Is Twitter A Better Corpus for Measuring Sentiment Similarity?

Shi Feng[1], **Le Zhang**[1], **Binyang Li**[2,3], **Daling Wang**[1], **Ge Yu**[1], **Kam-Fai Wong**[3]
[1]Northeastern University, Shenyang, China
[2]University of International Relations, Beijing, China
[3]The Chinese University of Hong Kong, Shatin, N.T., Hong Kong
{fengshi,wangdaling,yuge}@ise.neu.edu.cn, zhang777le@gmail.com
{byli,kfwong}@se.cuhk.edu.hk

Abstract

Extensive experiments have validated the effectiveness of the corpus-based method for classifying the word's sentiment polarity. However, no work is done for comparing different corpora in the polarity classification task. Nowadays, Twitter has aggregated huge amount of data that are full of people's sentiments. In this paper, we empirically evaluate the performance of different corpora in sentiment similarity measurement, which is the fundamental task for word polarity classification. Experiment results show that the Twitter data can achieve a much better performance than the Google, Web1T and Wikipedia based methods.

1 Introduction

Measuring semantic similarity for words and short texts has long been a fundamental problem for many applications such as word sense disambiguation, query expansion, search advertising and so on.

Determining the word's polarity plays a critical role in opinion mining and sentiment analysis task. Usually we can detect the word's polarity by measuring it's semantic similarity with a positive seed word se_p and a negative seed word se_n respectively, as shown in Formula (1):

$$SO(w) = sim(w, se_p) - sim(w, se_n) \quad (1)$$

where $sim(w_i, w_j)$ is the semantic similarity measurement method for the given word w_i and w_j. A lot of papers have been published for designing appropriate similarity measurements. One direction is to learn similarity from the knowledge base or concept taxonomy (Lin, 1998; Resnik, 1999). Another direction is to learn semantic similarity with the help of large corpus such as Web or Wikipedia data (Sahami and Heilman, 2006; Yih and Meek, 2007; Bollegala et al., 2011; Gabrilovich and Markovitch, 2007). The basic assumption of this kind of methods is that the word with similar semantic meanings often co-occur in the given corpus. Extensive experiments have validated the effectiveness of the corpus-based method in polarity classification task (Turney, 2002; Kaji and Kitsuregawa, 2007; Velikovich et al., 2010). For example, PMI is a well-known similarity measurement (Turney, 2002), which makes use of the whole Web as the corpus, and utilizes the search engine hits number to estimate the co-occurrence probability of the give word pairs. The PMI based method has achieved promising results. However, according to Kanayama's investigation, only 60% co-occurrences in the same window in Web pages reflect the same sentiment orientation (Kanayama and Nasukawa, 2006). Therefore, we may ask the question whether the choosing of corpus can change the performance of sim and is there any better corpus than the Web page data for measuring the sentiment similarity?

Everyday, enormous numbers of tweets that contain people's rich sentiments are published in Twitter. The Twitter may be a good source for measuring the sentiment similarity. Compared with the Web page data, the tweets have a higher rate of subjective text posts. The length limitation can guarantee the polarity consistency of each tweet. Moreover, the tweets contain graphical emoticons, which can be

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 897–902,
Seattle, Washington, USA, 18-21 October 2013.

considered as natural sentiment labels for the corresponding tweets in Twitter. In this paper, we attempt to empirically evaluate the performance of different corpora in sentiment similarity measurement task. As far as we know, no work is done on this topic.

2 The Characteristics of Twitter Data

As the world's second largest SNS website, at the end of 2012 Twitter had aggregated more than 500 million registered users, among which 200 million were active users . More than 400 million tweets are posted every day.

Several examples of typical posts from Twitter are shown below.

(1) *She had a headache and feeling light headed with no energy.* :(

(2) *@username Nice work! Looks like you had a fun day. I'm headed there Sat or Sun.* :)

(3) *I seen the movie on Direc Tv. I ordered it and I really liked it. I can't wait to get it for blu ray! Excellent work Rob!*

We observe that comparing with the other corpus, the Twitter data has several advantages in measuring the sentiment similarity.

Large. Users like to record their personal feelings and talk about the trend topics in Twitter (Java et al., 2007; Kwak et al., 2010). So there are huge amount of subjective texts with various topics generated in the millions of tweets everyday. Further more, the flexible Twitter API makes these data easy to access and collect.

Length Limitation. Twitter has a length limitation of 140 characters. Users have limited space to express their feelings. So the sentiments in tweets are usually concise, straightforward and polarity consistent.

Emoticons. Users tend to utilize emoticons to emphasize their sentiment feelings. According to the statistics, about 8.1% tweets contain at least one emoticon (Yang and Leskovec, 2011). Since the tweets have the length limitation, the sentiments expressed in these short texts are usually consistent with the embedded emoticons, such as the word *fun* and *headache* in above examples.

In addition to the above advantages, there are also some disadvantages for measuring sentiment similarity using Twitter data. The spam tweets that caused by advertisements may add noise and bias during the similarity measurement. The short length may also bring in lower co-occurrence probability of words. Some words may not co-occur with each other when the corpus is small. These disadvantages set obstacles for measuring sentiment similarity by using Twitter data as corpus. In the experiment section, we will see if we can overcome these drawbacks and get benefit from the advantages of Twitter data.

3 The Corpus-based Sentiment Similarity Measurements

The intuition behind the corpus-based semantic similarity measuring method is that the words with similar meanings tend to co-occur in the corpus. Given the word w_i, w_j, we use the notation $P(w_i)$ to denote the occurrence counts of word w_i in the corpus $\mathcal{C}$. $P(w_i, w_j)$ denotes the co-occurrence counts of word w_i and w_j in $\mathcal{C}$. In this paper we employ the corpus-based version of the three well-known similarity measurements: Jaccard, Dice and PMI.

$$CorpusJaccard(w_i, w_j) = \frac{P(w_i,w_j)}{P(w_i)+P(w_j)-P(w_i,w_j)} \quad (2)$$

$$CorpusDice(w_i, w_j) = \frac{2 \times P(w_i, w_j)}{P(w_i) + P(w_j)} \quad (3)$$

$$CorpusPMI(w_i, w_j) = \log_2\left(\frac{\frac{P(w_i,w_j)}{N}}{\frac{P(w_i)}{N}\frac{P(w_j)}{N}}\right) \quad (4)$$

In Formula (4), N is the number of documents in the corpus $\mathcal{C}$. The above similarity measurements may have their own strengths and weaknesses. In this paper, we utilize these classical measurements to evaluate the quality of the corpus in polarity classification task.

Google is the world's largest search engine, which has indexed a huge number of Web pages. Using the extreme large indexed Web pages as corpus, Cilibrasi and Vitanyi (2007) presented a method for measuring similarity between words and phrases based on information distance and Kolmogorov complexity. The search result page counts of Google were utilized to estimate the occurrence frequencies of the words in the corpus. Suppose w_i, w_j represent the candidate words, the Normalized Google

Distance is defined as:

$$NGD(w_i, w_j) = \frac{\max\{\log P(w_i), \log P(w_j)\} - \log P(w_i, w_j)}{\log N - \min\{\log P(w_i), \log P(w_j)\}} \quad (5)$$

where $P(w_i)$ denotes page counts returned by Google using w_i as keyword; $P(w_i, w_j)$ denotes the page counts by using w_i and w_j as joint keywords; N is the number of Web pages indexed by Google. Cilibrasi and Vitanyi have validated the effectiveness of Google distance in measuring the semantic similarity between concept words.

Based on the above formulas, we compare the Twitter data with the Web and Wikipedia data as the similarity measurement corpus. Given a candidate word w, we firstly measure its sentiment similarity with a positive seed word and a negative seed word respectively in Formula (1), and the difference of sim is used to further detect the polarity of w. The above four similarity measurements serve as sim with Web, Wikipedia and Twitter data as corpus. Turney (2002) chose *excellent* and *poor* as seed words. However, using isolated seed words may cause the bias problem. Therefore, we further select two groups of seed words that are lack of sensitivity to context and form a positive seed set PS and a negative seed set NS (Turney, 2003). The Formula (1) can be rewritten as:

$$\overline{SO}(w) = \sum_{se_p \in PS} sim(w, se_p) - \sum_{se_n \in NS} sim(w, se_n) \quad (6)$$

Based on the Formula(6) and the sentiment seed words, we can measure the sentiment polarity of the given candidate words.

4 Experiment

4.1 Experiment Setup

Corpus Preparing. The Twitter corpus corresponds to the *476 million Twitter tweets* (Yang and Leskovec, 2011), which includes over 476 million Twitter posts from 20 million users, covering a 7 month period from June 1, 2009 to December 31, 2009. We filter out the non-English tweets and the spam tweets that have only few words with URLs. The tweets that contain three or more trending topics are also removed. Finally, we construct the Twitter corpus that consists of 266.8 million English tweets. For calculating page counts in Web data, the candidate words were launched to Google from February 2013 to April 2013. We also conduct the experiments on the Google Web 1T data that consists of Google n-gram counts (frequency of occurrence of each n-gram) for $1 \leq n \leq 5$ (Brants and Franz, 2006). The Web 1T data provides a nice approximation to the word co-occurrence statistics in Web pages in a predefined window size ($1 \leq n \leq 5$). For example, the 5 gram Web1T data means the co-occurrence window size is 5. The English Wikipedia dump [1] we used was extracted at the end of March 2013, which contained more than 13 million articles. We extracted the plain texts of the Wikipedia data as the training corpus for the Formula (6).

Evaluation Method. Two well-know sentiment lexicons are utilized as gold standard for polarity classification task. The statistics of Liu's sentiment lexicon (Liu et al., 2005) and MPQA subjectivity lexicon (Wilson et al., 2005) are shown in Table 1. For each word w in the lexicons, we employ the Formula (6) to calculate the word's polarity using different corpora. If $\overline{SO}(w) > 0$, the word w is classified into the positive category. Otherwise if $\overline{SO}(w) < 0$, it is classified into the negative category. The accuracy of the classification result is used to measure the quality of the corpus.

	Positive#	Negative#
Liu	2,006	4,783
MPQA	2,304	4,153

Table 1: Lexicon size

4.2 Experiment Results

Firstly, we chose the seed words *excellent* and *poor* as Turney's (2002) settings. The polarity classification accuracies are shown in Table 2.

In Table 2, Google, Web1T, Wikipedia, Twitter represent the corpora that used in the experiment; CJ, CD, CP, GD represent the Formula (2) to Formula (5) respectively. We can see from the Table 2 that the Twitter based method can achieve the best performance. The rich sentiment information and

[1] http://en.wikipedia.org/

Lexicon	Corpus	CJ	CD	CP	GD
Liu	Google	0.5116	0.5117	0.5064	0.5076
	Web1T-5gram	0.3903	0.3903	0.3897	0.3864
	Web1T-4gram	0.3771	0.3771	0.3772	0.3227
	Wikipedia	0.5280	0.5280	0.5350	0.5412
	Twitter	**0.5567**	**0.5567**	**0.5635**	**0.5635**
MPQA	Google	0.4897	0.4890	0.4891	0.4864
	Web1T-5gram	0.3843	0.3843	0.3837	0.3783
	Web1T-4gram	0.3729	0.3729	0.3714	0.3225
	Wikipedia	0.5181	0.5181	0.5380	0.5344
	Twitter	**0.5421**	**0.5421**	**0.5493**	**0.5494**

Table 2: Polarity classification accuracies using *excellent* and *poor* as seed words

Lexicon	Corpus	CJ	CD	CP	GD
Liu	Google	0.4859	0.4936	0.4884	0.5060
	Web1T-5gram	0.5785	0.5785	0.3963	0.5782
	Web1T-4gram	0.5766	0.5766	0.3872	0.5775
	Wikipedia	0.6226	0.6225	0.5957	0.6145
	Twitter	0.6678	0.6678	0.6917	0.6457
	Twitter$^+$	**0.6921**	**0.6921**	**0.7273**	**0.6599**
MPQA	Google	0.5108	0.5225	0.5735	0.5763
	Web1T-5gram	0.5737	0.5737	0.4225	0.5718
	Web1T-4gram	0.5749	0.5749	0.3329	0.4797
	Wikipedia	0.6086	0.6085	0.5773	0.5985
	Twitter	0.6431	0.6431	0.6671	0.6253
	Twitter$^+$	**0.6665**	**0.6665**	**0.7001**	**0.6383**

Table 3: Polarity classification accuracies using the seed word groups

natural window size (140 characters) have a positive impact on determining the word's polarity. The Google based method gets a lower accuracy, this may be due to the length of Web documents which can not usually guarantee the semantic consistency in the returned data. Even though two words appear in one page (returned by Google), they might not be semantically related. Furthermore, the Google based method is time-consuming, because we have to periodically send queries in order to avoid being blocked by Google. The Web1T based method gets a much worse accuracy. After detailed analysis, we find that although the small window size (4 or 5) can guarantee the semantic consistency, the short length also brings in lower co-occurrence probability. Statistics show that about 38% $\overline{SO}$ values are zero when using Web1T corpus. Due to the short length, the Twitter data also suffers from the low co-occurrence problem.

To tackle the low co-occurrence problem, the seed word sets are selected as Turney's (2003) settings. The positive word set *PS*={*good, nice, excellent, positive, fortunate, correct, superior*} and negative word set *NS* = {*bad, nasty, poor, negative, unfortunate, wrong, inferior*} for the Formula (6). These seed words have been verified to be effective in Turney's paper for polarity classification. The experiment results are shown in Table 3.

Table 3 shows that the performance of Twitter corpus is much improved since the multiple seed words alleviate the problem of low co-occurrence probability in tweets. Generally, when using the seed word groups the Twitter can achieve a much better performance than all the other corpora. The improvements are statistically significant (p-value < 0.05). We further add the emoticons ':)' and ':(' into the seed word groups, denoted by Twitter$^+$ in Table 3. The emoticons are natural sentiment labels. We can see that the performances are further improved by considering emoticons as seed words. The above experiment results have validated the effectiveness of Twitter data as a better corpus for measuring the sentiment similarity. The results also reveal the potential usefulness of Twitter corpus in semantic similarity measurement.

5 Related Work

Detecting the polarity of words is the fundamental problem for most of sentiment analysis tasks (Hatzivassiloglou and McKeown, 1997; Pang and Lee, 2007; Feldman, 2013).

Many methods have been proposed to measure the words' or short texts similarity based on large corpus (Sahami and Heilman, 2006; Yih and Meek, 2007; Gabrilovich and Markovitch, 2007). Bollegala *et al*. (2011) submitted the word to the search engine, and the related result pages were employed to represent the meaning of the original word. Mihalcea *et al*. (2006) proposed a method to measure the semantic similarity of words or short texts, considering both corpus-based and knowledge-based information. Although the previous algorithms have achieved promising results, there are no work done on evaluating the quality of different corpora.

Mohtarami *et al*. (2012; 2013a; 2013b) introduced the concept of sentiment similarity, which was considered as different from the traditional semantic similarity, and more focused on revealing the underlying sentiment relations between words. Mo-

htarami *et al*. (2013b) proposed a hidden emotional model to calculating the sentiment similarity of word pairs. However, the impact of the different corpora is not considered for this task.

Mohammad *et al*. (2013) generated word-sentiment association lexicons from Tweets with the help of hashtags and emoticons. Pak and Paroubek (2010) collected tweets with happy and sad emoticons as training corpus, and built sentiment classifier based on traditional machine learning methods. Brody and Diakopoulos (2011) showed that lengthening was strongly associated with subjectivity and sentiment in tweets. Davidov *et al*. (2010) treated 50 Twitter tags and 15 smileys as sentiment labels and a supervised sentiment classification framework was proposed to classify the tweets. The previous literatures have showed that the emoticons can be treated as natural sentiment labels of the tweets.

6 Conclusion and Future Work

The quality of corpus may affect the performance of sentiment similarity measurement. In this paper, we compare the Twitter data with the Google, Web1T and Wikipedia data in polarity classification task. The experiment results validate that when using the seed word groups the Twitter can achieve a much better performance than the other corpora and adding emoticons as seed words can further improve the performance. It is observed that the twitter corpus is a potential good source for measuring sentiment similarity between words. In future work, we intend to design new similarity measurements that can make best of the advantages of Twitter data.

Acknowledgments

This research is partially supported by General Research Fund of Hong Kong (No. 417112) and Shenzhen Fundamental Research Program (JCYJ20130401172046450). This research is also supported by the State Key Development Program for Basic Research of China (Grant No. 2011CB302200-G), State Key Program of National Natural Science of China (Grant No. 61033007), National Natural Science Foundation of China (Grant No. 61370074, 61100026), and the Fundamental Research Funds for the Central Universities (N120404007).

References

Thorsten Brants and Alex Franz. 2006. Web 1t 5-gram version 1. *Linguistic Data Consortium*, ISBN: 1-58563-397-6, Philadelphia.

Samuel Brody and Nicholas Diakopoulos. 2011. Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 562–570, Edinburgh, UK, ACL.

Danushka Bollegala, Yutaka Matsuo, and Mitsuru Ishizuka. 2011. A Web Search Engine-Based Approach to Measure Semantic Similarity between Words. *IEEE Transactions on Knowledge and Data Engineering*, 23(7): 977–990.

Rudi Cilibrasi and Paul Vitanyi. 2007. The Google Similarity Distance. *IEEE Transactions on Knowledge and Data Engineering*, 19(3): 370–383.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced Sentiment Learning Using Twitter Hashtags and Smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 241–249, Beijing, China, ACL.

Ronen Feldman. 2013. Techniques and Applications for Sentiment Analysis. *Communications of the ACM*, 56(4):82–89.

Evgeniy Gabrilovich and Shaul Markovitch. 2007. Computing Semantic Relatedness Using Wikipedia-based Explicit Semantic Analysis. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*, pages 1606–1611, Hyderabad, India.

Vasileios Hatzivassiloglou and Kathleen McKeown. 1997. Predicting the Semantic Orientation of Adjectives. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics*, pages 174–181, Madrid, Spain, ACL.

Akshay Java, Xiaodan Song, Tim Finin, and Belle L. Tseng. 2007. Why We Twitter: An Analysis of a Microblogging Community. In *Proceedings of the 9th International Workshop on Knowledge Discovery on the Web and 1st International Workshop on Social Networks Analysis*, pages 118–138, San Jose, CA, USA, Springer.

Nobuhiro Kaji and Masaru Kitsuregawa. Building Lexicon for Sentiment Analysis from Massive Collection of HTML Documents. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1075–1083, Prague, Czech Republic, ACL.

Hiroshi Kanayama and Tetsuya Nasukawa. 2006. Fully Automatic Lexicon Expansion for Domain-oriented

Sentiment Analysis. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 355–363, Sydney, Australia, ACL.

Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue B. Moon. 2010. What is Twitter, a Social Network or a News Media? In *Proceedings of the 19th International Conference on World Wide Web*, pages 591–600, Raleigh, North Carolina, USA, ACM.

Dekang Lin. 1998. Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics*, pages 768–774, Montreal, Quebec, Canada, ACL.

Bing Liu, Minqing Hu, and Junsheng Cheng. 2005. Opinion Observer: Analyzing and Comparing Opinions on the Web. In *Proceedings of the 14th international conference on World Wide Web*, pages 342–351, Chiba, Japan, ACM.

Rada Mihalcea, Courtney Corley, and Carlo Strapparava. 2006. Corpus-based and Knowledge-based Measures of Text Semantic Similarity. In *Proceedings of the 21st National Conference on Artificial Intelligence and the 18th Innovative Applications of Artificial Intelligence Conference*, pages 775–780, Boston, Massachusetts, USA, AAAI Press.

Mitra Mohtarami, Hadi Amiri, Man Lan, Thanh Phu Tran, and Chew Lim Tan. 2012. Sense Sentiment Similarity: An Analysis. In *Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence*, pages 1706–1712, Toronto, Ontario, Canada, AAAI Press.

Mitra Mohtarami, Man Lan, and Chew Lim Tan. 2013a. From Semantic to Emotional Space in Probabilistic Sense Sentiment Analysis. In *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pages 711–717, Bellevue, Washington, USA, AAAI Press.

Mitra Mohtarami, Man Lan, and Chew Lim Tan. 2013b. Probabilistic Sense Sentiment Similarity through Hidden Emotions. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 983–992, Sofia, Bulgaria, ACL.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. NRC-Canada: Building the State-of-the-Art in Sentiment Analysis of Tweets. In *Proceedings of the seventh international workshop on Semantic Evaluation Exercises*, Atlanta, Georgia, USA, ACL.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a Corpus for Sentiment Analysis and Opinion Mining. In *Proceedings of the 2010 International Conference on Language Resources and Evaluation*, pages 1320–1326, Valletta, Malta, ELRA.

Philip Resnik. 1999. Semantic Similarity in a Taxonomy: An Information based Measure and Its Application to Problems of Ambiguity in Natural Language. *Journal of Artificial Intelligence Research*, 11:95–130.

Mehran Sahami and Timothy D. Heilman. 2006. A Web-based Kernel Function for Measuring the Similarity of Short Text Snippets. In *Proceedings of the 15th international conference on World Wide Web*, pages 377–386, Edinburgh, Scotland, UK, ACM.

Bo Pang and Lillian Lee. 2007. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135.

Peter D. Turney. 2002. 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*, pages 417–424, Philadelphia, PA, USA, ACL.

Peter Turney and Michael Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transaction Information System*, 21(4): 315–346.

Leonid Velikovich, Sasha Blair-Goldensohn, Kerry Hannan, and Ryan T. McDonald. The Viability of Web-derived Polarity Lexicons. In *Proceedings of the 2010 North American Chapter of the Association of Computational Linguistics*, pp. 777–785, Los Angeles, California, USA, ACL.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. In *Proceedings of the 2005 Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 347–354, Vancouver, British Columbia, Canada, ACL.

Jaewon Yang and Jure Leskovec. 2011. Patterns of Temporal Variation in Online Media. In *Proceedings of the Forth International Conference on Web Search and Web Data Mining*, pages 177–186, Hong Kong, China, ACM.

Wen-tau Yih and Christopher Meek. 2007. Improving Similarity Measures for Short Segments of Text. In *Proceedings of the 22nd AAAI Conference on Artificial Intelligence*, pages 1489–1494, Vancouver, British Columbia, Canada, AAAI Press.