Using Tweets for Assigning Sentiments to Regions

Erik Tjong Kim Sang
Meertens Institute
Amsterdam, The Netherlands
erik.tjong.kim.sang@meertens.knaw.nl

Abstract

We derive a sentiment lexicon for Dutch tweets and apply the lexicon for classifying Dutch tweets as positive, negative or neutral. The classifier enables us to test what regions in the Netherlands and Flanders express more positive sentiment on Twitter than others. The results reveal sentiment differences between Flemish and Dutch provinces, and expose municipalities which are a lot more negative than their neighborhood. The results of this study can be used for finding areas with local issues that might be expressed in tweets.

Keywords: Sentiment analysis, Twitter, Dutch

1. Introduction

Measuring sentiment of social media messages is an important application for organizations and individuals that want to track the impact of products, services, events and people on social media. However, the sheer volume of the data stream makes manual impossible for all but small selections of the data. An automatic analysis is difficult because of the ambiguity of language but there is no alternative when large volumes of data need to be processed.

In this paper we describe a two-stage process for identifying positive and negative Dutch tweets. First we create a Dutch sentiment lexicon based on the vocabulary observed in Dutch tweets. Next we use the lexicon for determining if Dutch tweets are positive, negative or neutral.

After this introduction and an overview of the related work, we present our method for deriving a sentiment lexicon from tweets. Next we apply the lexicon in a case study: a comparison of the average sentiment of different regions in The Netherlands and Flanders. The final section of the paper contains some concluding remarks.

2. Related work

The earliest references of work on automatic sentiment analysis are from 2002, for example the work of Pang et al. (2002), who use different machine learning techniques for determining if movie reviews are positive or negative. Application of sentiment analysis to tweets started seven years later, with among others the report of Go et al. (2009) who created a training corpus of 1.6 million positive and negative tweets by using emoticons as noisy labels. This approach has been used by several follow-up works, for example Pak and Paroubek (2010) . Sentiment analysis applied to Dutch tweets was only reported on in Tjong Kim Sang and Bos (2012), who performed a manual sentiment analysis of political Dutch tweets. In 2012, the company Incentro seemed to have developed a sentiment analysis module for Dutch (Incentro, 2012) but its current status is unknown. Sentiment analysis of tweets per region was first covered by Mislove et al. (2010) who studied the average mood of regions of the United States in the course of two days.

3. Sentiment lexicon

We use a lexicon of sentiment words for identifying positive and negative tweets. There are two reasons for favoring this approach over a machine learning approach with a training corpus of positive and negative examples. The first reason is portability: while we can share share a sentiment lexicon created from tweets with the research community, we would not be able to share annotated tweets because of the developer rules of the company Twitter (Twitter, 2011)[1]. The second reason is ease of implementation: our sentiment analysis is part of a parallel tweet search engine implemented on the Hadoop framework (White, 2012). Creating a lexicon-based analyzer required fewer resources than implementing a machine learner on Hadoop.

In order to collect words for the sentiment lexicon, we collected three sets of Dutch tweets, one with tweets that contained smileys – :-) or :) – one with tweets that contained frownies – :-(or :(– and one which did not contain any of the four emoticons. Our assumption is that when we compare the sets, words that express positive sentiment would predominantly be found in the first dataset while words associated with a negative sentiment would be found primarily in the second set. The third set will be used as an approximation of neutral tweets. We used the website twiqs.nl (Tjong Kim Sang and van den Bosch, 2013) for building the two sentiment tweet sets from the available tweets of January 2013 (1,724,642 and 999,685 tweets respectively). The third set was generated from the tweets of 16 January 2013 (2,569,203 tweets).

The search process produced frequency scores for 1,642,659 strings (words, names, punctuation signs and hash tags) from the sentiment tweets. We compared the frequencies of the two datasets with the t-test: $(f_1 - f_2)/\sqrt{f_1 - f_2}$ (Church et al., 1991, page 8), a measure for comparing the usage of a word in different texts. We created two ranked lists of words (strings without punctuation): one of the positive words versus the negative and neutral words and one of the negative words versus the positive and neutral words. We use add 0.5 smoothing to deal

[1] Twitter allows sharing the identification codes of tweets which can be used for retrieving the tweet text from their website. However the retrieval process will fail when tweets have been deleted by Twitter or by the author of the tweet.

Name	Lexicon size	Accuracy
baseline: 4 emoticons	4	87.6%
n-best t-scores, threshold 0	75,000	48.2%
n-best t-scores, threshold 5	4,400	53.2%
n-best t-scores, threshold 10	2,700	53.6%
incremental selection	644	84.2%
incr. sel. with 4 emoticons	100	93.2%
manual selection	338	82.1%

Table 1: Performance of the sentiment lexicon extracted from January 2013 tweets when tested on automatically annotated tweets from July 2013. For the n-best experiments, only the best results per threshold value are shown. The best results have been achieved with incremental selection of words suggested by the t-test combined with the four emoticons of the baseline: :-) :) :-(:(

with zero frequencies.

The t-test does not provide a perfect sentiment ranking: the top of the lists contained some character sequences that accidentally occur in a few positive or negative tweets. Therefore we experimented with frequency thresholds (0, 5 and 10) and removed words from the lexicon that occurred fewer times in either of the sentiment collections. We also tested building the lexicon incrementally, by only adding strings suggested by the t-test to the lexicon if they improved the sentiment performance of the lexicon on the test data.

Next, we devised a method for assigning sentiments to tweets based on the lexicon words. Tweets that do not contain any of the words will be neutral and tweets with words from only one sentiment set will be assigned that particular sentiment. In case a tweet contains both positive and negative words, the majority sentiment can be assigned. In case of a tie, the sentiment of the final sentiment word can be given preference, so that some cases of irony can be handled, like in the tweet: *so happy with with math grade :(*. Sentiment words immediately preceded by any of the words *not* (*niet*) and *no* (*geen*) are interpreted with their opposite sentiment value.

As test data we used a random selection of tweets from July 2013. We manually annotated 500 tweets with at least one of the two smileys, 500 tweets with at least one of the two frownies and 1000 tweets without any of the four emoticons. We selected the first 600 positive, negative and neutral tweets of this set as test set (a total of 1800). We tested different lexicons and measured their accuracy on classifying the tweets in test data with respect to the three sentiment classes positive, negative and neutral. A summary of these experiments can be found in Table 1.

Because of the method we used for selecting the test data, the baseline lexicon with only four emoticons already performed very well (accuracy 87.6%). Using the n words with the best t-scores did not perform as well (best accuracy 53.6%). Restricting the lexicon words to words which appeared at least 5 or 10 times in both positive and negative tweets, was a good idea (best accuracy 53.6% vs 48%). Adding only words to the lexicon which improved their performance on the test data worked very well, both without emoticons (accuracy 84.2%) and with emoticons in the lexicon (best accuracy 93.2%). We also evaluated a manually created sentiment lexicon and found that its performance was between the n-best approaches and the incremental selection methods.

In our experiments, incrementally adding words suggested by the t-test worked best. Words are only added to the lexicon if they improve the performance on the test set. However, this approach amounts to tuning the lexicon to the test data which may lead to performances which cannot be reproduced for other datasets. An inspection of the words in the two lexicons showed that several of the included words did not express sentiment in isolation but they were only added because they appeared in a positive or negative tweet in the test data. For this reason we did not select the lexicons generated with this method but we continued with the manually selected lexicon. This lexicon also has the advantage that it finds more sentiment tweets than the baseline lexicon and the incrementally built lexicons.

4. Measuring sentiment per region

As an application, we measured the average sentiment of regions in The Netherlands and Flanders. The results of this measurement could complement the research on living conditions periodically performed by the Dutch and Belgian government and press. For this purpose we selected the tweets with geolocation information from the period 1 to 31 January 2014 and measured their sentiment using the sentiment lexicon described in the previous section. About 5% of the tweets in the selected time frame contain geolocation information, a total of 2,071,851 tweets.

We started with examining regions. Dutch is spoken in The Netherlands (12 provinces) and Flanders (5 provinces). We made crude map of the 17 provinces and linked the boundaries to longitude and latitude figures from the coordinate system used in the tweet meta data (degrees with decimal part). Next, we determined for every tweet coordinate to which province it belonged using the point-in-polygon algorithm (Sutherland et al., 1974). We found tens of thousands of tweets per province, the lowest number for Belgian Limburg (29,489) and the highest number for South-Holland (313,312). The associated sentiment scores can be found in Figure 1.

The most striking observation that can be made from the map in Figure 1 is the difference between The Netherlands and Flanders. With the exception of the most southern Dutch province Limburg, all Dutch provinces obtain a sentiment score of nine or higher. Meanwhile the maximum score of the Flemish provinces is nine. This is not an isolated feature: we made similar observations for earlier months. There is no obvious reason why people in Flanders would tweet more negatively that people in The Netherlands. Cornips (2014) has suggested that the measured differences might be caused by dialect differences. If people from the southern regions of the map use words for expressing sentiment that are not part of our sentiment lexicon then the sentiment scores measured for their region will be closer to zero than the the scores measured in the northern regions. Since the average sentiment is positive, it will

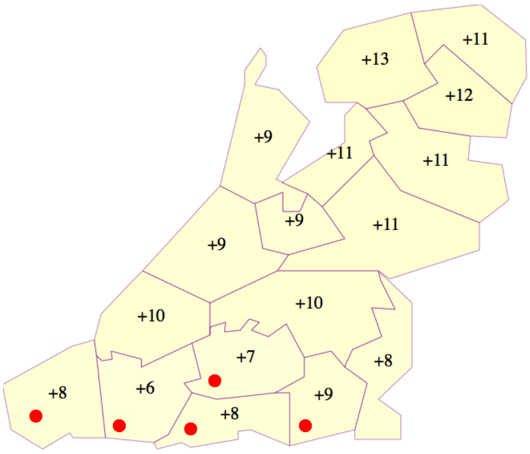

Figure 1: Twitter sentiment scores in the 17 Dutch-speaking provinces of Flanders and The Netherlands measured in January 2014. There is a clear difference between the sentiment scores of the provinces of The Netherlands on one side and the provinces of Flanders (marked with dots) on the other side.

appear that their tweets are less positive while this need not be the case.

Next, we performed sentiment analysis tweets originating from municipalities. The number of municipalities is too large to represent in a easily drawable map so we applied for an official Dutch municipality map from the Dutch mapping registry Kadaster. They offered the digital version of 2012 which was already outdated (417 instead of 403 municipalities). We mapped the tweets to municipalities using the point-in-polygon algorithm. The number of tweets per municipality was lower than for the provinces, with a minimum of 221 for Ouderkerk. For this reason, we computed the average of the sentiments per user (36 for Ouderkerk) rather than per tweet, otherwise one user with many tweets could have a large impact on the sentiment score of a municipality.

The resulting map for January 2014 can be found in Figure 2. Some interesting observations can be made. First, the larger population centers achieve sentiment scores at or below average: neither Amsterdam (+15), Rotterdam (+12), The Hague (+15), Eindhoven (+14), Tilburg (+15), Almere (+13), Breda (+13) nor Nijmegen (+14) does better than average (+15). Utrecht (+17) and Groningen (+16) are the only two of the ten most populated Dutch municipalities that achieve an above-average score.

A second observation is that all five Frisian islands in the north achieve very positive scores with the island of Schiemonnikoog appearing as one of the two most positive municipalities of The Netherlands. The fact that the islands are a popular tourist attraction probably has a positive influence on their mood on Twitter.

A third observation is that there are large differences between some neighboring municipalities. Voerendaal (+3) in the south has a relatively low score but neighboring Gulpen-Wittem (+22) is very positive. Grootegast (+5) in the north also has a relatively low score but it is surrounded by municipalities with scores around +20. Further study of the tweets involved is necessary to see if they mention important local issues that cause discomfort for their inhabitants.

5. Concluding remarks

We have described a sentiment analysis method for Dutch tweets based on a sentiment lexicon automatically derived from tweets. Words in the lexicon have been selected based on a comparison of positive and negative tweets with the t-test (Church et al., 1991). Several versions of the lexicon have been tested. We chose a manually developed lexicon of 338 tweets as the most appropriate for further experiments.

The sentiment lexicon has been used for determining the average sentiment of Dutch-speaking regions: provinces in Flanders and The Netherlands and municipalities in The Netherlands. The province sentiments revealed a surprising difference between Flemish and Dutch regions, most likely caused by differences in tweet vocabularies between the two areas. In the municipality results, we observed neutral busy regions and happy holiday regions. We also found some areas which were much less positive than their neighbors, a possible indication of local problems.

In all cases, one should be cautious in drawing conclusions from the sentiment measurements. They have been performed automatically and contain a certain degree of error. But one should also take into consideration that the demographics of Twitter is different from that of the Dutch-speaking community. This especially true for the users behind the tweets studied: people that freely share their location in their tweets. This group is predominantly male (66%) and over 25 years of age (56%). Only manual study of the tweets themselves can give an insight in why the users are positive or negative.

An obvious followup of this work is to try to find more indicators than the two used in this study (positive/negative), for example, crime, recreation, traffic, pollution, education and politics. A live view of local opinions of these topics would be interesting for policy makers. The main challenge here would be to collect enough tweets to be able to say something meaningful about the topics for all regions. Present day Twitter will probably not be able to satisfy that information need completely but it should prove to be a useful addition to other information sources.

6. References

Church, K., Gale, W., Hanks, P., and Hindle, D. (1991). Using Statistics in Lexical Analysis. In Zernik, U., editor, *Lexical Acquisition: Exploiting On-Line Resources to Build a Lexicon*. Lawrence Erlbaum Associates.

Cornips, L. (2014). Zeur-tweets. *Dagblad De Limburger*, 25 January.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. Technical report, Technical Report, Stanford Digital Library Technologies Project.

Incentro. (2012). Sentiment Analysis for the Dutch Language. `http://www.incentro.com/en/inspiration/blogs/sentiment-analysis-dutch-language`. Web page, accessed 3 February 2014.

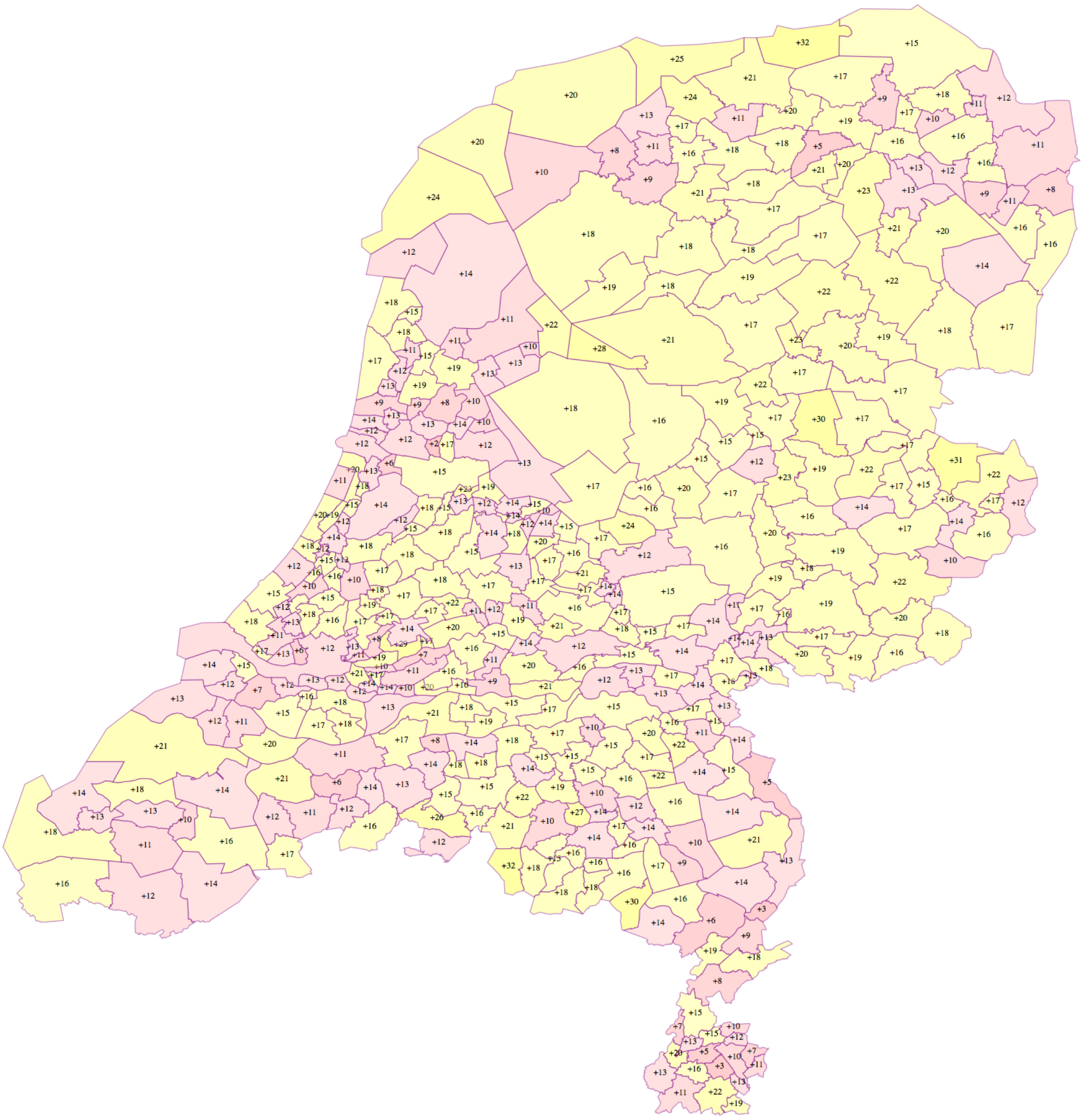

Figure 2: Twitter sentiment scores in the 417 Dutch municipalities measured in January 2014. Municipality areas include both land and water. Areas in yellow achieved an average (+15) or above average score while red areas achieved a score below average. Because of the low number of tweets, sentiment scores have been averaged over users rather than over tweets.

Mislove, A., Lehmann, S., Ahn, Y.-Y., Onnela, J.-P., and Rosenquist, J. N. (2010). Pulse of thge Nation: U.S. Mood Throughout the Day inferred from Twitter. http://www.ccs.neu.edu/home/amislove/twittermood/. Web page, accessed 4 February 2014.

Pak, A. and Paroubek, P. (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC-2010)*. European Language Resources Association (ELRA).

Pang, B., Lee, L., and Vaithyanathan, S. (2002). Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP-2002*, pages 79–86. ACL.

Sutherland, I. E., Sproull, R. F., and Schumacker, R. A. (1974). A Characterization of Ten Hidden-Surface Algorithms. *ACM Computing Surveys*, 6(1):1–55.

Tjong Kim Sang, E. and Bos, J. (2012). Predicting the 2011 Dutch Senate Election Results with Twitter. In *Proceedings of SASN 2012, the EACL 2012 Workshop on Semantic Analysis in Social Networks*, pages 53–60. ACL, Avignon, France.

Tjong Kim Sang, E. and van den Bosch, A. (2013). Dealing with Big Data: the Case of Twitter. *Computational Linguistics in the Netherlands Journal*, 3:121–134. ISSN: 2211-4009.

Twitter. (2011). Developer Rules of the Road. https://dev.twitter.com/terms/api-terms. Web page, accessed 4 February 2014.

White, T. (2012). *Hadoop: the definite guide*. O'Reilly.