Quantifying the role of the opinion lexicon in sentiment analysis

Yannick Marchand
Faculty of Computer Science
Dalhousie University,
6050 University Avenue
Halifax, NS, B3H 1W5, Canada
ymarchan@dal.ca

Vlado Keselj
Faculty of Computer Science
Dalhousie University
6050 University Avenue
Halifax, NS, B3H 1W5, Canada
vlado@cs.dal.ca

Evangelos Milios
Faculty of Computer Science
Dalhousie University
6050 University Avenue
Halifax, NS, B3H 1W5, Canada
eem@cs.dal.ca

Michael Shepherd
Faculty of Computer Science
Dalhousie University
6050 University Avenue
Halifax, NS, B3H 1W5, Canada
shepherd@cs.dal.ca

ABSTRACT
A widespread approach to determine the sentiment of a text is to use a pre-established opinion lexicon that includes negative and positive lexical entities. However, it is not clear how the choice of this lexicon impacts the classifier performance. In this paper, a comparison of seven opinion lexicons on six sentiment datasets (movie reviews and tweets) is conducted. Results suggest that increasing the lexicon size by semantic expansion as well as assigning an interval value to the words of the opinion lexicon significantly increases the classification performance on short texts (e.g. tweets).

General Terms
Measurement, Performance.

Keywords
Sentiment analysis, Opinion lexicon, Tweets.

1. INTRODUCTION
In the past decade there has been a tremendous growth in user-generated content on the World Wide Web in the form of social media such as blogs, reviews, social networks, and forums. The immense amount of text data that accumulates as a result of social media has recently been advantageously exploited as a valuable resource to detect real world events (e.g., disasters [1], influenza epidemics [2], stock market variations [3], and oil price changes [4]). In addition, through social media, it is possible to understand both public opinion and sentiment regarding these events. As such, opinions expressed within the context of social media can be highly influential. For instance, one survey [5] found that 61% of people rely on information from reviews when making a purchase decision.

Influence12 - Symposium on Measuring Influence on Social Media, September 28-29, 2012, Halifax, NS, Canada.

One popular, simple and widely-used approach by companies in analytics is to determine the sentiment of a text by counting the occurrences of its positive and negative terms (e.g. [6]). In the case of a binary classification, a text is labeled positive if it contains more positive than negative words and vice versa. This matching technique generally relies on the existence of an opinion lexicon, in which a pre-established list has already been generated of both positive and negative terms. Surprisingly, in spite of the growing body of work in the field of opinion mining, no study appears to have studied and quantified the impact of the opinion lexicon on sentiment classification performance.

2. OBJECTIVE AND DATASETS
Our paper addresses this issue by comparing seven opinion lexicons on six sentiment datasets. The opinion lexicons that are evaluated in the present study consist of twitrratr [6], a Financial Dictionary [7], Subjclues [8], UIC [9]), as well as three versions of LabMT1.0. The dictionary LabMT1.0 [10] consists of a list of more than 10,000 words, which are classified on an "index of happiness," which is based on a 1 to 9 integer scale (1= least happy, 5 = neutral, and 9 most happy).

In regard to the six sentiment datasets, three are based on movie reviews [11, 12, 13] and three consist of twitter data [14, 15, 16]. Tables 1 and 2 summarize the key features for the opinion lexicons and sentiment datasets, respectively.

3. RESULTS
An "index of encapsulation" (defined for two sets as the percentage of elements of the smallest set that are also present within the largest set) as well as the Jaccard index were calculated between each pair of the opinion lexicons. The results demonstrated that the seven lexicons under investigation were highly dissimilar (Mean of the Jaccard index = 0.1, Standard deviation = 0.15).

Table 1. Summary of the opinion lexicons (KW: keywords)

Lexicon	Negative KW	Positive KW	Total KW
Financial Dictionary [6]	2350	354	2074
Subjclues [7]	4147	2298	6445
Twitrratr [5]	142	117	259
UIC [8]	4780	2003	6783
LabMT1.0 (v.1) [9]	1063	2668	3731
LabMT1.0 (v.2) [9]	410	598	1008
LabMT1.0 (v.3) [9]	51	26	77

In an effort to better understand the reason of this discrepancy, we expanded the lexicons by using the semantic relations of synonymy and antonymy of WordNet [17]. By applying this semantic propagation approach, both indexes increased significantly (p-value < 0.05). This suggests that the lexicons overlap more in terms of semantic field than in terms of lexical entities.

Table 2. Summary of the sentiment datasets

Data	Negative labels	Positive labels	Total labels
Movie reviews [6]	1000	1000	2000
Movie reviews [7]	5331	5331	10662
Movie reviews [6]	25000	25000	50000
Tweets [9]	177	182	359
Tweets [9]	1193	705	1898
Tweets [8]	571	522	1093

To illustrate this finding, let's take the simple example of two lexicons with only two terms: Lex1={'good': Positive; 'death': Negative} and Lex2={'great': Positive; 'hate': Negative}. Before semantic propagation, there is no overlap at all. Imagine that through the operator of synonymy (the case of antonymy is discarded on purpose for the sake of simplicity) the two following new lexicons are obtained: SYNONYMY(Lex1)={'good': Positive; 'great': Positive; 'death': Negative} and SYNONYMY(Lex2)={'good': Positive; 'great': Positive; 'hate': Negative}. There is now overlap (i.e. 'good' and 'great' belong to the two dictionaries) and the two indexes subsequently increase.

Accuracy performances on the movie review datasets were shown to be higher than on the Twitter datasets (p-value < 0.01). This finding is not surprising due to the fact that Twitter data is known, from a natural language processing perspective, to be challenging because it is of limited length (maximum 140 characters), unstructured, and extremely noisy (e.g. repeated letters, abbreviations, and misspellings). It was also found that the classification accuracy was higher when the lexicons were semantically expanded relative to their original form. These results were statistically significant for the Twitter data but not for the movie review datasets.

Finally and interestingly, it was possible to investigate whether or not the type of scale of measurement of positivity (or negativity) for a word matters for sentiment analysis. To assess this, the index of happiness of the words of LabMT1.0 was applied to the other lexicons that contained the same words [which were previously labeled with only a nominal value (positive or negative)]. Our results demonstrated that this value transfer was beneficial to the binary sentiment classification and the accuracy significantly improved (p-value < 0.01).

4. CONCLUSION

When conducting sentiment analysis using keyword technique, our results highlight the strong benefits of, firstly, taking into consideration the opinion lexicon to be selected, secondly, increasing the lexicon size by semantic expansion and lastly, assigning an interval value to the words of the opinion lexicon.

5. ACKNOWLEDGMENTS

This work was supported by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC Engage Grant). The authors thank Dr. Celeste Lefebvre for her feedback on this paper.

6. REFERENCES

[1] S. Doan, B-K. Ho Vo, and N. Collier (2011). An analysis of Twitter messages in the 2011 Tohoku Earthquake. In *4th ICST International Conference on eHealth*.

[2] A. Culotta (2010). Towards detecting influenza epidemics by analyzing Twitter messages. In *Proceedings of the First Workshop on Social Media Analytics*.

[3] J. Bollen, H. Mao, and X. Zeng (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), pp. 1-8.

[4] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, pp. 122-129.

[5] R. Petersen (2011). http://barnraisersllc.com/2011/07/34-case-studies-prove-social-commerce-roi/

[6] Twitrratr (2012). http://twitrratr.com/

[7] T. Loughran and B. McDonald (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-ks. *The journal of finance*. Vol. LXVI, nº1, pp. 35-65.

[8] T. Wilson, J. Wiebe, and P. Hoffmann (2005). Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of HLT/EMNLP 2005*.

[9] M. Hu and B. Liu. (2004). Mining and Summarizing Customer Reviews. In *KDD*, pp. 168-177.

[10] I. Kloumann, C. Danforth, K. Harris, C. Bliss, and P. Dodds (2012). Positivity of the English language. PLoS ONE 7(1), pp. 1-7.

[11] B. Pang and L. Lee (2004). A sentimental education: Sentimental analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual*

Meeting of the Association for Computational Linguistics, pp. 271-278.

[12] B. Pang and L. Lee (2005). Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pp. 115-124.

[13] A. Maas, R. Daly, P. Pham, D. Huang, A. Ng, and C. Potts (2011). Learning Word Vectors for Sentiment Analysis. In *The 49th Annual Meeting of the Association for Computational Linguistics.*

[14] A. Go, R. Bhayani, and L. Huang (2009). Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford.

[15] D.A. Shamma, L. Kennedy, and E. Churchill (2009). Tweet the debates. *ACM Multimedia Workshop on Social Media*.

[16] N. Sanders (2011). http://www.sananalytics.com/

[17] G. Miller (1995). Wordnet: a lexical database for English. *Communications of the ACM*. Vol. 38 (11), pp. 39-41.