provides a view of the relation between number of tweets and number of followers, which has been used as an indication of influence as presented in chapter 2.1. On the right-hand side of the figure is the activity level divided into five different classes, representing the number of tweets per tweeter for the period from January till July 2011.

As can be seen, the more often a tweeter posts a tweet, i.e. the higher the activity level, the more follower one has. This is in coherence with the Spearman's correlation coefficient of 0,667 for followers count (FO) and number of tweets (ST) shown in **Figure 9**.

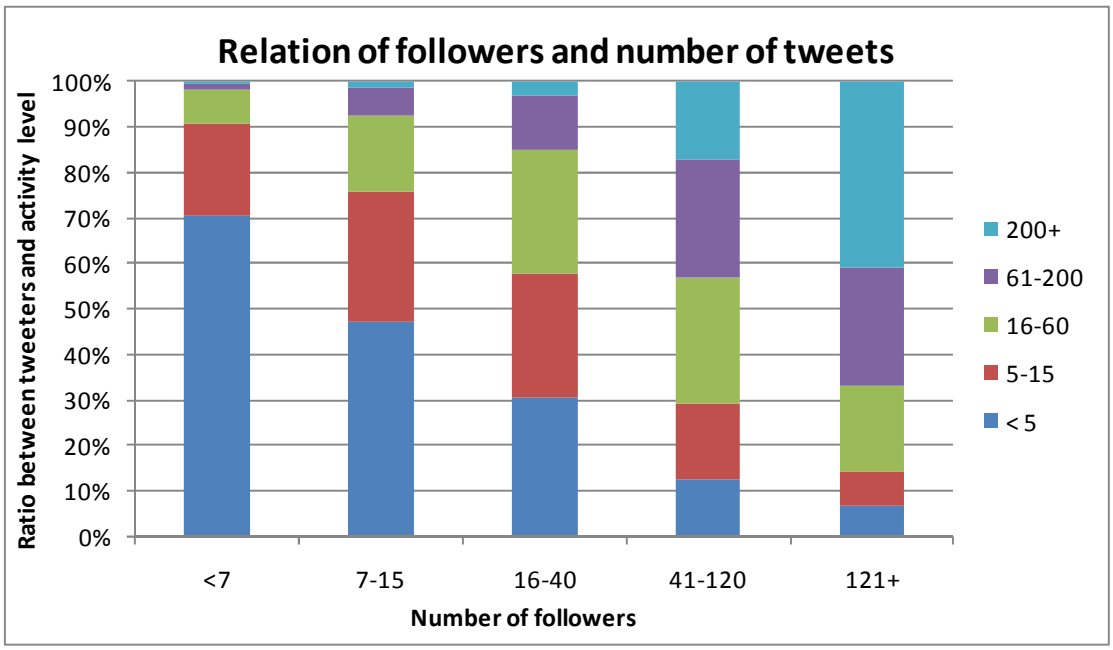

Figure 10 - Relation of followers and tweets

The average Icelandic tweeter follows 75 other tweeters, thereof 23 Icelandic, but is followed by 76 tweeters, thereof 23 Icelandic. What is rather unique about the Icelandic Twitter network, in relation to the relationship between tweeters and followers, is that in a network where only few tweeters have more than 10.000 followers there is one that has over 100.000 followers and another that has over 200.000 followers. These two outlaws play a considerable large role in the creation of the characteristics of the Icelandic Twitter network. A brief study on cluster formulation of the network showed three major clusters; the first included artists, particularly musicians, the second included athletes and sportswriters and the third consisted of large groups such as companies and news media. What may be surprising is that politicians, except for one, are not among active Icelandic tweeters. This is in contrast with for example politicians in the US.

5.2 Sentiment analysis

As already introduced, this project is based on two basic methods for sentiment analysis; bag-of-words approach and machine learning approach, where the latter focuses on the use of the Naïve Bayes classifier included in the nltk Python Package. Both of these methods were tested on the manually labeled tweets that fell into 220 positive, 702 neutral and 220 negative, 1142 in total. **Figure 11** reveals, in more detail, the distribution of those manually labeled tweets.

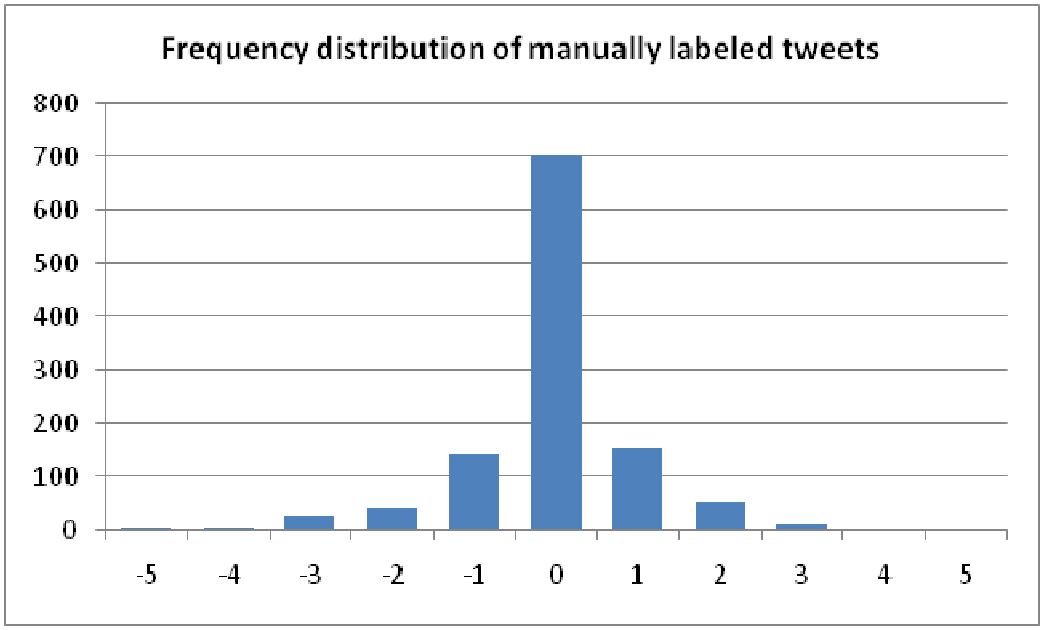

Figure 11 - Distribution of manually labeled tweets

5.2.1 Bag-of-words approach

The bag-of-words approach for identifying sentimental strength involves the use of both the files and the methods described in chapter 4.2. Manually labeled tweets were run through the analyzer and different scores were saved based on the usage of different combinations of files. The objective was to find out if there was a difference between those scores. This can be seen by observing the eight different correlation matrices in **Figure 12**. For a brief description, the vertical axis shows how the tweets were manually labeled as positive, neutral and negative and the horizontal axis shows how the algorithm classified the tweets. The accuracy of every outcome for every combination is also shown

and this reflects the cases where the algorithm classifies the tweets into the same classes as they were when manually labeled. Additionally, the accuracy for a plus/minus one and plus/minus two classes was calculated. The former class involves the cases where the algorithm wrongly categorizes the tweets but without doing so by identify a positive tweet as a negative tweet and vice versa. Instead this would for example mean manually labeled tweets as positive but categorized by the algorithm as neutral. The latter class, however, reflects the cases when the algorithm classifies tweets in total opposite of what they were manually labeled as. Thus, the algorithm would classify a manually labeled positive tweet as a negative one and a manually labeled negative tweet as a positive one. Further division on the eight matrices can be seen in the figure, but what should be pointed out is that there is a difference between the four matrices on the left-hand side and the four on the right-hand side. Those on the right-hand side all include the emoticon score, while the other four on the left-hand side do not.

Emotion list

	pos	neu	neg	sum
pos	156	49	15	220
neu	158	445	99	702
neg	32	66	122	220
sum	346	560	236	1142

accuracy	= class:	723	0,633
	±1 class:	372	0,326
	±2 class:	47	0,041

Emotion and emoticon lists

	pos	neu	neg	sum
pos	159	46	15	220
neu	173	430	99	702
neg	34	65	121	220
sum	366	541	235	1142

accuracy	= class:	710	0,622
	±1 class:	383	0,335
	±2 class:	49	0,043

Emotion and negation lists

	pos	neu	neg	sum
pos	156	50	14	220
neu	163	443	96	702
neg	33	65	122	220
sum	352	558	232	1142

accuracy	= class:	721	0,631
	±1 class:	374	0,327
	±2 class:	47	0,041

Emotion, negation and emoticon lists

	pos	neu	neg	sum
pos	159	47	14	220
neu	178	428	96	702
neg	36	63	121	220
sum	373	538	231	1142

accuracy	= class:	708	0,620
	±1 class:	384	0,336
	±2 class:	50	0,044

Emotion and booster lists

	pos	neu	neg	sum
pos	156	48	16	220
neu	157	447	98	702
neg	32	68	120	220
sum	345	563	234	1142

accuracy	= class:	723	0,633
	±1 class:	371	0,325
	±2 class:	48	0,042

Emotion, booster and emoticon lists

	pos	neu	neg	sum
pos	159	45	16	220
neu	172	432	98	702
neg	34	67	119	220
sum	365	544	233	1142

accuracy	= class:	710	0,622
	±1 class:	382	0,335
	±2 class:	50	0,044

Emotion, negation and booster lists

	pos	neu	neg	sum
pos	156	48	16	220
neu	160	446	96	702
neg	34	68	118	220
sum	350	562	230	1142

accuracy	= class:	720	0,630
	±1 class:	372	0,326
	±2 class:	50	0,044

Emotion, negation, booster and emoticon lists

	pos	neu	neg	sum
pos	159	45	16	220
neu	175	431	96	702
neg	37	66	117	220
sum	371	542	229	1142

accuracy	= class:	707	0,619
	±1 class:	382	0,335
	±2 class:	53	0,046

Figure 12 - Correlation matrics, 3 classes

The results shown by the eight matrices are somewhat surprising, especially considering how similar the outcomes for the use of different combinations of files are. Thus, there is a very little difference between the scores. The only exception is related to emoticons where the lower accuracy is more visible than in the other outcomes of different combinations. As reflected by the matrices on the right-hand side, the algorithm has a tendency to give a higher positive score for when emoticons occur than for when they are not included in the tweets. This has a decreasing effect on the accuracy since too many tweets are scored as positive and too few are scored as negative. There is, however, nothing abnormal about this considering that about five percent of the 1142 manually labeled tweets included emoticons, thereof 48 positive. But this makes it difficult to consider emoticons as a reliable factor to identify sentimental strength.

Booster and negation words also have a negative effect, i.e. both decreases the accuracy, but the total effect is very limited, unless when emoticons are added to the emotional, booster and negation lists. Then the accuracy decreases by one to two percent. Considering the booster words, it was envisaged to some degree that those would not have strong effect. The main reason is the fact that the tweets are short and for example if tweets only contain one emotion word it does not have any effect on the accuracy when classifying tweets as positive, neutral or negative. Booster words occurred in 60 tweets of the total number of 1142 manually labeled tweets but only had an effect on the accuracy in 20 occurrences of those 60. On the other hand, it was expected in the beginning that the usage of negation list in combination with the emotion list would increase the accuracy from that of only using the latter. Of the total of 1142 tweets, 221 contained a negation word. Nonetheless, this does not seem to have much effect on the accuracy. By taking a closer look at possible reasons a part of the results explained this by revealing that even though 221 tweets included a negation word it was only 37 times that the negation word stood next to the emotion word and therefore it only had an effect those number of times.

What gives the highest accuracy of all the possible combinations of the different files is when the emotion list alone is used, i.e. without combining it with the booster, negation or the emoticon lists. Of the total number of the 1142 manually labeled tweets 601 of them contained emotion words, with the division of 384 tweets having one emotion word, 132 tweets having two emotion words and 85 tweets having three or more emotion words.

Figure 13 shows an example of how an Icelandic tweet is classified by the algorithm based on the usage of all of the above mentioned combinations of different files. Every score includes the usage of the emotion list and then various combinations of negation list (N), booster list (B) and emoticon list (E) are added to it. This mixing of lists gives the total number of eight lines shown in the figure, which correspond to the eight correlation matrices presented above, in **Figure 12**. The sentence includes three emotion words, one booster word, one negation word and an emoticon and it is scored as a positive one in all instances, no matter what kind of a combination of different files is used.

Tweet:	Ágæt	mynd,	***ekki***	eins	góð	og	þættirnir.	En	ég	bíð	*mjög*	spenntur	eftir	7.	seríu	;)	Rounded
Translation:	Fine	movie,	***not***	as	good	as	the shows.	But	I'm	waiting	*very*	excited	for	7th	season	;)	Score:
	2	0	0	0	2	0	0	0	0	0	0	3	0	0	0	0	2
N	2	0	**N**	0	**-2**	0	0	0	0	0	0	3	0	0	0	0	1
B	2	0	0	0	2	0	0	0	0	0	**B**	**6**	0	0	0	0	3
NB	2	0	**N**	0	**-2**	0	0	0	0	0	**B**	**6**	0	0	0	0	2
E	2	0	0	0	2	0	0	0	0	0	0	3	0	0	0	**E**	3
NE	2	0	**N**	0	**-2**	0	0	0	0	0	0	3	0	0	0	**E**	2
BE	2	0	0	0	2	0	0	0	0	0	**B**	**6**	0	0	0	**E**	4
NBE	2	0	**N**	0	**-2**	0	0	0	0	0	**B**	**6**	0	0	0	**E**	3
Naive Bayes	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	POS

Figure 13 - Icelandic sentence

In addition to discussing the accuracy only by looking at the classification into positive, neutral and negative scores, as was done above, there is another approach that involves identifying how close to the manually labeled scores the algorithm classifies the tweets. This approach is reflected in the eight matrices shown in **Figure 14** below and is based on the same data as those shown in **Figure 12**. The difference though is that below is a focus on 11 different classes, in which the scores were classified, instead of only the following three classes, positive, neutral and negative.

The manually labeled classes were compared to the results from the algorithm with the aim of trying to identify if emoticons, booster words and negation words scores move any closer to the classes as they were scored when manually labeled in the beginning. The results are similar to those shown in **Figure 12**. For example, when emoticons are used it decreases the accuracy. The usage of the booster and negation lists also gives similar results, except for the instances where the emotion list is used in combination with the negation list. Then there is a slight increase in accuracy from only using the emotion list.

Consequently, negation would score better, or give higher accuracy, if the focus would be on this view instead of the one provided above.

Emotion list		
	frequency:	accuracy:
= class	526	0,461
±1 class	862	0,755
±2 classes	1056	0,925
±3 classes	1120	0,981
	1142	

Emotion and emoticon lists		
	frequency:	accuracy:
= class	509	0,446
±1 class	853	0,747
±2 classes	1050	0,919
±3 classes	1117	0,978
	1142	

Emotion and negation lists		
	frequency:	accuracy:
= class	526	0,461
±1 class	867	0,759
±2 classes	1058	0,926
±3 classes	1119	0,980
	1142	

Emotion, negation and emoticon lists		
	frequency:	accuracy:
= class	509	0,446
±1 class	858	0,751
±2 classes	1051	0,920
±3 classes	1116	0,977
	1142	

Emotion and booster lists		
	frequency:	accuracy:
= class	525	0,460
±1 class	853	0,747
±2 classes	1047	0,917
±3 classes	1116	0,977
	1142	

Emotion, booster and emoticon lists		
	frequency:	accuracy:
= class	508	0,445
±1 class	844	0,739
±2 classes	1041	0,912
±3 classes	1113	0,975
	1142	

Emotion, negation and booster lists		
	frequency:	accuracy:
= class	526	0,461
±1 class	859	0,752
±2 classes	1050	0,919
±3 classes	1113	0,975
	1142	

Emotion, negation, booster and emoticon lists		
	frequency:	accuracy:
= class	509	0,446
±1 class	850	0,744
±2 classes	1043	0,913
±3 classes	1110	0,972
	1142	

Figure 14 - Correlation matric, 11 classes

The matrices above are based on classification using the emotion file created for this project and by testing four files, containing different numbers of emotion words, on manually labeled data it was possible to see how the increasing amount of emotion words affected the manually labeled tweets. Here, this is presented in the learning curve shown in **Figure 15**. For explanation, the green curve reflects the proportion of manually labeled positive tweets as negative and manually labeled negative tweets as positive. Tweets that

are scored as neutral are not included here. The blue curve also takes into account only positive and negative scores, i.e. it does not include the neutral ones, but it reflects the proportion of correctly scored tweets into positive and negative instead of wrongly scored. The orange curve is for the overall accuracy. In the beginning, 61,5% of the manually labeled data were classified as neutral and when there are no words in the emotion word list they continue to be classified as neutral. This is the reason for the curve to start as a reflection of the proportion 702:1142. The curve ends in 63% accuracy which means that the result is almost a straight line. The last curve was created to show that if there would have been equal number of positive, neutral and negative scores in the beginning, the curve would have begun with 33% accuracy.

A reason for the shape of the curves can be related to the data consisting of 61% neutral scores, 19% positive scores and 19% negative scores. Another reason that the curves do not have the shape of a typical learning curve is because the Icelandic words used in the emotion word list are received by translating an English list of words in an alphabetical order. Thus, the most common words did not appear first, which they would have if the word list would have been created from scratch. However, during the translation, common Icelandic slang was added to the list of words, which resulted in a slightly curved line instead of ending as a completely straight line.

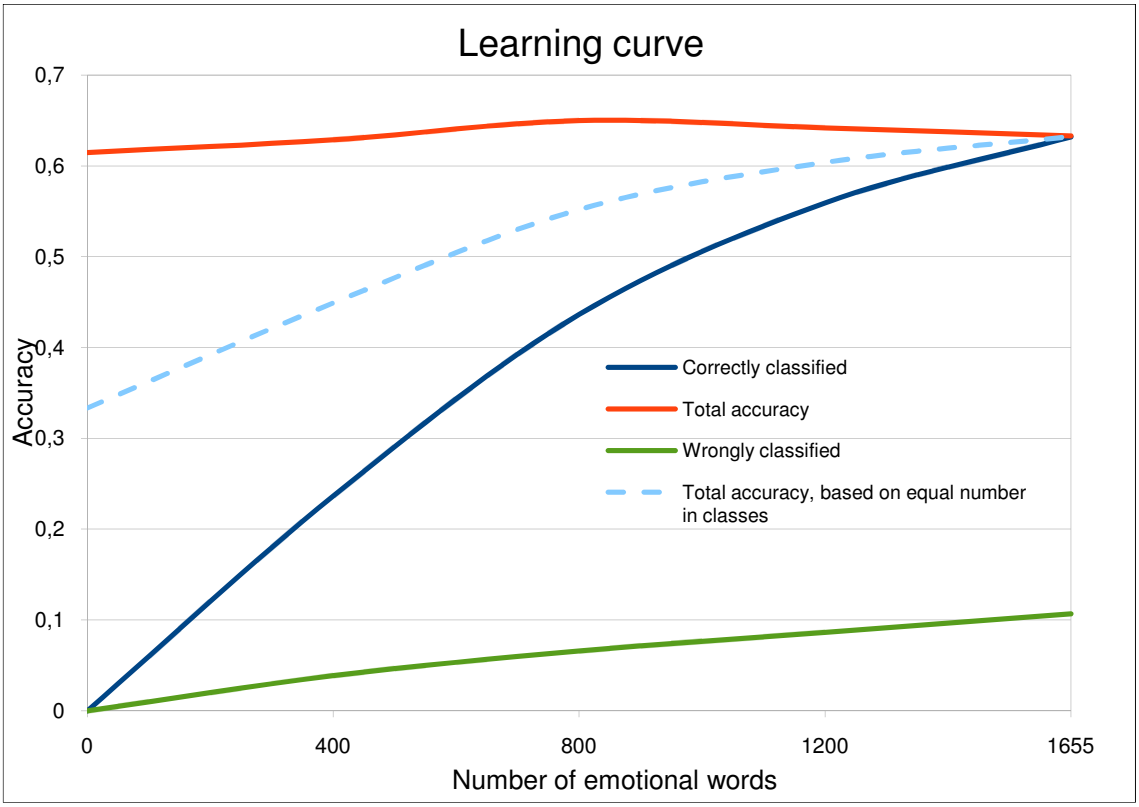

Figure 15 - Learning curve

5.2.2 Machine learning approach

In addition to the bag-of-words approach, a machine learning approach was used to analyze the tweets, as already discussed. Based on a comparison of how those two approaches perform, a decision on which one to use for prior classifier is then made. Under the machine learning approach the Naïve Bayes algorithm from the Python package nltk was used. All of the 1142 tweets were used as training and test data where the ¾ were used as training data and ¼ as test data. The classifier was trained by clearing the data in a number of different ways in order to identify whether or not any particular clearing would increase the accuracy more than other. Before moving on to the results this will be described briefly. Before the classifier was trained and tested, the data was cleared by removing symbols (s), digits (D), words that have only one letter (W) and words that occurred in the stop word list (S). Every possible combinations of these clearing approaches was made, which results in a total of 16 different combinations. This was done for both bigram and unigram, so the ending result was a total of 32 different combinations, which each was constructed 30 times in order to get more reliable average. In **Figure 16** the average accuracy and standard deviation for each version is showed.

Manually labeled tweets				
	Unigram		Bigram	
Clearing:	avg acc:	std:	avg acc:	std:
SDWs	0,4802	0,0245	0,4754	0,0235
SDs	0,4759	0,0239	0,4726	0,0240
SWs	0,4802	0,0245	0,4754	0,0235
DWs	0,4301	0,0303	0,4153	0,0328
Ss	0,4759	0,0239	0,4726	0,0240
Ds	0,4285	0,0297	0,4040	0,0329
Ws	0,4301	0,0303	0,4153	0,0328
s	0,4285	0,0297	0,4040	0,0329
SDW	0,4852	0,0281	0,4809	0,0285
SD	**0,4864**	0,0287	**0,4813**	0,0293
SW	0,4822	0,0287	0,4782	0,0291
DW	0,4372	0,0363	0,4236	0,0305
S	0,4836	0,0281	0,4790	0,0286
D	0,4404	0,0354	0,4169	0,0313
W	0,4355	0,0371	0,4224	0,0317
	0,4386	0,0344	0,4163	0,0313

Figure 16 - Naive Bayes results

When results for bigrams and unigrams, shown in the table, are compared, it is obvious that the former always gives lower accuracy than the latter. The difference is on average between two and three percent. It is also notable that by not clearing the words in the stop word list it results in a much lower accuracy than if they are cleared. Similarly, it is interesting to see that the standard deviation, and thereupon the distribution, is always higher when the stop words are not cleared from the data. Other combinations give similar results, as can be seen. The highest accuracy is when unigrams are used and when digits and stop words are cleared from the data, but since the overall accuracy is quite low it would be interesting to see if it would have an increasing effect on the accuracy to have a lot more manually labeled tweets.

To find out if there is a possibility to increase the accuracy without analyzing more tweets, the emotional words from the emotional file were added to the classifier as training data along with the labeled tweets. After completing this, the training set consisted of 75% of all the 1142 number of tweets in addition to all the emotional words, which gave about 2500 training features in total. The disadvantage of combining the

tweets and the words consists in mixing two different statistical distributions together. **Figure 17** is comparable to **Figure 16**, except that the training data is different, as already described.

Manually labeled tweets and emotion words				
	Unigram		Bigram	
Clearing:	avg acc:	std:	avg acc:	std:
SDWs	0,5888	0,0232	0,5873	0,0237
SDs	0,5896	0,0235	0,5879	0,0237
SWs	0,5888	0,0232	0,5873	0,0237
DWs	0,6106	0,0275	0,6141	0,0271
Ss	0,5896	0,0235	0,5879	0,0237
Ds	0,6148	0,0265	**0,6178**	0,0263
Ws	0,6106	0,0275	0,6141	0,0271
s	**0,6148**	0,0265	**0,6178**	0,0263
SDW	0,6018	0,0235	0,5996	0,0236
SD	0,6071	0,0223	0,6056	0,0230
SW	0,6013	0,0233	0,5996	0,0228
DW	0,6064	0,0269	0,6089	0,0256
S	0,6074	0,0221	0,6065	0,0223
D	0,6118	0,0247	0,6148	0,0238
W	0,6058	0,0271	0,6083	0,0257
	0,6118	0,0247	0,6149	0,0238

Figure 17- Naive Bayes results including emotion words

The results are somewhat surprising, especially considering that the accuracy increased quite a lot by adding the emotional words to the classifier. By using the emotional words both the precision and recall decreases, even though the accuracy increases. It is interesting to see that the accuracy is the most when symbols and digits are cleared from the data and when using bigram, but not when using unigram as was the case in the former table.

6 Discussion and conclusions

The Icelandic Twitter network is a distinct network that cannot be considered very large compared to many other national Twitter networks around the world. However, it was very interesting to work with a data of this scope and nature. By focusing on a specific country or a specific nationality particular characterizing features follow, language being a good example. The Icelandic language played a major role in the sentimental analysis part of this project but besides that the general characteristics of the Icelandic Twitter network was studied. Similar to the prior work of Krishnamurthy, Gill and Arlitt (2008) factors such as the growth of the network and activity level were presented. The relationship between different variables was also discussed.

Two different approaches were used in order to detect sentiment in Icelandic tweets, the bag-of-word approach and the machine learning approach, which both were based on tweets classified into positive, neutral and negative classes. The main conclusion of this study is that the former approach performed better than the latter. By using the machine learning approach it gave an accuracy of 48,6% while by using the bag-of-words approach the accuracy was 63,3%. This performance involves using only emotion words. Thus, the best performance resulted in not using negation words, booster words or emoticons, which gave an accuracy of 63,3%. This is higher compared to Thelwall et al. (2009) who got an accuracy of 60,6% with the use of bag-of-words approach. However, it should be kept in mind that they use a different scale and they also label the tweets in a different way than done in this project. They also use the SentiStrength algorithm which is developed for English language and therefore it is difficult to compare those two results.

There were certain challenges related to the detection of polarity in the Icelandic Twitter statuses. For the first it was difficult to detect sentiment in such short texts as characterizes tweets, with its length limit of 140 characters. This was because many tweets did not include a single emotional word, which made it difficult to detect for emotion. It was also challenging because Icelandic tweets were often written with the use

of the English alphabet and therefore it was not possible to use the Icelandic language to detect the sentiment. This can be categorized as wrongly written Icelandic, when for example English letters are used instead of indigenous Icelandic letters. Example of this is when the word football is written like this *fotbolti* instead of fótbolti. Last, but not least, it is considered a disadvantage that the same person both labeled the tweets and the emotion word list, which effects the correctness of the classification. This is in context with the discussion of Thelwall et al. (2009) about different challenges related to identifying positive and negative sentiment detection in informal text language. Those included for example language creativity, various interpretations between human coders and expressions of sentiment without emotion-bearing words.

References

Bautin, M., Vijayarenu, L. and Skiena, S. (2008). International Sentiment Analysis for News and Blogs. In E. Adar, M. Hurst, T. Finin, N.S. Glance, N. Nicolov and B.L. Tseng (eds.) *Proceedings of the 2nd International Conference on Weblogs and Social Media (ICWSM), 30 March-2 April 2008* (p. 19-27). Seattle: AAAI Press.

Bollen, J., Mao, H. and Pepe, A. (2009). Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, 17-21 July 2011*. Barcelona, Spain: Association for the Advancement of Artificial Intelligence.

Boyd, D., Golder, S. and Lotan, G. (2010). Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter. *Proceedings of the 43rd Hawaii International Conference on System Sciences, 5 August 2010*. Hawaii: IEEE Computer Society.

Boiy, E., Hens, P., Deschacht, K. and Moens M.F. (2007). Automatic Sentiment Analysis in On-line Text. In L. Chan and B. Martens (eds.) *Proceedings of the 11th International Conference on Electronic Publishing, 13-15 June 2007* (p. 349-360). Vienna, Austria: ÖKK-Editions.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. *Computer Network and ISDN Systems, 30* (1-7), 107-117.

Cheong, M. and Lee, V. (2009). Integrating Web-based Intelligence Retrieval and Decision-making from the Twitter Trends Knowledge Base. *Proceeding of the 2nd ACM workshop on Social web search and mining, 2 November 2009*. Hong Kong, China: ACM.

Hatzivassiloglou, V. and McKeown, K. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of 35th Meeting of the Association for Computational Linguistics, July 1997* (p. 174-181). Madrid, Spain: ACL.

Heerschop, B., van Iterson, P., Hogenboom, A., Frasincar, F. and Kaymak, U. (2011). Accounting for Negation in Sentiment Analysis. *Proceedings of the 11th Dutch-Belgian Information Retrieval Workshop (DIR 2011), 4 August 2011* (p. 38-40). Amsterdam: DIR.

Java, A., Song, X., Finin, T. and Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, 12 August 2007* (p. 56-65). New York: ACM Press.

Krishnamurthy, B., Gill, P. and Arlitt, M. (2008). A few chirps about Twitter. In *Proceedings of the first workshop on Online social networks, 18 August 2008* (p. 19-24). New York: ACM Press.

Kwak, H., Lee, C., Park, H. and Moon, S. (2010). What is Twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web, 26-30 April 2010* (p. 591-600). New York: ACM Press.

Manning, C.D. and Schuetze, H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: MIT Press.

Na, J.C., Sui, H., Khoo, C., Chan, S. and Zhou, Y. (2004). Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *Advances in Knowledge Organization, 9*, 49-54.

Nielsen, F.Å. (2011). *AFINN-111*. Retrieved from http://www2.imm.dtu.dk/pubdb/views/publication_details.php?id=6010

Pang, B. and Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval, 1* (1-2), 1-135.

Pang, B., Lee, L. and Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, 6-7 July 2002* (p. 79-86). Morristown, NJ: Association for Computational Linguistics.

Saurí, R. (2008). *A Factuality Profiler for Eventualities in Text*. Ph.D. Thesis: Brandeis University. Retrieved from http://pages.cs.brandeis.edu/~roser/pubs/sauriDiss_1.5.pdf

Thelwall, M., Buckley, K. and Paltoglou, G. (2011). Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology, 62* (2), 406-418.

Thelwall, M., Buckley, K., Paltoglou, G. and Cai, D. (2009). Sentiment Strength Detection in Short Informal Text. *Journal of the American Society for Information Science and Technology, 61* (12), 2544-2558).

Tong, R.M. (2001). An operational system for detecting and tracking opinions in on-line discussions. *Working Notes of the ACM SIGIR 2001 Workshop on Operational Text* Classification (p. 1-6). New York: ACM.

Weng, J., Lim, E.P., Jiang, J. and He, Q. (2010). TwitterRank: Finding Topic-sensitive Influential Twitterers. In *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining (WSDM 2010), 4-6 February 2010* (p. 261-270). New York: ACM.

Wilson, T., Wiebe, J. and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 6-8 October 2005* (p. 354). Vancouver, CA: Association for Computational Linguistics.