Context-Sensitive Sentiment Classification of Short Colloquial Text

Norbert Blenn, Kassandra Charalampidou, and Christian Doerr

Department of Telecommunication
TU Delft, Mekelweg 4, 2628CD Delft, The Netherlands
{N.Blenn, C.Doerr}@tudelft.nl, K.Charalampidou@student.tudelft.nl

Abstract. The wide-spread popularity of online social networks and the resulting availability of data to researchers has enabled the investigation of new research questions, such as the analysis of information diffusion and how individuals are influencing opinion formation in groups. Many of these new questions however require an automatic assessment of the sentiment of user statements, a challenging task further aggravated by the unique communication style used in online social networks.
This paper compares the sentiment classification performance of current analyzers against a human-tagged reference corpus, identifies the major challenges for sentiment classification in online social applications and describes a novel hybrid system that achieves higher accuracy in this type of environment.

Keywords: Online Social Networks, Sentiment Analysis, Text Classification

1 Introduction

The amble availability of data from online social networks in machine-readable format has made it possible to investigate and evaluate a whole new set of research questions at a large scale, such as "how do trends form?", "what determines how influential a person is?" or "how do our friends and contacts shape our opinion?". Many research questions posed in online social network analysis thus require to be able to assess the context and meaning of a user's statements, identifying in the simplest case whether a sentence is a neutral, objective comment or a subjective opinion, or in more advanced scenarios tracing and quantifying the development and flow of positive/negative thoughts across a user's various communication threads.

In recent years, a number of methods for sentiment analysis have been proposed; these techniques have however been developed for and are consequently geared towards the extraction of meaning in large text corpora, such as product reviews, letters or articles. User communication in online social networks such as Facebook or Twitter has however very specific and challenging features: 1) Messages are short and highly abbreviated and therefore only a small angle of attack for an automatic classifier, 2) Text is very colloquial and typically deprived of any context information, thus making it difficult to infer the subject and reference of the sentiment.

Due to these special characteristics, traditional sentiment analysis methods do not provide sufficiently accurate results when applied towards online social network communications. This paper deviates from these established sentiment classification approaches and describes an alternative method utilizing additionally both grammatical and contextual information for increased detection accuracy. The contributions of this paper are two-fold: First, we evaluate a set of available sentiment classifiers against a dataset obtained from the social microblogging platform Twitter, and measure the detection performance of these automatic tools against a human classification. Second, we identify common problems in sentiment analysis and demonstrate how an alternative approach can provide a higher detection accuracy than previous context-less classifiers, and beside pure identification of sentiment polarization, can additionally provide a magnitude quantification of sentiments.

The remainder of this paper is structured as follows: Section 2 overviews previous work in sentiment classification, with a specific focus on online social network sentiment classification. Section 3 describes the evaluation corpus, compares the detection performance of existing approaches and discusses common problems with short colloquial text analysis. Section 4 introduces our hybrid approach and outlines additional application use cases. Section 5 summarizes our findings.

2 Related Work

Sentiment analysis, i.e., the extraction of an opinion's overall polarization and strength towards a particular subject matter, is a recent research direction [1, 2], and typically approached from a statistical, or machine-learning angle. Attention has been given particularly in the domain of movies [3, 4], by analysis of social media data, as reflection of common opinion. It is found that prices of the movies industry have a strong correlation with observed outcome frequencies, and therefore they are considered as good indicator of future outcomes. Most recently published work either perform unsupervised learning on a provided corpus of perceived positive and negative texts such as product reviews [5, 6], or use a set of curated keywords with positive or negative connotations to classify input [7, 1].

Another common approach [5, 6] is measuring sentence similarity between given data input and texts of specific polarity, which explores the hypothesis that opinion sentences will be more similar to other opinion sentences that to factual ones. Additionally, previous work [8, 9] was focused on learning extraction patterns associated with objectivity (and subjectivity) in order to be used as features of objective/subjective classifiers. It is shown that this approach achieves higher recall and comparable precision than previous techniques. Apart from that, recent publications [10, 11], introduced the use of Natural Language

Processing modules in order to extract concepts from the processed text and eventually derive sentiment out of them. In the very recent past, several of these general approaches have been specifically extended towards the mining of sentiments from online social media sources, in particular the microblogging platform Twitter [12]. For our analysis, we focus on those approaches for which the original authors made a reference implementation available to us, specifically we compare the classification accuracy with the following classifiers:

Twitter Sentiment We used the bulk classification service available on the Twitter Sentiment website [13] in order to classify our test-set. This tool attaches to each tweet a polarity value: 0 for negative, 4 for positive and 2 for neutral- therefore we consider the first to describe subjective tweets, while neutral is for objective tweets. The main idea behind Twitter Sentiments approach is the use of emoticons as noisy labels for the training data which is shown that it increases the accuracy of different machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM). It is noted that the web service of Twitter Sentiment uses a Maximum Entropy classifier.

Tweet Sentiments Our test-set was also tested through the API of Tweet-Sentiments [14], a well known tool for analysing Twitter data and provide sentiment analysis on tweets. TweetSentiments is based on Support Vector Machines (SVM) and is using the LIBSVM library developed at Taiwan National University. It classifies tweets as positive, negative or neutral and these values are treated as stated previously.

Lingpipe We also used the Sentiment Analysis tool of the LingPipe [15] package which focuses on the subjective/objective (as well as positive/negative) sentence categorisation especially on the movie-review domain. This approach uses the usual machine learning algorithms (Naive Bayes, Maximum Entropy, SVM) and a Java API of the classifier is available online. Even though it comes with its own training set, we used the half of our hand-classified set to train the classifier, in order to have better results. The other half was used as test-set and results were compared to the corresponding hand-classified tweets.

3 Methodology, Test Corpus and Performance Evaluation

To evaluate the performance of established sentiment classifiers and create a benchmark for our developed solution, we randomly sampled a set of some 1,000 publicly readably messages from the microblogging platform Twitter. Prime use cases for sentiment analysis are for example research questions revolving around the spread of information, opinion formation and identification of influential relationships in social networks, and such processes are typically believed to be present in discussions around product and media such as music, book or movie reviews.

Data Acquisition and Processing. For our evaluation, we therefore collected a data-set of 1,073 randomly chosen tweets related to the five most popular films of the 83rd Academy Awards. We used the language detection library of Cybozu Labs [16], in order to eliminate the tweets written in any language other than English, while we also tried to remove advertising tweets out of the set. Multiple retweets of the same text were also removed to prevent performance over- or underestimation, as well as unnecessary tokens like link urls, "@" tags for mentioning a user, 'RT' tags etc. Each tweet of this test-set was classified by hand before the begin of the evaluation into an objective or subjective statement; this corpus is used throughout this work as a reference benchmark.

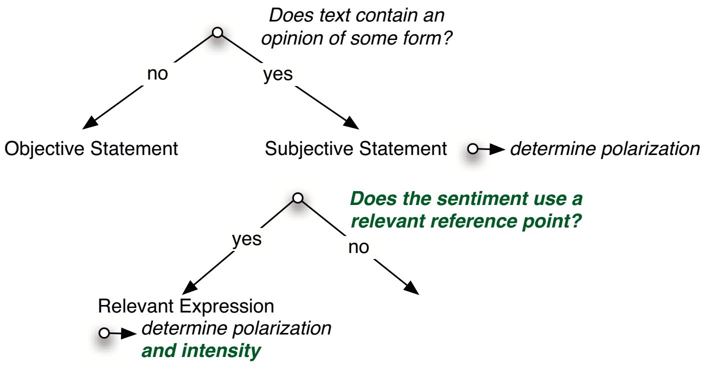

Fig. 1. Sentiment classification involves a multi-stage process, in which not only the existence of a sentiment should be checked, but also the reference point and strength should be assessed. (Contributions of this work are indicated in green.)

Components in Sentiment Analysis. Starting point for any sentiment analysis (as shown in figure 1) is the detection of any form of opinion in written text. If the author expresses some form of judgement, the input can be considered a subjective statement, otherwise the data is classified as an objective claim. Example cases distinguished by such test are for example "I liked *The King's Speech*" versus, "*The King's Speech* was a really long movie.", respectively. Another case is given by subjective messages where the sentiment is not based on the relevant reference point (the title of the movie). For example the tweet "I like you, even when watching *The King's Speech*" is a positive tweet, but its not the opinion about the movie that is positive. Once the existence of a sentiment has been established, typically a classification step is performed to determine whether the speaker is expressing a positive or negative opinion over a particular subject matter.

Challenges for Classification in Short Colloquial Text. In many types of inputs, and specifically in micro-texts such as tweets or chats, however a problem arises: conversations are highly abbreviated. As tweets offer only 140 characters of payload, messages are reduced to a bare minimum and several different thoughts - for example reactions to previous incoming messages - frequently are abbreviated and intertwined: "Watched King's Speech today in class. I love the end of the term." This results in a very small footprint on which sentiment analysis can be conducted, at least compared to the essay- and article-type classification previously used for polarization analysis. Previously established approaches, which for example operate using a statistical word-frequency analysis, are therefore less suited, as the low quantities of text and the high concept compression ratios are resulting in very high statistical fluctuations and noise during the detection. For this reason, we pursue a different approach in this work and analyze the grammatical structure of messages. By detecting which concepts a particular sentiment is referencing to, we can make more fine-grained decisions and consequently achieve a higher prediction accuracy especially in dense, intertwined texts. In the example above this concept is the end of the term (and therefore potentially a looser schedule) rather than the movie itself.

Automatic Detection and Tuning of Polarization Intensity. Finally, many applications and research hypotheses can be better served if not only the existence of a sentiment and its general polarization is known, but an absolute notion of "how positive" or negative a particular opinion can be derived. This would allow both a better assessment of how opinions are propagated and adopted, as a person with a strong negative attitude towards a particular concept is first expected to become less and less negative before developing a positive sentiment if at all. Without a quantification of polarization such trends would go unnoticed. Additionally, a quantitative measurement of attitude would allow of differentiation between alternatives, which in sum are all considered positively, but in a pairwise comparison are not equal.

If considered in previous work, this aspect is typically approached using manually curated word lists, as for example in [2]. When following this strategy in the application context of micro-messages, we however discovered two fundamental difficulties: 1) Users utilize a rich set of vocabulary to describe their opinions about concepts. Capturing and maintaining an accurate ranking of evaluative comments would require a significant effort in a practical setting. 2) Expressions indicating positive and negative sentiments and their relative differences are neither stable over time nor between different people, thus a method to re-adjust and "normalize" sentiment baselines over time or between say generally very positively oriented, neutral or pessimistic speakers will provide an advantage.

Evaluation of the State-of-the-Art. To evaluate the performance of existing sentiment classifiers, we let the set of available classifiers described in section 2 analyze and distinguish a reference body of tweets. As most methods do not allow for a sentiment quantification, we limited this evaluation to only a general polarization detection which is supported by all systems. Comparing the output against the previous human classification, we measured the overall accuracy of the automatic classifiers in distinguishing subjective from objective statements as shown in figure 2(a). Figure 2(b) shows the overall performance in correctly and incorrectly classified statements.

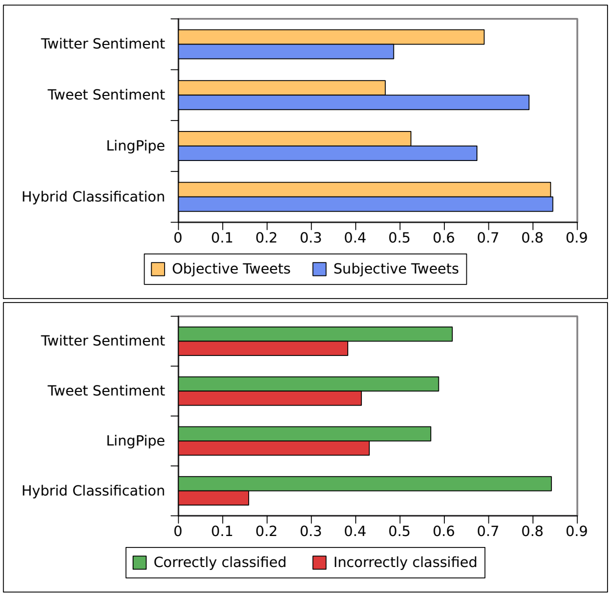

Fig. 2. Classification accuracy of different sentiment analyzation methods.

As can be seen in the figure, the classification accuracy of all statistical sentiment analyzers is between 55 and 60%, whereas the proposed statistical-grammatical hybrid approach yields a correct classification accuracy of about 85%, a 40% gain over previous work. Note also that the accuracy of existing system also varies significantly between the type of input data: Twitter Sentiment [13] for example is much stronger identifying objective statements compared to subjective ones, while Tweet Sentiment [14] shows exactly the opposite behavior. The proposed hybrid solution on the other hand does not show any significant bias.

4 Sentiment Classification & Polarity Estimation

This section describes in detail the underlying concept of the proposed hybrid sentiment classifier. As shown in figure 1, the general task of sentiment analysis can be conducted in two general phases, first the detection of opinions in general (yielding to a categorization in objective and subjective text), after which a quantification of the polarity can be attempted. The following discussion mirrors these steps.

4.1 Grammatical Sentiment Classification

In order to classify a given message into subjective and objective we analyze the grammatical structure of a tweet. Subjectivity is mostly based on adjectives or verbs expressing the polarity related to the subject of the message. This means if directly expressing an emotion one usually uses a verb like "like/love/hate" whereas expressing the mood about something usually contains an adjective. Consider the examples: *I'm feeling sick today*, *I liked the movie*.

To determine the existence of a sentiment, we therefore invert the grammatical structure of a given text to detect the presence of verbs carrying an emotional meaning or to find the adjectives associated with the keywords we are performing the sentiment analysis on, in our case the titles of movies. Grammatical structure analysis is a mature research area and we use Klein and Manning's lexicographical parser [17] to determine the structure of the English texts of tweets after removing special characters as described in section 3. For a given text, this tool is estimating the grammatical structure. An example for the tweet "I liked the movie" is given in the following:

```
[I, liked, the, movie]
(ROOT
  (S
    (NP (PRP I))
    (VP (VBD liked)
      (NP (DT the) (NN movie)))))

nsubj(liked-2, I-1)
det(movie-4, the-3)
dobj(liked-2, movie-4)
```

Here, the parser is reasoning that "I" is the nominal subject of "liked", and "movie" is the direct object of the verb "liked". As mentioned, most tweets expressing a sentiment have this kind of structure. If an adjective is referring to a subject the likelihood is quite high that this tweet expresses the mood about something. Note however that it is in general possible that the speaker was using sarcasm or irony, which could not be detected from a grammatical viewpoint.

In a second step, we cross-check whether the word referring to the subject of a tweet is an adjective. This can be done using either a lexical database such as WordNet [18] or a part-of-speech Tagger [19], which will be used further in this discussion. Through such a tool, every word in a given sentence can be annotated with a tag identifying its purpose in the sentence [20], so the example "I liked the movie" is marked as:

```
I/PRP liked/VBD the/DT movie/NN
```

The part of speech tagger tells us that "I" is a personal pronoun, "liked" a verb in past tense, "the" a determiner and "movie" a noun. By connecting the so gathered information of a message we build simple rules to detect if a message is a subjective statement whereas all the others by inversion have to be objective:

1. if an adjective is referring to the subject
2. if an verb out of a list is referring to the subject
3. adjective + [movie, film]
4. [movie, film] is/looks [adjective]
5. love/hate + [movie, film]

4.2 Automatic Polarity Estimation

After the existence of a subjective component has been established, it is necessary to determine the overall polarity of the sentiment and if possible also the magnitude of the sentiment. In order to estimate the general polarity direction of words in the corpus, we used an unsupervised approach based on word correlations. This approach is inspired by the way a person is learning to judge which words have a positive or negative meaning, which is essentially a result of a lot of exposure to speech and written text, from which the learner infers which words appear in a positive or negative context.

The same basic principle, inferring which words appear together in a positive or negative context, can however be easily mirrored in a machine as well. Here, a computer would simply need to count how often a particular adjective has been encountered with a positive meaning compared to the frequency it has been observed with a negative connotation. To begin such an automatic classification, some notion of what is deemed positive or negative will be necessary. In our work, we have explored two general options, first by manually specifying a set of keywords one would associate with positive expressions such as “fantastic”, “incredible”, “amazing”, which can read from existing databases such as [18], and second by looking at the most basic positive/negative expression commonly used in online messages such as chats, emails or microblogs: a positive smiley :-) and a negative smiley :-(. In the following, we will discuss the results for this second alternative.

From a list of one million tweets, we search for all tweets containing a positive smiley :-), :) or =) – in the following referred to as *positive keywords* – and created a list of texts containing at least one of those symbols. Similarly, a list of all tweets containing a negative smiley such as :-(, :(or =(– deemed *negative keywords* – was prepared. Using the techniques discussed above, we dissect all individual statements and count the number of co-occurrences between every detected word and the positive or negative keywords. To correct for differences in length of those two lists – as users typically write more positive than negative statements –, the two values are then normalized by the number of words in the list of messages containing positive words and the list of messages containing negative words. This results in a relative assessment of a particular word to appear in a positive or a negative context, where context is defined by the positive and negative keywords, respectively.

To arrive at a relative polarity of a particular word between the two extremes “positive” and “negative”, it is now simply enough to subtract the relative frequencies. This number is positive if the word is typically used within a positive context and negative if the word typically occurring with a negative meaning.

As this number is biased by the number how often a word is used in general, a final correction step is executed in which each rating is multiplied by the term frequency measured in all tweets: the emphasis of frequently observed words is therefore reduced, and the value of unusual ones is lifted.

Fig. 3 shows the output of this simple procedure conducted over a body of one million tweets, and using just positive and negative smileys as corresponding positive and negative keywords. As can be seen, this unsupervised process leads to a clear and meaningful ranking of adjectives. We have repeated this analysis over a selection of datasets of different duration and verified the automatically generated polarity estimation against those done by a human. Typically, less than 7% of the words were considered wrongly placed in the overall order; out of a list of 30 adjectives between one and two placement were deemed higher or lower in the ranking by a human observer than by a machine. As this method can be executed without manual intervention, this new methodology can be continuously conducted to detect the general development of sentiments in entire online communities, as well as to identify whether the polarity of certain words shift in strength over time.

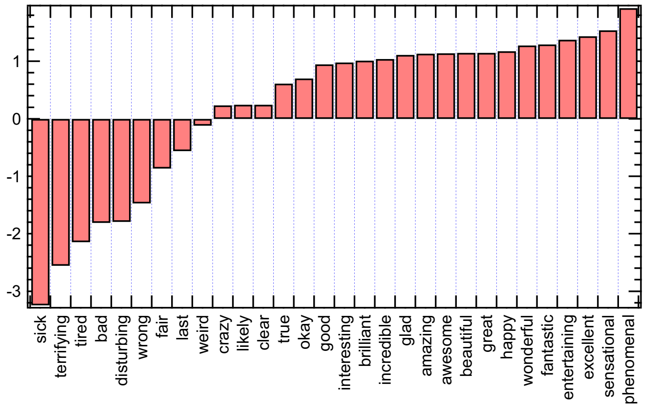

Fig. 3. Automatic polarity ranking of adjectives based on general Twitter messages.

A more comprehensive analysis is needed to see if the sentiment of adjectives change over time. However taking twitter data of durations of one week, one and two months the polarity seems to stabilize the longer the duration of observation.

4.3 Detecting Networks of Concepts

This general method is however not limited to determine the polarity strength of words in general, but can be used to detect and identify common concepts and their associated sentiments in general. To do so, it is simply necessary to swap out the two sets of keywords (which in the last section were *:-), :), =)* and *:-(, :(, =(*, respectively), and replace them with those terms and synonyms relevant to a particular study.

Consider for example a situation where one would to determine the associations made with the brands and products of two hypothetical tea manufacturers: *McArrow's orange-peppermint tea* and *DrBrew's strawberry-melon tea*. Here, one would populate *keyword group 1* with words from the first area, i.e., McArrows, orange-peppermint, etc. and analogously *keyword group 2* with words such as DrBrew, strawberry-melon, etc., and by the same means described above, this method would derive the set of words frequently used in combination with any of those keyword terms as well as the strength of their typical common appearance as shown in figure 4.

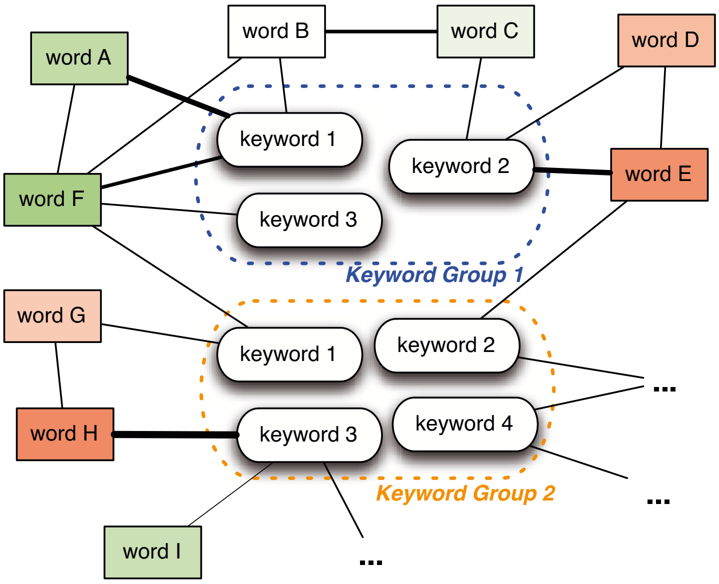

Fig. 4. The technique can be broadened to determine the concepts commonly associated with any keyword as well as the particular strength of the association.

The words $A - I$ discovered to be co-located can however be themselves further interpreted, for example, 1) depending on how positively/negatively they are (as discussed above), or 2) which general topic areas or word field the concepts come from. Imagine for example words A and F in figure 4 to be "taste" and "flavor", while words D and E are "packaging" and "price". Clearly, such combined word co-localization, polarization and word-field analysis will provide a significant insight to our hypothetical tea manufacturer, which can also be easily repeated over time to track its overall development, but also to the researcher interested in how particular opinions form, are spread and change over time.

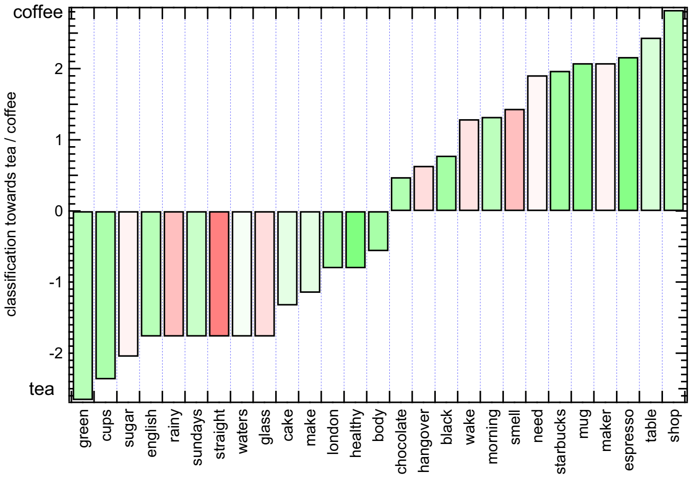

Fig. 5. Polarization analysis of commonly used words with coffee and tea based on general Twitter messages.

As a general example, we have applied this techniques towards the keywords "coffee" and "tea", and let the system arrange the resulting associated words according to their overall polarity strength. The overall abstract network can then be drawn in a similar manner as figure 3, and figure 5 shows the 27 strongest connections commonly made the terms coffee and tea. All bars indicating their affiliation towards the two beverages are colored by their polarity value. Red indicates a negative, green a positive and white a neutral polarity.

5 Conclusion

In this paper, we have presented an alternative method to determine the existence and strength of subjective opinions in short colloquial text, an application domain where existing approaches do not yield a high detection accuracy. The proposed system works through a combination of grammatical analysis with traditional word frequency analysis, does not need supervised training and improves the accuracy of previous work by about 40%.

Acknowledgements. This project was partially supported by the Transsector Research Academy for Complex Networks and Services (TRANS).

References

1. B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," in *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2002.
2. P. D. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002.
3. G. Mishne and N. Glance, "Predicting movie sales from blogger sentiment," in *AAAI 2006 Spring Symposium on Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, 2006.
4. S. Asur and B. Huberman, "Predicting the future with social media," in *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, vol. 1, pp. 492–499, IEEE, 2010.
5. H. Yu and V. Hatzivassiloglou, "Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 129–136, 2003.
6. L. Barbosa and J. Feng, "Robust sentiment detection on twitter from biased and noisy data," in *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pp. 36–44, 2010.
7. R. Cilibrasi and P. Vitanyi, "Automatic meaning discovery using google," *Manuscript, CWI*, 2004.
8. E. Riloff and J. Wiebe, "Learning extraction patterns for subjective expressions," in *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pp. 105–112, Association for Computational Linguistics, 2003.
9. J. Wiebe and E. Riloff, "Creating subjective and objective sentence classifiers from unannotated texts," *Computational Linguistics and Intelligent Text Processing*, pp. 486–497, 2005.
10. L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification," *Proc. 49th ACL: HLT*, vol. 1, pp. 151–160, 2011.
11. E. Cambria, A. Hussain, C. Havasi, and C. Eckl, "Sentic computing: Exploitation of common sense for the development of emotion-sensitive systems," *Development of Multimodal Interfaces: Active Listening and Synchrony*, pp. 148–156, 2010.
12. "Twitter, http://www.twitter.com."
13. A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," tech. rep., Stanford University, 2009.
14. "Tweet sentiments, http://tweetsentiments.com."
15. "Lingpipe, http://alias-i.com/lingpipe."
16. N. Shuyo, *Language Detection Library.* http://code.google.com/p/language-detection/, 2010.
17. D. Klein and C. D. Manning, "Accurate unlexicalized parsing," in *Proceedings of the 41st Meeting of the Association for Computational Linguistics*, 2003.
18. "Wordnet. a lexical database for english. http://wordnet.princeton.edu/."
19. D. K. C. M. Kristina Toutanova and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of HLT-NAACL*, 2003.
20. "The university of pennsylvania (penn) treebank tag-set http://www.comp.leeds.ac.uk/ccalas/tagsets/upenn.html."