Chapter 17

SMARTPHONE MESSAGE SENTIMENT ANALYSIS

Panagiotis Andriotis, Atsuhiro Takasu, and Theo Tryfonas

Abstract Humans tend to use specific words to express their emotional states in written and oral communications. Scientists in the area of text mining and natural language processing have studied sentiment fingerprints residing in text to extract the emotional polarity of customers for a product or to evaluate the popularity of politicians. Recent research focused on micro-blogging has found notable similarities between Twitter feeds and SMS (short message service) text messages. This paper investigates the common characteristics of both formats for sentiment analysis purposes and verifies the correctness of the similarity assumption. A lexicon-based approach is used to extract and compute the sentiment scores of SMS messages found on smartphones. The data is presented along a timeline that depicts a sender's emotional fingerprint. This form of analysis and visualization can enrich a forensic investigation by conveying potential psychological patterns from text messages.

Keywords: SMS messages, Twitter feeds, emotion, timeline

1. Introduction

The evolution of mobile phones from simple telephone devices to sophisticated handheld computers has transformed how users communicate. Modern smartphones are widely used to send email and chat with friends via instant messages. However, the older SMS (short message service) is still one of the most popular mobile phone services because of its simplicity. The textual information in SMS messages convey the thoughts and emotions of the sender.

A forensic investigation typically involves the examination of electronic devices such as computers, hard drives, tablets and smartphones. The goal is to retrieve relevant information, recover deleted files and

G. Peterson and S. Shenoi (Eds.): Advances in Digital Forensics X, IFIP AICT 433, pp. 253–265, 2014.

present the extracted data in an efficient and intuitive manner. Text messages are an integral part of mobile communications and potentially constitute valuable evidence.

Research efforts in the areas of natural language processing and information retrieval have studied micro-blogging services such as Twitter because of their openness, ubiquity and information content [11, 14]. The extraction of the emotional polarity of text messages is an active research topic. Other researchers (e.g., [19]) note the marked similarity between a tweet (Twitter post) and an SMS message, claiming that Twitter feed analysis techniques can be used to analyze SMS messages. The two formats do indeed have common characteristics. Both involve a limited number of characters (140 for tweets and 160 for SMS messages). In addition, the "@" feature in Twitter can be viewed as a direct message to a specific person, just like an SMS message sent to a contact. Moreover, special symbols such as emoticons and other abbreviations are widely used in Twitter and SMS communications.

This paper extends the textual sentiment analysis work to the domain of digital forensics. It has four main contributions. First, the efficiency of a lexicon-based sentiment analysis algorithm on various datasets (Twitter feeds and SMS messages) is investigated, and the claim that mood analysis methods used on Twitter feeds can also be used on SMS messages is evaluated. Second, the importance of word pre-processing methods is examined along with their contribution to sentiment scoring. Third, the significance of a lexicon is studied and the optimization of a lexicon-based method to improve SMS mood scores is discussed. Finally, a forensic tool that performs sentiment analysis is described; the tool can be used by investigators to search for keywords, compute the sentiment scores of messages and present the sentiment scores in a timeline view.

2. Background

Sentiment analysis is the process of identifying positive and negative opinions about a subject or topic from a piece of text. Prior work in this field used lexical knowledge and the emotional valence of the words included in vocabularies [12].

Knowledge-based methods use linguistic models to categorize the sentiment of passages of text. These methods construct and use dictionaries in order to capture the sentiments of words. The models can be crafted in a manual [4] or semi-automated manner [20]. Pang, *et al.* [16] implemented three classifiers (naive Bayes, maximum entropy and support vector machines) to evaluate their efficiency in categorizing movie

reviews as negative or positive; they conclude that support vector machines work better in most cases. They emphasize that using a more complicated linguistic model such as n-grams does not dramatically improve the results, suggesting that a unigram approach may be sufficient.

Sentiment analysis has been applied to diverse problems ranging from online forum hotspot detection [11] to sentiment classification in microblogs [2, 5]. Melville, *et al.* [12] have developed a framework for sentiment analysis of blogs that combines lexical knowledge with supervised learning for text categorization; they report that it is preferable to incorporate both methods to obtain good results for blog analysis. Taboada, *et al.* [19] have used dictionaries of words with predefined characteristics (polarity and strength) and show that lexicon-based methods for sentiment analysis are robust and can be used in a variety of domains without any training with domain-specific data. Their lexicon-based implementation performs well on diverse tasks ranging from video game reviews to blog post classification.

Twitter is an open micro-blogging service that makes all posts available to the public. The public nature of the data provides researchers with the opportunity to use Twitter as a corpus [14]. Bermingham and Smeaton [2] suggest that it is easier to infer the sentiment polarity of micro-blogging posts compared with blogs, which have richer textual content (they also note the similarity between micro-blogging posts and SMS messages). However, the automatic processing of micro-blogging posts can be problematic because of the use of non-standard words and unusual punctuation [9]. Leong, *et al.* [10] use sentiment mining to analyze SMS messages in teaching evaluations.

In the area of digital forensics, text analysis has been used to extract patterns from email and to construct user-profiles from text [6]. Several researchers (see, e.g., [3]) have focused on investigations and the modeling of texting languages. Despite an extensive search of the literature on mood analysis, we were unable to find any research that engaged principles from opinion and text mining to conduct mobile device forensics.

3. Experimental Setup and Datasets

This section describes the datasets used in the research, the algorithm and the experimental setup.

The Twitter dataset (TWT) used in the research contained 6,566 tweets collected from 33 popular Twitter accounts on August 5, 2013. The accounts belonged to musicians, actors, athletes and managers. A dataset (SENT140) of tweets classified as positive (PoSENT140) and negative (NegSENT140) was also used [7].

The SMS dataset contained 5,574 messages with duplicates [1]. Removing the duplicates yielded 4,827 unique messages that were manually classified as conveying positive or negative sentiments. Negative messages were assumed to express anger, fear, sadness, disgust and boredom (919 messages). Positive messages were assumed to express joy and happiness (1,867 messages). Numerous messages that were not classified as positive or negative were marked as neutral.

The algorithm uses a bag-of-words approach and employed three lexicons containing words linked to positive or negative emotions. The AFINN lexicon [8] contains words with valences ranging from –5 to 5. The Wordnet-Affect lexicon [18] consists of synsets linked to affective labels and words. The NRC lexicon [13] was originally used in a competition on sentiment analysis on Twitter feeds. The lexicons were sanitized by eliminating hashtag symbols (#) and words that appeared to be irrelevant to sentiment analysis (e.g., "5yo" and plain letters like "t"). The sanitized vocabularies consisted of 25,675 words and abbreviations corresponding to positive emotions, and 20,636 corresponding to negative emotions.

The methodology for calculating sentiment scores used the three lexicons to score each tweet and SMS. Let L_p be the set of positive textual markers (positive lexicon) with $l_{pi} \in L_p$ for $i = 1, 2, \ldots, m$ and let L_n be the set of negative textual markers (negative lexicon) with $l_{nj} \in L_n$ for $j = 1, 2, \ldots, n$. The corpus C consists of tweets and SMS messages and $t_k \in C$ denotes an individual message for $k = 1, 2, \ldots, q$. If a positive marker l_{pi} appears in a tweet or SMS (t_k) in the corpus:

$$l_{pi}(t_k) = 1. \tag{1}$$

Otherwise, $l_{pi}(t_k)$ is set to zero. The same calculations were performed for negative markers l_{nj}. The Boolean assignment is based on the fact that the documents are of limited size; thus, there is a limited range of markers in each tweet that contribute to the tweet sentiment score. The tweet sentiment score $s(t_k)$ is equal to the number of positive markers found in a tweet minus the number of negative markers found in the tweet:

$$s(t_k) = \sum_i l_{pi}(t_k) - \sum_j l_{nj}(t_k). \tag{2}$$

4. Experimental Results

This section describes the experimental results related to determining the impact of stemming on the dataset, identifying the lexicon that results in the best classification accuracy, and analyzing the impact of emoticons.

Table 1. Effect of stemming on Twitter sentiment scores.

	Dataset	None	Total	Neutral
No Stem	TWT	39.1%	60.9%	48.0%
	PoSENT140	28.4%	71.6%	41.5%
	NegSent140	25.3%	74.7%	39.3%
Stem	TWT	31.8%	68.2%	42.3%
	PoSENT140	23.3%	76.7%	36.6%
	NegSent140	18.7%	81.3%	34.7%

4.1 Stemming Results

Stemming is a popular text pre-processing method to obtain the root of a word. For example, the words "connect," "connected" and "connection" have the same root that can be represented by the lemma "connect" [17]. The stemming test discussed in this section evaluates if stemming combined with Equation (2) increases classification accuracy. The experiments used the AFINN lexicon on the Twitter dataset and subsequently on the SMS dataset.

Table 2. Distribution of textual markers in the Twitter datasets.

	Dataset	1	2	3	4	5	6	7+
No Stem	TWT	47.7%	28.7%	14.3%	5.9%	2.1%	0.8%	0.5%
	PoSENT140	40.8%	29.0%	16.1%	8.0%	3.6%	1.5%	1.0%
	NegSent140	43.4%	28.3%	15.4%	7.4%	3.4%	1.3%	0.8%
Stem	TWT	41.6%	28.6%	16.6%	7.8%	3.3%	1.3%	0.8%
	PoSENT140	36.6%	28.0%	17.7%	9.5%	4.7%	2.1%	1.4%
	NegSent140	36.9%	28.1%	17.4%	9.4%	4.7%	2.1%	1.4%

Each tweet from the TWT and SENT140 datasets was stemmed using Porter's stemming algorithm from the Apache Lucene Library prior to calculating the sentiment score. Table 1 compares the sentiment scores obtained with and without stemming. The column "None" lists the percentages of tweets that did not contain matching words from the lexicon. "Total" lists the percentages of tweets that contained at least one matching word (positive or negative). "Neutral" lists the percentages of tweets that had a sentiment score of zero. Table 2 shows the number of textual markers found in a single tweet.

Table 1 shows that stemming improves the identification of tweets. Another observation is that the Boolean sentiment scores have high "Neutral" classifications. The algorithm sorted 48% of the tweets as

Table 3. Effect of stemming on SMS sentiment scores.

	None	Total	Neutral $s(t_k)$
No Stem	40.6%	59.4%	50.2%
Stem	31.9%	68.1%	42.8%

Table 4. Distribution of textual markers in the SMS dataset.

	1	2	3	4	5	6	7	8/9	10+
No Stem	45.1%	25.2%	14.5%	6.8%	3.7%	2.5%	0.8%	1.0%	0.4%
Stem	42.2%	25.0%	14.6%	7.5%	4.4%	2.6%	1.7%	1.6%	0.4%

neutral, but the percentage of the messages that did not contain any matching words was 39.1%. This difference is because several tweets have zero sentiment scores (i.e., equal numbers of positive and negative markers). In fact, Table 2 shows that the positive and negative datasets have almost the same numbers of matching words. This finding is an indication that a balanced lexicon (i.e., AFINN) is used in the bag-of-words approach.

Tables 3 and 4 present the results of the same experiment conducted on the SMS dataset. The lexicon used the AFINN vocabulary and the SMS dataset was first analyzed without stemming and then with stemming after applying Porter's algorithm. Table 3 shows that stemming enhances the efficiency of SMS sentiment analysis because the number of classified messages is increased from 59.4% to 68.1%. Table 4 supports the assessment that AFINN is a balanced lexicon with similar numbers of positive and negative scores. The results also confirm that the methods used for Twitter sentiment analysis help discern the emotional trends of users during forensic examinations of smartphones.

4.2 Lexicon Results

Three vocabularies were compared to test the generalizability of the sentiment score: (i) the AFINN lexicon (Tables 1 through 4); (ii) the WordNet-Affect lexicon, which contains formal words that are not widely used in the micro-blogging world; and (iii) the NRC lexicon, which contains many words and hashtags (#) associated with positive or negative emotions. The NRC lexicon was sanitized to fit the bag-of-words approach, but many lemmas such as "okayyy" were present in both categories (positive and negative), making us skeptical about its use. The NRC lexicon also included symbols and Internet slang.

In the case of the WordNet-Affect lexicon, roughly one to three matching words were found in each tweet and 68.5% of the tweets had neutral sentiment scores. For the NRC lexicon, 20.6% of the tweets had neutral emotional scores while 96.9% of the tweets contained at least one word from the lexicon. However, the distributions of matching words were very different compared with the previous results. For example, there were as many as 20 "matching" words in a single tweet. This occurred because many textual markers were present in the positive and negative lexicons. The fact that such an extensive and diverse lexicon could not decrease the percentage of tweets rated with $s(t_k) = 0$ motivated us to use the AFINN lexicon in the remainder of the study.

4.3 Emoticon Results

Many people, particularly young smartphone users, include symbols and special words to abbreviate their messages and produce more concise text. The most popular symbols are called "emoticons." An emoticon is a symbol that expresses an emotion; for example, happiness is represented using a smiling face ":)" emoticon. Based on the characteristics of the SMS dataset, the AFINN lexicon was extended to include emoticons.

Due to the number of neutral sentiment scores, the computation of sentiment scores was modified to include valence. In lexicons such as NRC and AFINN, words are presented with their emotional valence. In AFINN, for example, the textual marker "amazing" has a valence of 4 and the word "approval" has a valence of 2. The maximum valence is 5 and the minimum valence is –5, denoting the most powerful negative emotion. We rated the emoticons as having higher valences on the assumption that users add these symbols to their text messages to express strong feelings. Hence, emoticons should have the highest weight when incorporated in the sentiment scores. In this case, Equation (2) does not change, and the difference arises in the manner in which the contribution of each marker to the final score is computed. Instead, $l_{pi}(t_k) = 1$ in Equation (1) is changed to $l_{pi}(t_k) = v$ where $v \in \{-5, -4, \ldots, 4, 5\}$ is the valence of each textual marker.

Table 5 shows the results obtained after repeating the tests on the SMS dataset with the emoticon-enhanced AFINN lexicon and the revised sentiment score computation that includes valence. Note that the addition of emoticons and valence both have a positive effect on the success rates and reduce the numbers of false positives and false negatives.

Table 5. Success and failure rates of lexicon-based sentiment analysis.

	Datasets	Total	FPR	FNR	Hit Rate
AFINN	Positive	1,867	12.1%	32.6%	55.3%
	Negative	919	22.7%	32.4%	44.8%
Emoticons	Positive	1,867	11.5%	31.1%	57.4%
	Negative	919	22.0%	31.7%	46.3%
Emoticons and Valence	Positive	1,867	7.3%	23.9%	68.8%
	Negative	919	29.2%	25.0%	45.8%

5. Combining Sentiment Analysis with Forensics

A common practice during forensic analysis is to use Linux commands such as `strings` and `grep` to view and elaborate text. Additionally, an analyst may use open source tools such as the SQLite Database Browser and SQLiteman to view the contents of databases. Our forensic tool extends the database view and incorporates the sentiment analysis approach presented in this paper. This enables the tool to be used to search for keywords in a database, compute the sentiment scores of messages, and present the sentiment scores in a timeline view.

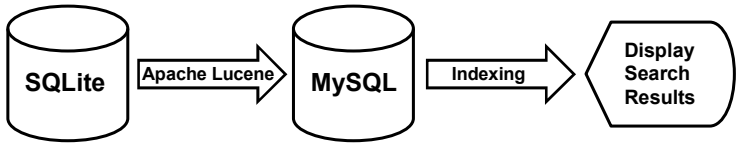

Figure 1. Schematic diagram of the forensic tool.

Figure 1 shows a schematic diagram of the forensic tool. First, all the requested content from an SMS SQLite database (`/data/com.android.providers.telephony/databases/mmssms.db`) is stored in a MySQL database. Next, Apache Lucene is used to produce unstemmed SMS message keywords for searching, and stemmed words for SMS mood analysis. After all the data is processed, search indexes are constructed to answer text queries posed by forensic examiners.

The forensic tool was tested on an image extracted from a Samsung Galaxy Y S5360 smartphone running the Android 2.3.5 operating system. Unfortunately, the database contained non-English messages; since these messages had no lexical matches, no sentiment scores were generated. We substituted SMS messages in `mmssms.db` with random tweets from the TWT database. After running the tool, we had a MySQL

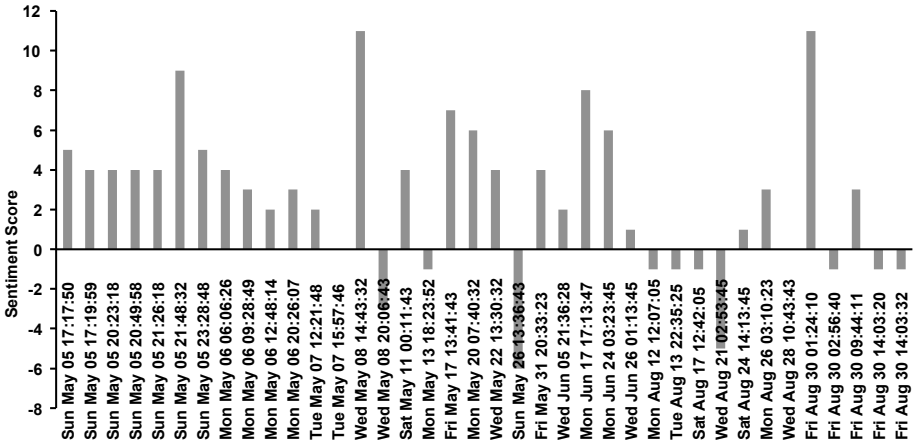

Figure 2. Timeline emotional analysis of SMS messages.

database containing the data from `mmssms.db` and the keywords extracted from each "message" along with the computed sentiment scores.

The forensic tool is designed to provide forensic examiners with a timeline analysis interface that conveys the emotion of each SMS message found on a smartphone. This could provide a quick indication of the emotional states of a user over time. For example, if the person being investigated was generally in a good mood, except for a certain period of time when negative emotions were observed, it might be worthwhile for a forensic examiner to focus on that specific time period.

Figure 2 shows the mood timeline of SMS messages extracted from the SQLite database, including the messages sent to and received from all contacts. Note that the smartphone user mostly exchanged messages that have positive emotional fingerprints. However, there are some periods (e.g., from 12th August until 17th August) when the sentiment scores are negative. The timeline view provides the forensic examiner with a general perspective of user behavior as well as the opportunity to focus on periods that may be of special interest.

The forensic tool enables an examiner to focus on the communications between two parties, or to view the sentiment scores of only the received or sent messages. For example, Figure 3(a) shows the sentiment scores of all the messages exchanged with a specific device. Figure 3(b) shows the scores of all the messages sent by the user.

Finally, the MySQL database may be searched using the index. Figure 4 shows the results obtained for the search term "happy." The tool returned four hits, each with details such as the message ID in the origi-

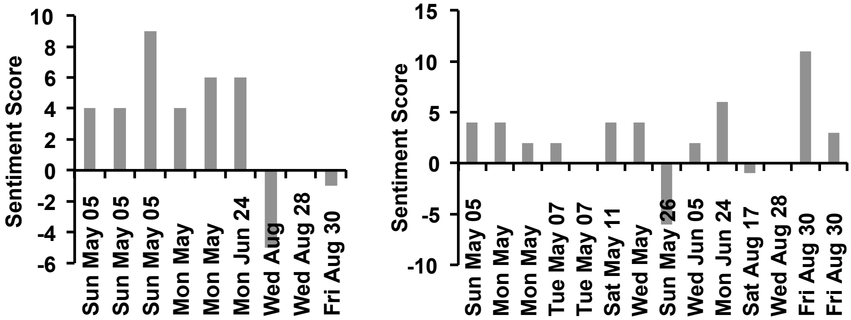

Figure 3. Timeline emotional analysis of SMS messages.

```
Found 4 hits.
1. It is a boy! So happy for my cousin Kate and the future King of England!
---- Database id: 11 on Mon May 06 20:26:07 JST 2013 received from: +3[redacted]48
2. Happy Halloween from me &amp; the dummy  http://t.co/jQxPvHqi
---- Database id: 32 on Mon Aug 26 03:10:23 JST 2013 received from: +4[redacted]94
3. Happy 4th of July! In 1776, Betsy Ross signed the Gettysburg Address and celebrated at the Boston Tea Party. It was a magical day.
---- Database id: 19 on Mon May 20 07:40:32 JST 2013 received from: +3[redacted]8
4. Congratulations to my dear friend @JimmyKimmel and his beautiful new wife! I am so happy for him. I had no idea he was straight.
---- Database id: 14 on Wed May 08 14:43:32 JST 2013 received from: +3[redacted]40
```

Figure 4. Searching the MySQL database.

nal SQLite database, the date the message was sent or received, and the contact involved in the communication.

6. Conclusions

Current methods used for Twitter sentiment analysis are very useful for depicting the emotional polarity of SMS messages. The results obtained using emoticons and the sentiment valence of words demonstrate the value of a bag-of-words approach. Furthermore, the results can be enhanced by employing a dynamic lexicon.

The digital forensic tool described in this paper merges traditional string searches with text mining to expedite the retrieval of sentiment indications from SMS messages. In particular, the tool can be used by investigators to search for keywords, compute the sentiment scores of messages and present the sentiment scores in a timeline view. This functionality can enrich a forensic investigation by revealing the psychological patterns of users from their text messages.

Our future research will employ the mood analysis approach to produce a timeline view of all smartphone content, including email, instant messages, notes and social network activities. Also, we plan to investigate the application of support vector machine classifiers to enhance the efficiency of the digital forensic tool.

Acknowledgement

This research was supported by the European Union's Prevention of and Fight against Crime Programme "Illegal Use of Internet" – ISEC 2010 Action Grants (HOME/2010/ISEC/AG/INT-002), the National Institute of Informatics, the Systems Centre of the University of Bristol and Project NIFTy (HOME/2012/ISEC/AG/INT/4000003892).

References

[1] T. Almeida, J. Hidalgo and A. Yamakami, Contributions to the study of SMS spam filtering: New collection and results, *Proceedings of the Eleventh ACM Symposium on Document Engineering*, pp. 259–262, 2011.

[2] A. Bermingham and A. Smeaton, Classifying sentiment in microblogs: Is brevity an advantage? *Proceedings of the Nineteenth ACM International Conference on Information and Knowledge Management*, pp. 1833–1836, 2010.

[3] M. Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar and A. Basu, Investigation and modeling of the structure of texting language, *International Journal of Document Analysis and Recognition*, vol. 10(3-4), pp. 157–174, 2007.

[4] S. Das and M. Chen, Yahoo! for Amazon: Sentiment extraction from small talk on the web, *Management Science*, vol. 53(9), pp. 1375–1388, 2007.

[5] X. Ding, B. Liu and P. Yu, A holistic lexicon-based approach to opinion mining, *Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 231–240, 2008.

[6] D. Estival, T. Gaustad, S. Pham, W. Radford and B. Hutchinson, Author profiling for English emails, *Proceedings of the Tenth Conference of the Pacific Association for Computational Linguistics*, pp. 263–272, 2007.

[7] A. Go, R. Bhayani and L. Huang, Twitter Sentiment Classification using Distant Supervision, CS224N Final Project Report, Department of Computer Science, Stanford University, Stanford, California, 2009.

[8] L. Hansen, A. Arvidsson, F. Nielsen, E. Colleoni and M. Etter, Good friends, bad news-affect and virality in Twitter, in *Future Information Technology*, J. Park, L. Yang and C. Lee (Eds), Springer, Berlin-Heidelberg, Germany, pp. 34–43, 2011.

[9] G. Laboreiro, L. Sarmento, J. Teixeira and E. Oliveira, Tokenizing micro-blogging messages using a text classification approach, *Proceedings of the Fourth Workshop on Analytics for Noisy Unstructured Text Data*, pp. 81–88, 2010.

[10] C. Leong, Y. Lee and W. Mak, Mining sentiments in SMS texts for teaching evaluation, *Expert Systems with Applications*, vol. 39(3), pp. 2584–2589, 2012.

[11] N. Li and D. Wu, Using text mining and sentiment analysis for online forums hotspot detection and forecast, *Decision Support Systems*, vol. 48(2), pp. 354–368, 2010.

[12] P. Melville, W. Gryc and R. Lawrence, Sentiment analysis of blogs by combining lexical knowledge with text classification, *Proceedings of the Fifteenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1275–1284, 2009.

[13] S. Mohammad, S. Kiritchenko and X. Zhu, NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets, *Proceedings of the Seventh International Workshop on Semantic Evaluation Exercises*, 2013.

[14] A. Pak and P. Paroubek, Twitter as a corpus for sentiment analysis and opinion mining, *Proceedings of the Seventh International Conference on Language Resources and Evaluation*, pp. 1320–1326, 2010.

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

[16] B. Pang, L. Lee and S. Vaithyanathan, Thumbs up? Sentiment classification using machine learning techniques, *Proceedings of the Association for Computational Linguistics Conference on Empirical Methods in Natural Language Processing*, vol. 10, pp. 79–86, 2002.

[17] M. Porter, An algorithm for suffix stripping, *Program: Electronic Library and Information Systems*, vol. 14(3), pp. 130–137, 1980.

[18] C. Strapparava and R. Mihalcea, Semeval-2007 Task 14: Affective text, *Proceedings of the Fourth International Workshop on Semantic Evaluations*, pp. 70–74, 2007.

[19] M. Taboada, J. Brooke, M. Tofiloski, K. Voll and M. Stede, Lexicon-based methods for sentiment analysis, *Computational Linguistics*, vol. 37(2), pp. 267–307, 2011.

[20] H. Yu and V. Hatzivassiloglou, Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences, *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 129–136, 2013.