A World List for Sentiment Analysis of Dutch Microblog Messages

Jeffrey Luppes, Radboud University Nijmegen, the Netherlands

Keywords: *Sentiment Analysis, Social networking sites, Language resources*

Abstract

While there are many resources for sentiment analysis of English messages there is relatively little material for the Dutch language. In this paper a new word list resource for sentiment analysis of messages in Dutch is proposed. This sentiment Lexicon is specifically designed to handle microblog messages such as Twitter and scores words based on emotional valence. This resource was generated by collecting words from a corpus of 5648 Dutch tweets, which were then annotated manually. The resulting list contains 2192 words scored on a Likert scale of between 1 and 5 for positive words, and -1 to -5 for negative words, and includes many words not found in previous lists. The word list was evaluated on a small collection of Dutch tweets and is capable of detecting positive or negative sentiment. The sentiment lexicon has particular applications for the analysis of sentiment in Dutch language in microblog messages on social media.

1. Introduction

Sentiment analysis is a subfield of natural language processing (NLP) that became popular with the mass availability of opinionated content on the internet [1, 2]. People feel a need to give their opinion online, and the vast amount of machine-readable opinion resources such as blogs, reviews, tweets, and posts has increased the potential of sentiment analysis [2]. Sentiment analysis can provide valuable insight into customer opinion and is therefore also the field of marketing specialists and analysts [3].

Particularly on the microblogging website Twitter users may post short opinionated bits of text (tweets). Tweets are limited to 280 characters and were, until recently, limited to 140 characters[1]. Twitter's switch to 280 from 140 characters is a very recent change and is outside the scope of this document. However, this character limit has considerable influence on the use of language: quite often messages seem to ignore the rules of spelling and grammar [4].

Although there are many resources and different approaches developed for the sentiment analysis of the English language the amount of resources for Dutch is limited, and that what is available does not translate well to the domain of the analysis of microblog postings.

2. Related Work

There are several approaches towards sentiment analysis that often include building a lexicon for analysis. Lexicon-based methods rely on a pre-constructed lexicon in which words are listed with scores in one or multiple dimensions, often called polarity or emotional valence [5]. This lexicon is then used by algorithmic tools to analyze the sentiment found in a source message. The benefit is that this approach is computationally inexpensive and still accurate, but it cannot deal with the fact that a word may have different polarities in different contexts and cannot handle sarcasm adequately [1]. Examples for of

[1] See https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html

sentiment lexicons for English include ANEW [6], the un-related AFINN [5], and a 6800-word list maintained by Liu [7]. Using emoticons to gauge sentiment in short messages is also a known approach, especially for detecting positive sentiment [8, 9].

It is also possible to forsake the manual creation of a lexicon and employ supervised or semi-supervised machine learning to learn about positive and negative messages [2]. Examples are SentiStrength which employed an algorithmic approach to find emotion in MySpace messages [4], and the development of classifiers to detect polarity [10].

For Dutch there are fewer available lexicons. The first is a 1994 data set by Hermans and De Houwer and contained 740 words [11]. It was validated in a 2013 research paper that resulted in a data set of 4300 words, scored for valence and several other categories, on a Likert scale from one to seven [12]. The DuOMAn Project [13, 14] resulted in a data set of over 9000 words scored for polarity by two human annotators. It should be mentioned that the aim of the DuOMAn project was not to create a list of human-annotated polarity assessments, but it was used as a seed with a PageRank approach to generate subjectivity scores for a data set composed out of the Dutch Wordnet database and the Cornetto database of over 100.000 words [13, 15]. These datasets are available under license, but are not available for open source projects.

Pattern is a web mining application programmed in the Python language developed by the Computational Linguistics and Psycholinguistics Research Center (CLiPS) of the University of Antwerp [16]. It incorporates a widely used sentiment analysis library that uses a Dutch word list based on 3900 adjectives, which are taken from the Cornetto database [15]. In Pattern's data set words are scored based on polarity.

A comparison of the Dutch language lexicons mentioned here is shown in table 1.

Dataset	**Year**	**Number of words**	**scale**
Moors et al.	2013	4300	1 to 7
DuOMAn	2009	9088	-- to ++ (5 levels)
Pattern	2014[2]	3918	-1 to +1

Table 1: a comparison of Dutch language lexicons for sentiment analysis.

There are also several commercial products available that offer sentiment analysis in Dutch. These resources are not considered here as they are not publicly available.

3. Approach

Tweets were collected from the Twitter search API[3] based on a collection of 288 user accounts. Users were selected on the basis of having a publicly accessible profile and predominately Dutch language tweets and re-tweets. The API option of showing only Dutch tweets was not used, as it had many false positives.

In order to create a diverse collection the accounts were selected from multiple categories (such as sports, news, politics) but also from what appeared to be personal accounts. For each account the last twenty tweets were requested, or all if there were less tweets available. This resulted in a collection of 5648 tweets. Each tweet was further processed and stripped of everything but the message. Messages were also stripped of any mention of account names (e.g. @name) that are frequently featured in postings, but did collect words from hashtags (e.g. #fun). Emoticons (e.g. :-)) were removed. After applying a filter

[2] *Based on Tom de Smedt's commit 6417d13 on Sep 12, 2014, to https://github.com/clips/pattern/blob/master/pattern/text/nl/nl-sentiment.xml*

[3] For Twitter's API reference see: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-search-tweets.html

for only alphabetical content, 9511 words remained to be further analyzed.

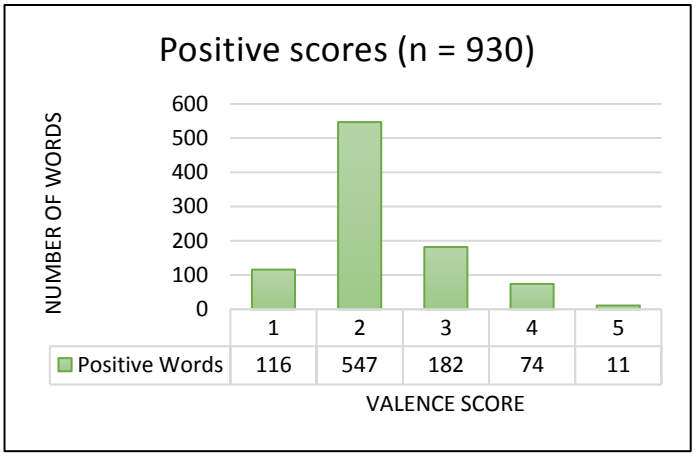
This set of words was then inspected and non-Dutch words, as well as misspellings of Dutch words and Dutch words that were ambiguous in meaning or did not contain emotion were pruned. This resulted in a collection of 2192 words for annotation.

To score the collected words the method of SentiStrength [4] was used, which provides instructions to score words based on a Likert scale of 1 to 5 for both positive and negative emotion. Here, 1 indicates little to no emotional payload, while 5 contains very strong emotion. To ease computation, this was transformed into a valence scale of -5 to +5, with most positive and negative words scored as +2 or -2 respectively. No word was scored 0. This is essentially the same approach as used by Nielsen, who also worked alone, in the construction of AFINN [5]. Words were annotated for valence by the author of this paper, who is a native speaker of the Dutch language.

The resulting set contains 1262 (58%) negative words, and 930 (42%) positive words. The most common score for both negative and positive words was 2, which is consistent with the findings with AFINN which was constructed with the same method [5]. Histograms of both the negative and positive words are shown in figure 1.

The resulting lexicon contains many new negative words such as "*kech*" (a relatively new word translating to "slut"), "*trol*", "*hoax*", and "*nepnieuws*" ("fake news"). When the lexicon is compared to DuOMAn and Pattern there are 1806 (82.4%) words that do not appear in the Pattern lexicon, and 1535 (70%) words do not appear in the much larger DuOMAn lexicon. It should however be noted that neither Pattern nor DuOMAn was created for short informal text.

The sentiment lexicon does not include emoticons and nor does it consider emojis.



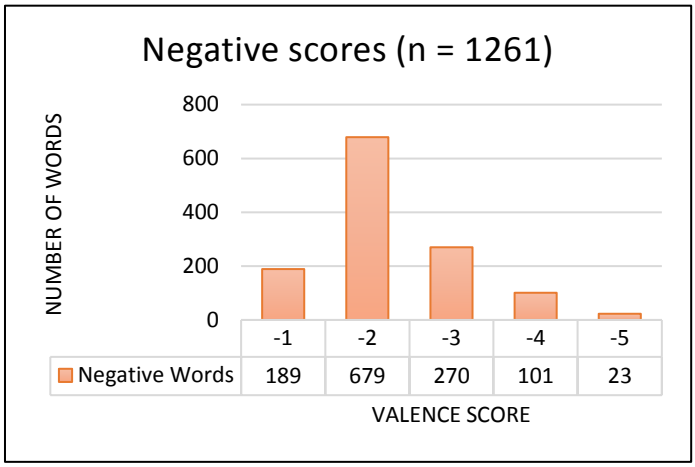


Figure 1a: Histogram of positive scores in the lexicon.
Figure 1b: Histogram of negative scores in the lexicon.

Phrases are also not included, due to the tokenization used when collecting the words from tweets. However, it is quite possible to combine the list with an emoji lexicon such as [8], which already tackled this problem.

4. Evaluation

A small test data set of 80 tweets was collected from twitter and annotated for positive (1) or negative polarity (0) by a second student. These tweets were not used for the construction of the lexicon and were collected from a different time frame.

For judging the sentiment score of a tweet the individual words were scored and the tweet's score was the sum of all of its words. The Rand Index [17] was used to evaluate the accuracy of the resulting cluster of scores. Here a score close to 1 means that two clusters are exactly alike while a score close to zero means clusters

are different on every point. When treating the annotated labels as the true class of sentences and the valence scores of the tweets as classes, the Rand index is 0.7, indicating a decent correspondence to the annotated labels. This is comparable to the reported accuracy of implementations of AFINN on Amazon, Yelp and IMDB reviews, which varied between 0.67 and 0.76 [18].

When evaluating performance it appeared that 10 of the 80 tweets (12.5%) did not contain any words currently present in the sentiment lexicon. Upon inspection it seems that there are many words that could have made it to the word list such as "*gedood*" (killed), "*coderood*" (code red), "*onthoofde*" (decapitated) and "*sick fucks*". There are also many positive words that were not found, such as "*indrukwekkendste*" (most impressive) and "*prachtstad*" (beautiful city).

5. Discussion

The results of the short evaluation show that the lexicon could benefit from more words. The data set was relatively small, as the UCI data sets mentioned in the comparative evaluation for English for AFINN, all contain 1000 comments [19]. Additionally, the Rand index used does not account for chance generating the results, which might not be considered a problem given that the sentiment ratings are discrete. The lexicon seems to be quite useful for finding sentiment in messages and covers many words that were not present in other sets.

Some bias is introduced through the annotation process as it is based on the language understanding of the author of the lexicon. It also does not include many words that on twitter appear almost exclusively as negative terms but that still refer to real-life persons, processes, and minorities outside of twitter.

6. Conclusion

In this paper a new sentiment lexicon for the Dutch language was introduced. Despite the small size, the data set contains many words not present in other, state-of-the-art sets. It could therefore be a good supplement to existing resources. Existing word lists were not used as a basis to increase the lexicon size due to conflicting licenses and methodologies.

The list is arguably most useful for the analysis of short informal texts such as twitter and social media comments, as it was collected from a body of 5648 tweets. The annotated word list could also serve as a basis for training algorithms. The sentiment score might be implemented in software in many different ways and could also serve as an input feature for classifiers.

The performance of the list would be improved by annotating more words, including emoticons and phrases. It has been released into the public domain for use.

7. References

[1] Bing Liu. 2015. *Sentiment analysis: Mining opinions, sentiments, and emotions.* Cambridge University Press, Cambridge, UK.

[2] Bo Pang and Lillian Lee. 2008. *Opinion mining and sentiment analysis.* Hanover, MA: Now Publishers.

[3] Tom De Smedt. 2013. *Modeling Creativity: Case Studies in Python*. Ph.D. Dissertation. University Press Antwerp, Antwerp.

[4] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. 2010. Sentiment strength detection in short informal text. *Journal of the Association for Information Science and Technology 61, no. 12* (2010): 2544-2558.

[5] Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages.* Volume 718 in CEUR Workshop Proceedings: 93-98. 2011 May. Matthew Rowe, Milan Stankovic, Aba-Sah Dadzie, Mariann Hardey (editors)

[6] Margaret M. Bradley, and Peter J. Lang. 1999. *Affective norms for English words (ANEW): Instruction manual and affective ratings*. Technical report C-1, the center for research in psychophysiology, University of Florida, FL.

[7] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004.

[8] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. *PloS one* 10, no. 12 (2015): e0144296.

[9] Jonathon Read. 2005. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pp. 43-48. Association for Computational Linguistics, 2005.

[10] Andrea Esuli and Fabrizio Sebastiani. 2005. Determining the semantic orientation of terms through gloss classification. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 617-624. ACM, 2005.

[11] Dirk Hermans, and Jan De Houwer. 1994. Affective and subjective familiarity ratings of 740 Dutch words. *Psychologica Belgica 34,* no. 2-3 (1994): 115-139.

[12] Agnes Moors, Jan De Houwer, Dirk Hermans, Sabine Wanmaker, Kevin Van Schie, Anne-Laura Van Harmelen, Maarten De Schryver, Jeffrey De Winne, and Marc Brysbaert. 2013. Norms of valence, arousal, dominance, and age of acquisition for 4,300 Dutch words. Behavior research methods 45, no. 1 (2013): 169-177.

[13] Valentin Jijkoun and Katja Hofmann. 2009. Generating a non-english subjectivity lexicon: Relations that matter. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 398-405. Association for Computational Linguistics.

[14] Valentin Jijkoun and Katja Hofmann. 2008. *Task-based evaluation report: Building a Dutch subjectivity lexicon*. ILPS-ISLA, University of Amsterdam, the Netherlands.

[15] Piek Vossen, Isa Maks, Roxane Segers, Hennie Van Der Vliet, Marie-Francine Moens, Katja Hofmann, Erik Tjong Kim Sang, and Maarten de Rijke. 2013. *Cornetto: a combinatorial lexical semantic database for Dutch*. In Essential Speech and Language Technology for Dutch, pp. 165-184. Springer Berlin Heidelberg.

[16] Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *Journal of Machine Learning Research 13*, no. Jun (2012): 2063-2067.

[17] William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association* 66.336 (1971): 846-850.

[18] Andrew Sliwinski. 2018. AFINN-based sentiment analysis for Node.js. Retrieved from https://github.com/thisandagain/sentiment/.

[19] Dimitrios Kotzias, Misha Denil, Nando De Freitas, and Padhraic Smyth. 2015. From group to individual labels using deep features. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 597-606. ACM.