The Impact of Z_score on Twitter Sentiment Analysis

Hussam Hamdan*,,*****
*LSIS
Aix-Marseille Université CNRS
Av. Esc. Normandie Niemen,
13397 Marseille Cedex 20,
France
hussam.hamdan@lsis.org

Patrice Bellot*,**
**OpenEdition
Aix-Marseille Université CNRS
3 pl. V. Hugo, case n°86
13331 Marseille Cedex 3,
France
patrice.bellot@lsis.org

Frederic Béchet***
***LIF
Aix-Marseille Université CNRS
Avenue de Luminy
13288 Marseille Cedex 9,
France
frederic.bechet@lif.univ-mrs.fr

Abstract

Twitter has become more and more an important resource of user-generated data. Sentiment Analysis in Twitter is interesting for many applications and objectives. In this paper, we propose to exploit some features which can be useful for this task; the main contribution is the use of Z-scores as features for sentiment classification in addition to pre-polarity and POS tags features. Our experiments have been evaluated using the test data provided by SemEval 2013 and 2014. The evaluation demonstrates that Z_scores features can significantly improve the prediction performance.

1 Introduction

The interactive Web has changed the relation between the users and the web. Users have become an important source of content. They express their opinion towards different issues. These opinions are important for others who are interested in understanding users' interests such as buyers, sellers and producers.

Twitter is one of the most important platforms in which the users express their opinions. Many works have exploited this media for predicting valuable issues depending on Sentiment Analysis (SA). The authors in (Asur and Huberman 2010) predicted the box-office revenues of movies in advance of their releases using the tweets talking about them. In (Bae and Lee 2012) Sentiment Analysis has been used to study the impact of 13 twitter accounts of famous persons on their followers and also for forecasting the interesting tweets which are more probably to be reposted by the followers (Naveed, Gottron et al. 2011).

Sentiment Analysis can be done in different levels; Document level; Sentence level; Clause level or Aspect-Based level. SA in Twitter can be seen as a sentence level task, but some limitations should be considered in such sentences. The size of tweets is limited to 140 characters, informal language, emotion icons and non-standard expressions are commonly used, and many spelling errors can be found due to the absence of correctness verification.

Three different approaches can be identified in the literature of Sentiment Analysis in Twitter, the first approach is lexicon based, using specific types of lexicons to derive the polarity of a text, this approach suffers from the limited size of lexicon and requires human expertise to build manual lexicon (Joshi, Balamurali et al. 2011), in the other hand the automatic lexicons are not so efficient. The second one is machine learning approach which uses annotated texts with a given labels to learn a classification model, an early work was done on a movie review dataset (Pang, Lee et al. 2002). Both lexicon and machine learning approaches can be combined to achieve a better performance (Khuc, Shivade et al. 2012). These two approaches are used for SA task but the third one is specific for Twitter or social content, the social approach exploits social network properties and data for enhancing the accuracy of the classification (Speriosu, Sudan et al. 2011).

In this paper, we exploit machine learning algorithm with the aid of some features:

- The original Terms: the terms representing the tweet after the tokenization and stemming;

Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 636–641,
Dublin, Ireland, August 23-24, 2014.

- Pre-polarity features: the number of negative, positive and neutral words extracted from two sentiment lexicons;
- POS tags: the number of adjectives, connectors, verbs, nouns, adverbs in the tweet;
- Z-score: The numbers of terms having Z-score value more than three for each class positive, negative and neutral.

We extended the original terms with these last features. We also constructed a dictionary for the abbreviations and the slang words used in Twitter in order to overcome the ambiguity of the tweets. We tested the performance of every possible combination of these features.

The rest of this paper is organized as follows. Section 2 outlines previous work that focused on sentiment analysis in Twitter. Section 3 presents the Z_score features and the others which we used for training a classifier. Our experiments are described in section 4, conclusion and future work is presented in section 5.

2 Related Works

We can identify three main approaches for sentiment analysis in Twitter. The lexicon based approaches which depend on sentiment lexicons containing positive, negative and neutral words or expressions; they calculate the polarity according to the number of common opinionated words between the lexicons and the text. Many dictionaries have been created manually such as ANEW (Affective Norms for English Words) or automatically such as SentiWordNet (Baccianella, Esuli et al. 2010). Four lexicon dictionaries were used to overcome the lack of words in each one (Joshi, Balamurali et al. 2011; Mukherjee, Malu et al. 2012). Automatically construction of a Twitter lexicon was implemented by (Khuc, Shivade et al. 2012).

Machine learning approaches were employed from annotated tweets by using Naive Bayes, Maximum Entropy *MaxEnt* and Support Vector Machines (SVM). The authors (Go, Bhayani et al. 2009) reported that SVM outperforms other classifiers. They tried a unigram and a bigram model in conjunction with parts-of-speech (POS) features; they noted that the unigram model outperforms all other models when using SVM and that POS features decrease the quality of results. The authors in (Kouloumpis, Wilson et al. 2011) found that N-gram with lexicon features and micro-blogging features are useful but POS features are not. In contrast, in (Pak and Paroubek 2010) they reported that POS and bigrams both help. In (Barbosa and Feng 2010) the authors proposed the use of syntax features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS tags, in (Agarwal, Xie et al. 2011) this approach was extended by using real valued prior polarity and by combining prior polarity with POS. Authors in (Saif, He et al. 2012) proposed to use the semantic features, therefore they extracted the named entities in the tweets. Authors in (Hamdan, Béchet et al. 2013) used the concepts extracted from DBpedia and the adjectives from WordNet, they reported that the DBpedia concepts are useful with Naïve-Bayes classifier but less useful with SVM.

The third main approach takes into account the influence of users on their followers and the relation between the users and the tweets they wrote. It assumes that using the Twitter follower graph might improve the polarity classification. In (Speriosu, Sudan et al. 2011) they demonstrated that using label propagation with Twitter follower graph improves the polarity classification. In (Tan, Lee et al. 2011) they employed social relation for user-level sentiment analysis. In (Hu, Tang et al. 2013) a Sociological Approach to handling the Noisy and short Text (SANT) for supervised sentiment classification is used; they reported that social theories such as Sentiment Consistency and Emotional Contagion could be helpful for sentiment analysis.

3 Feature Selection

We used different types of features in order to improve the accuracy of sentiment classification.

- Bag of words (Terms)

The most commonly used features in text analysis are the bag of words which represent a text as unordered set of words or terms. It assumes that words are independent from each other and also disregards their order of appearance. We stemmed the words using Porter Stemmer and used them as a baseline features.

- Z_score Features (Z)

We suggest using a new type of features for Sentiment Analysis, Z_score can distinguish the importance of each term in each class. We compute the number of terms having Z_score more than three for each class over each tweet. We assume that the term frequencies follow the multinomial distribution. Thus, Z_score can be seen as a standardization of the term. We compute the

Z_score for each term t_i in a class Cj (t_{ij}) by calculating its term relative frequency tfr_{ij} in a particular class C_j, as well as the mean ($mean_i$) which is the term probability over the whole corpus multiplied by n_j the number of terms in the class C_j, and standard deviation (sd_i) of term t_i according to the underlying corpus (see Eq. (1,2)).

$$Z_{score(t_{ij})} = \frac{tfr_{ij} - mean_i}{sdi} \quad \text{Eq. (1)}$$

$$Z_{score(t_{ij})} = \frac{tfr_{ij} - n_j * p(ti)}{\sqrt{nj * p(ti) * (1 - p(ti))}} \quad \text{Eq. (2)}$$

The term which has salient frequency in a class in compassion to others will have a salient Z_score. Z_score was exploited for SA by (Zubaryeva and Savoy 2010) , they choose a threshold (>2) for selecting the number of terms having Z_score more than the threshold, then they used a logistic regression for combining these scores. We use Z_scores as added features for classification because the tweet is too short, therefore many tweets does not have any words with salient Z_score. The three following figures 1,2,3 show the distribution of Z_score over each class, we remark that the majority of terms has Z_score between -1.5 and 2.5 in each class and the rest are either vey frequent (>2.5) or very rare (<-1.5). It should indicate that negative value means that the term is not frequent in this class in comparison with its frequencies in other classes. Table1 demonstrates the first ten terms having the highest Z_scores in each class. We have tested to use different values for the threshold, the best results was obtained when the threshold is 3.

positive	Z_score	negative	Z_score	Neutral	Z_score
Love	14.31	Not	13.99	Httpbit	6.44
Good	14.01	Fuck	12.97	Httpfb	4.56
Happy	12.30	Don't	10.97	Httpbnd	3.78
Great	11.10	Shit	8.99	Intern	3.58
Excite	10.35	Bad	8.40	Nov	3.45
Best	9.24	Hate	8.29	Httpdlvr	3.40
Thank	9.21	Sad	8.28	Open	3.30
Hope	8.24	Sorry	8.11	Live	3.28
Cant	8.10	Cancel	7.53	Cloud	3.28
Wait	8.05	stupid	6.83	begin	3.17

Table1. The first ten terms having the highest Z_score in each class

- Sentiment Lexicon Features (POL)

We used two sentiment lexicons, MPQA Subjectivity Lexicon(Wilson, Wiebe et al. 2005) and Bing Liu's Opinion Lexicon which is created by (Hu and Liu 2004) and augmented in many latter works. We extract the number of positive, negative and neutral words in tweets according to these lexicons. Bing Liu's lexicon only contains negative and positive annotation but Subjectivity contains negative, positive and neutral.

- Part Of Speech (POS)

We annotate each word in the tweet by its POS tag, and then we compute the number of adjectives, verbs, nouns, adverbs and connectors in each tweet.

4 Evaluation

4.1 Data collection

We used the data set provided in SemEval 2013 and 2014 for subtask B of sentiment analysis in Twitter(Rosenthal, Ritter et al. 2014) (Wilson, Kozareva et al. 2013). The participants were provided with training tweets annotated as positive, negative or neutral. We downloaded these tweets using a given script. Among 9646 tweets, we could only download 8498 of them because of protected profiles and deleted tweets. Then, we used the development set containing 1654 tweets for evaluating our methods. We combined the development set with training set and built a new model which predicted the labels of the test set 2013 and 2014.

4.2 Experiments

Official Results

The results of our system submitted for SemEval evaluation gave 46.38%, 52.02% for test set 2013 and 2014 respectively. It should mention that these results are not correct because of a software bug discovered after the submission deadline, therefore the correct results is demonstrated as non-official results. In fact the previous results are the output of our classifier which is trained by all the features in section 3, but because of index shifting error the test set was represented by all the features except the terms.

Non-official Results

We have done various experiments using the features presented in Section 3 with Multinomial Naïve-Bayes model. We firstly constructed feature vector of tweet terms which gave 49%, 46% for test set 2013, 2014 respectively. Then, we augmented this original vector by the Z_score

features which improve the performance by 6.5% and 10.9%, then by pre-polarity features which also improve the f-measure by 4%, 6%, but the extending with POS tags decreases the f-measure. We also test all combinations with these previous features, Table2 demonstrates the results of each combination, we remark that POS tags are not useful over all the experiments, the best result is obtained by combining Z_score and pre-polarity features. We find that Z_score features improve significantly the f-measure and they are better than pre-polarity features.

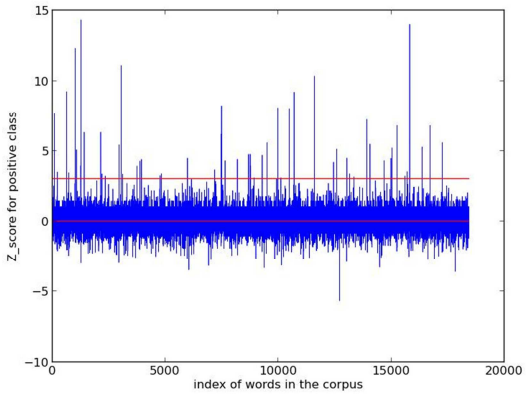

Figure 1 Z_score distribution in positive class

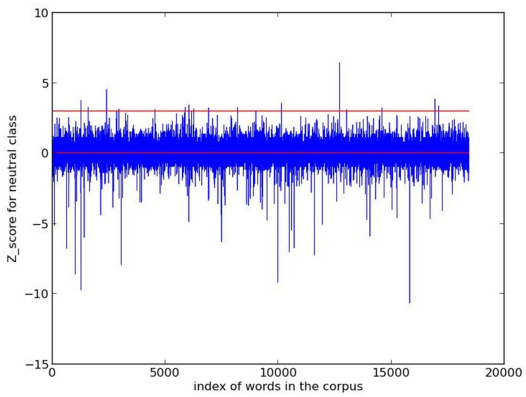

Figure 2 Z_score distribution in neutral class

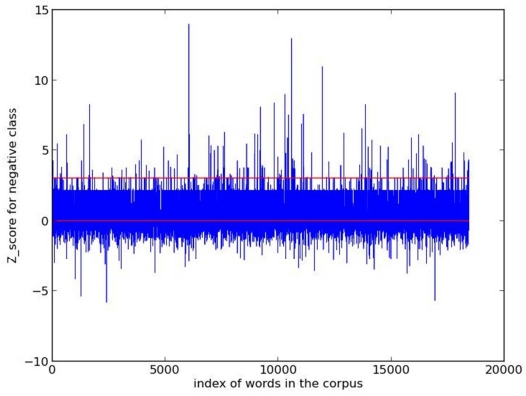

Figure 3 Z_score distribution in negative class

Features	F-measure	
	2013	**2014**
Terms	49.42	46.31
Terms+Z	55.90	57.28
Terms+POS	43.45	41.14
Terms+POL	53.53	52.73
Terms+Z+POS	52.59	54.43
Terms+Z+POL	**58.34**	**59.38**
Terms+POS+POL	48.42	50.03
Terms+Z+POS+POL	55.35	58.58

Table 2. Average f-measures for positive and negative classes of SemEval2013 and 2014 test sets.

We repeated all previous experiments after using a twitter dictionary where we extend the tweet by the expressions related to each emotion icons or abbreviations in tweets. The results in Table3 demonstrate that using that dictionary improves the f-measure over all the experiments, the best results obtained also by combining Z_scores and pre-polarity features.

Features	F-measure	
	2013	**2014**
Terms	50.15	48.56
Terms+Z	57.17	58.37
Terms+POS	44.07	42.64
Terms+POL	54.72	54.53
Terms+Z+POS	53.20	56.47
Terms+Z+POL	**59.66**	**61.07**
Terms+POS+POL	48.97	51.90
Terms+Z+POS+POL	55.83	60.22

Table 3. Average f-measures for positive and negative classes of SemEval2013 and 2014 test sets after using a twitter dictionary.

5 Conclusion

In this paper we tested the impact of using Twitter Dictionary, Sentiment Lexicons, Z_score features and POS tags for the sentiment classification of tweets. We extended the feature vector of tweets by all these features; we have proposed new type of features Z_score and demonstrated that they can improve the performance.

We think that Z_score can be used in different ways for improving the Sentiment Analysis, we are going to test it in another type of corpus and using other methods in order to combine these features.

Reference

Apoorv Agarwal,Boyi Xie,Ilia Vovsha,Owen Rambow and Rebecca Passonneau (2011). Sentiment analysis of Twitter data. Proceedings of the Workshop on Languages

in Social Media. Portland, Oregon, Association for Computational Linguistics**:** 30-38.

Sitaram Asur and Bernardo A. Huberman (2010). Predicting the Future with Social Media. Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01, IEEE Computer Society**:** 492-499.

Stefano Baccianella,Andrea Esuli and Fabrizio Sebastiani (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh Conference on International Language Resources and Evaluation (LREC'10), European Language Resources Association (ELRA).

Younggue Bae and Hongchul Lee (2012). "Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers." J. Am. Soc. Inf. Sci. Technol. **63**(12): 2521-2535.

Luciano Barbosa and Junlan Feng (2010). Robust sentiment detection on Twitter from biased and noisy data. Proceedings of the 23rd International Conference on Computational Linguistics: Posters. Beijing, China, Association for Computational Linguistics**:** 36-44.

Alec Go,Richa Bhayani and Lei Huang Twitter Sentiment Classification using Distant Supervision.

Hussam Hamdan,Frederic Béchet and Patrice Bellot (2013). Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), Atlanta, Georgia, USA.

Minqing Hu and Bing Liu (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. Seattle, WA, USA, ACM**:** 168-177.

Xia Hu,Lei Tang,Jiliang Tang and Huan Liu (2013). Exploiting social relations for sentiment analysis in microblogging. Proceedings of the sixth ACM international conference on Web search and data mining. Rome, Italy, ACM**:** 537-546.

Aditya Joshi,A. R. Balamurali,Pushpak Bhattacharyya and Rajat Mohanty (2011). C-Feel-It: a sentiment analyzer for micro-blogs. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. Portland, Oregon, Association for Computational Linguistics**:** 127-132.

Vinh Ngoc Khuc,Chaitanya Shivade,Rajiv Ramnath and Jay Ramanathan (2012). Towards building large-scale distributed systems for twitter sentiment analysis. Proceedings of the 27th Annual ACM Symposium on Applied Computing. Trento, Italy, ACM**:** 459-464.

E. Kouloumpis,T. Wilson and J. Moore (2011). Twitter Sentiment Analysis: The Good the Bad and the OMG! Fifth International AAAI Conference on Weblogs and Social Media.

Subhabrata Mukherjee,Akshat Malu,Balamurali A.R. and Pushpak Bhattacharyya (2012). TwiSent: a multistage system for analyzing sentiment in twitter. Proceedings of the 21st ACM international conference on Information and knowledge management. Maui, Hawaii, USA, ACM**:** 2531-2534.

Nasir Naveed,Thomas Gottron,J\'Er\^Ome Kunegis and Arifah Che Alhadi (2011). Bad News Travels Fast: A Content-based Analysis of Interestingness on Twitter. Proc. Web Science Conf.

Alexander Pak and Patrick Paroubek (2010). Twitter as a Corpus for Sentiment Analysis and Opinion Mining. Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10), Valletta, Malta, European Language Resources Association (ELRA).

Bo Pang,Lillian Lee and Shivakumar Vaithyanathan (2002). Thumbs up?: sentiment classification using machine learning techniques. Proceedings of the ACL-02 conference on Empirical methods in natural language processing - Volume 10, Association for Computational Linguistics**:** 79-86.

Sara Rosenthal,Alan Ritter,Veselin Stoyanov and Preslav Nakov (2014). "SemEval-2014 Task 9: Sentiment Analysis in Twitter." In Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval'14).August 23-24, Dublin, Ireland.

Hassan Saif,Yulan He and Harith Alani (2012). Semantic sentiment analysis of twitter. Proceedings of the 11th international conference on The Semantic Web - Volume Part I. Boston, MA, Springer-Verlag**:** 508-524.

Michael Speriosu,Nikita Sudan,Sid Upadhyay and Jason Baldridge (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. Proceedings of the First Workshop on Unsupervised Learning in NLP. Edinburgh, Scotland, Association for Computational Linguistics**:** 53-63.

Chenhao Tan,Lillian Lee,Jie Tang,Long Jiang,Ming Zhou and Ping Li (2011). User-level

sentiment analysis incorporating social networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, USA, ACM**:** 1397-1405.

Theresa Wilson,Zornitsa Kozareva,Preslav Nakov,Alan Ritter,Sara Rosenthal and Veselin Stoyanov (2013). "SemEval-2013 Task 2: Sentiment Analysis in Twitter." Proceedings of the 7th International Workshop on Semantic Evaluation. Association for Computational Linguistics.

Theresa Wilson,Janyce Wiebe and Paul Hoffmann (2005). Recognizing contextual polarity in phrase-level sentiment analysis. Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Vancouver, British Columbia, Canada, Association for Computational Linguistics**:** 347-354.

Olena Zubaryeva and Jacques Savoy (2010). "Opinion Detection by Combining Machine Learning & Linguistic Tools." In Proceedings of the 8th NTCIR, Workshop Meeting on Evaluation of Information Access Technologies: InformationRetrieval, Question Answering and Cross-Lingual Information Access.