Sentiment Lexicon-Based Features for Sentiment Analysis in Short Text

Hussam Hamdan,[1,2,3] Patrice Bellot,[1,2] Frederic Bechet[3]

[1] Aix Marseille Université, CNRS, ENSAM,
[2] Université de Toulon, LSIS UMR 7296,13397, Marseille, France
hussam.hamdan,patrice.bellot@lsis.org
[3] Aix Marseille Université, CNRS, LIF UMR 7279, Marseille, France
frederic.bechet@lif.univ-mrs.fr

Abstract. Sentiment lexicon-based features have proved their performance in recent work concerning sentiment analysis in Twitter. Automatic constructed lexicon features seem to be enough influential to attract the attention. In this paper, we propose a new metric to estimate the word polarity score, called natural entropy (ne), in order to construct a new sentiment lexicon based on Sentiment140 corpus. We derive six features from the new lexicon and show that (ne) metric outperforms the PMI metric which has been used for the same purpose. For evaluation, we build a state-of-the-art system for sentiment analysis in short text using a supervised classifier trained on several groups of features including n-gram, sentiment lexicons, negation, Z score and semantic features. This system has been one of the best systems in both tasks of SemEval-2015: Sentiment Analysis in Twitter and Aspect-Based Sentiment Analysis. We investigate the impact of the lexicon-based features extracted from existing manual and automatic constructed lexicons on the system performance and also the impact of the proposed metric (ne).

1 Introduction

The interactive Web has changed the relation between the users and the web. Users have become an important source of content. This content includes users' opinions about events, products, or people. The availability of such information has attracted growing attention from those who want to understand the opinions and preferences of individuals which may be useful in various domains.

Sentiment Analysis (SA) has become more and more interesting since the year 2000 [1], many techniques in Natural Language Processing have been used to understand the expressed sentiment on an entity. The basic task in sentiment analysis is the polarity classification which determines the polarity of a given text, i.e. whether the expressed opinion is positive, negative or neutral. This analysis can be done at different levels of granularity: Document Level, Sentence Level or Aspect Level.

Early work in that area includes [2] and [3] applied different methods for detecting the polarity; they proposed unsupervised, semi-supervised and supervised

pp. 217–226; rec. 2015-02-01; acc. 2015-02-27

methods. The document representation is a critical component in SA, several publications have focused on the term weighting, others on the feature extraction and selection.
In this paper, we focus on the impact of automatic constructed lexicons on the supervised sentiment classification in short text. Therefore, we build a new automatic sentiment lexicon using new metric called natural entropy (ne) which has recently been proposed for supervised term weighting [4], this lexicon is built from sentiment140 corpus [5] which contains 1.6 millions automatically collected tweets classified as positive or negative depending on the involved emoticons. After getting this new lexicon, six features have been extracted and have replaced the six features extracted from sentiment140 lexicon, which is built from the same sentiment140 corpus but using PMI (pair-wise mutual information) metric instead. The new (ne) lexicon features outperform those extracted from the original PMI sentiment140 lexicon. For evaluation purpose, we use two known data set, the first extracted from twitter, the second from restaurant reviews. We build two state-of-the-art systems using a Logistic Regression classifier with different types of features including word n-gram, twitter dictionary, Z score, sentiment lexicon and semantic features, we also adapts a weighting schema for tuning the parameters of our classifiers. These two systems are among the best systems participating in SemEval 2015 task of Sentiment Analysis in Twitter and Aspect-based Sentiment Analysis. We investigate the effect of sentiment lexicon features in the two classifiers, as these features play an import role, we have been enough motivated to propose the using of (ne) metric.
The rest of this paper is organized as follows. Section 2 outlines existing work in sentence level sentiment analysis. Section 3 describes the data and resources that have been used. The features we used for representing the document are presented in Section 4. Our experiments are described in section 5, and future work is presented in Section 6.

2 Related Work

Two main approaches for sentiment analysis can be identified. The lexicon based approach which depends on sentiment lexicons containing positive, negative and neutral words or expressions; the polarity is computed according to the number of common opinionated words between the lexicons and the text. Many lexicons have been created manually such as MPQA Subjectivity Lexicon [6] or automatically such as SentiWordNet [7].
The second one is the Machine Learning approach which adapts different classifiers and features. Naive Bayes, Maximum Entropy MaxEnt and Support Vector Machines (SVM) were adapted in [5], the authors reported that SVM outperforms other classifiers, they tried a unigram and a bigram model in conjunction with part-of-speech (POS) features; they noted that the unigram model outperforms all other models when using SVM and that POS features decrease the results. The Authors in [8] found that n-gram with lexicon features and microbloging features are useful but POS features are not. In contrast, in [9] the

authors reported that POS and bigrams both help. In [10] the authors proposed the use of specific features of tweets like retweet, hashtags, link, punctuation and exclamation marks in conjunction with features like prior polarity of words and POS tags. Authors in [11] 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. Many features were used with SVM including the lexicon-based features in [12] which seem to get the most gain in performance. Other work has also proved the importance of lexicon-based features with logistic regression classifier [13, 14].

Some sentiment lexicons are manually constructed. A label (positive, negative) or a polarity score is assigned by human annotators. While automatically constructed ones assign a score indicating the association with the positive or negative sentiment, this score is computing with the aid of an annotated corpus or a linguistic resource. These scores make us capable of ranking the terms according to their association to the sentiment. Authors in [2] estimated the sentiment orientation (SO) of the extracted phrases using the Pointwise Mutual Information (PMI). The sentiment orientation of a phrase is computed based on its association with the positive reference word ”excellent” and the negative reference word ”poor”. Authors in [15] used SO to compute the sentiment orientation of a given word. In [16] Authors collected a set of 775,000 tweets to generate a large word-sentiment association lexicon; a tweet was considered positive if it has one of 32 positive hashtagged seed words, and negative if it has one of 36 negative hash-tagged seed words; the association score for a term was calculated using SO. Authors in [12] used similar method on the sentiment140 corpus [5], a collection of 1.6 million tweets that contains positive and negative emoticons; the tweets are labeled positive or negative according to the emoticons.

3 Data and Resources

3.1 Training and Testing Data

We have used two data sets, the first one is extracted from Twitter which has been provided in SemEval 2013 for subtask B of sentiment analysis in Twitter [17]. The participants have been provided with training tweets annotated positive, negative or neutral. We downloaded these tweets using the given script. We obtained 9646 tweets, the whole training data set is used for training, the provided development set containing 1654 tweets is used for tuning the machine learner. The test data set provided in SemEval-2015 containing about 2390 tweets [18] is used for evaluating our system.

The second data set is extracted from restaurant reviews, provided by SemEval 2015 ABSA organizers [19] where each review is composed of several sentences and each sentence may contain several Opinion Target Expression OTE which we want to detect their polarities. Table 1 shows the distribution of each label in each data set.

Table 1. Sentiment labels distribution in the training, testing and development data sets in Twitter and Restaurant.

Data	All	Positive	Negative	Neutral
Twitter				
train	9684	3640	1458	4586
dev	1654	739	340	575
test	2390	1038	365	987
Restaurant				
train	1655	1198	403	53
test	845	454	346	45

3.2 Sentiment Lexicons

In this section, we describe the manual and automatic constructed sentiment lexicons which have been used for realizing our system, and also explain the new lexicon we have constructed using the natural entropy metric.

Manual Constructed Sentiment Lexicons:

1- MPQA Subjectivity Lexicon: MPQA (Multi-Perspective Question Answering) Subjectivity Lexicon is maintained by Theresa Wilson, Janyce Wiebe, and Paul Hoffmann [6], a lexicon of over 8,000 subjectivity single-word clues, each clue is classified as positive or negative.

2- Bing Liu Lexicon: A list of positive and negative opinion words or sentiment words for English (around 6800 words). This list was compiled over many years starting from this paper [20].

Automatic Constructed Sentiment Lexicons:

1- NRC Hashtag Sentiment Lexicon: NRC Hashtag Sentiment Lexicon [16] contains tweet terms with scores, positive score indicates association with positive sentiment, whereas negative score indicates association with negative sentiment. It has entries for 54,129 unigrams and 316,531 bigrams; the scores are computed using PMI over a corpus of tweets.

2- Sentiment140 Lexicon: Sentiment140 Lexicon [12] contains tweet terms with scores, Sentiment140 has entries for 62,468 unigrams, 677,698 bigrams. the scores are computed using PMI over sentiment140 tweet corpus .

3- SentiWordNet: SentiWordNet [7] is the result of automatically annotating all WORDNET synsets according to their degrees of positivity, negativity, and neutrality.

4- Our Lexicon: PMI metric has been widely used to compute the semantic orientation (SO) of words in order to construct the automatic lexicons. Sentiment140 lexicon is constructed using SO but Sentiment140 corpus is a balanced corpus, it contains the same number of positive and negative tweets. Therefore, SO can be rewritten as following:

$$SO(w) = PMI(w,+) - PMI(w,-) = log(\frac{p(w,+)}{p(w).p(+)}) - log(\frac{p(w,-)}{p(w).p(-)})$$

As $p(+) = p(-)$=0.5 in the balanced corpus:

$$So(w) = 1 + log(p(+|w)) - 1 - log(p(-|w)) = log(a/c)$$

where + stands for the positive class, - stands for negative class, a is the number of documents containing the word w in the positive class, c is the number of documents containing the word w in the negative class. Thus, the SO is positive if a>c else it is negative. We should note that the probability of the classes does not affect the final SO score, therefore we propose another metric which depends on the distribution of the word over the classes which seems more relevant in the balanced corpus. We constructed a lexicon from sentiment140 corpus [5] a collection of 1.6 million tweets that contain positive and negative emoticons, we calculated Natural Entropy score for each term in this manner:

$$ne(w) = 1 - (-(p(+|w).log(p(+|w)) - p(-|w).log(p(-|w)))) \qquad (1)$$

where p(+| w): The probability of the positive class given the word w, p(−| w): The probability of the negative class given the word w. The more uneven the distribution of documents where a term occurs, the larger the Natural Entropy (ne) of this term is. Thus, the entropy of the term can express the uncertainty of the classes given the term. One minus this degree of uncertainty boosts the terms that unevenly distributed between the two classes [4]. ne score is always between 0 and 1, and it assigns a high score for the words unevenly distributed over the classes, but it cannot discriminate the positive words from the negative ones. therefore, we have used the a and c for discriminating the positive words from the negative ones; if a>c then the word is considered positive else it is considered negative.

3.3 Twitter dictionary

We constructed a dictionary for the abbreviations and the slang words used in Twitter in order to overcome the ambiguity of these terms. This dictionary maps certain twitter expressions and emotion icons to their meaning or their corresponding sentiment (e.g. gr8 replaced by great, :) replaced by very-happy).

4 Feature Extraction

Text representation for sentiment analysis can be enriched by many types of features. Adding syntactic and semantic features may result in important improvement on the system performance. In this section, we present several groups of features which have proved their performance in many experiments in SemEval 2015.

Word n-grams

unigram and bigram are extracted for each word in text without any stemming or stop-word removing, all terms with occurrence less than 3 are removed from the feature space.

Negation features

The rule-based algorithm presented in Christopher Potts' Sentiment Symposium Tutorial is implemented. This algorithm appends a negation suffix to all words that appear within a negation scope which is determined by a negation key and a punctuation. All these words have been added to the feature space.

Twitter Dictionary

All terms presented in the text and in the twitter dictionary (presented in 3.3) are mapped to their corresponding terms in the dictionary and added to the feature space.

Sentiment Lexicons

The system extracts four features from the manual constructed lexicons and six features from the automatic ones. For each sentence the number of positive words, the number of negative ones, the number of positive words divided by the number of negative ones and the polarity of the last word are extracted from manual constructed lexicons. In addition to the sum of the positive scores and the sum of the negative scores from the automatic constructed lexicons.

Z Score

Z score can distinguish the importance of each term in each class, their performances have been proved in [21]. We assume as in the mentioned work that the term frequencies are following a multi-nomial distribution. Thus, Z score can be seen as a standardization of the term frequency using multi-nomial distribution. We compute the Z score for each term ti in a class C_j (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 the number of terms in the class C_j, and standard deviation (sd_i) of term ti according to the underlying corpus. Like in [11] we tested different thresholds for choosing the words which have higher Z score.

$$Zscore(ti) = \frac{tfr_{ij} - mean_i}{sd_i} \tag{2}$$

Thus, we added the number of words having Z score higher than the threshold in each class positive,negative and neutral, the two classes which have the maximum number and minimum number of words having Z score higher than the threshold. These 5 features have been added to the feature space.

Semantic Features

The semantic representation of a text may bring some important hidden information, which may result in a better text representation and a better classification system.

-Brown dictionary features Each word in the text is mapped to its cluster in Brown, 1000 features are added to feature space where each feature represents

the number of words in the text mapped to each cluster. The 1000 clusters is provided in Twitter Word Clusters of CMU ARK group. 1000 clusters were constructed from approximately 56 million tweets.
-Topic Features Latent dirichlet association or topic modeling is used to extract 10 features. Lda-c is configured with 10 topics and the training data is used for training the model, then for each sentence in the test set, the trained model estimates the number of words assigned to each topic.
-Semantic Role Labeling Features Authors in [22] encode semantic role labeling features in SVM classifier. Our system also extract two types of features, the names: the whole term which represents an argument of the predicate and the tags: the type of each argument in the text (A0 represents the subject of predicate, A1 the object, AM-TMP the time, AM-ADV the situation, AM-loc the location). These encodings are defined by the tool which we used (Senna). We think that the predicate arguments can constitute a multi-word expression which may be helpful in Sentiment Classification.

5 Experiments

5.1 Experiment Setup

We used L1-regularized Logistic regression classifier implemented in LibLinear [23], this classifier has given good results in recent work [24] [14]. We learned two classifiers one from twitter data set using all features of Section 4 with the three polarities (positive, negative, and neutral) as labels and the second from restaurant review data set using only the following features (word n-gram, negation, lexicon-based, Z score, Brown cluster). A weighting schema is adapted for each class, we use the weighting option $-w_i$ which enables a use of different cost parameter C for different classes. Since the training data is unbalanced, this weighting schema adjusts the probability of each label. Thus, we tuned the classifier in adjusting the cost parameter C of Logistic Regression, weight wpos of positive class and weight Wneg of negative class. We used the twitter development set and 10% from the training data of restaurants for tuning the three parameters, all combinations of C in range 0.1 to to 4 by step 0.1, w_{pos} in range 1 to 8 by step 0.1, w_{neg} in range 1 to 8 by step 0.1 are tested. The combination C=0.2, w_{pos}=5.2, w_{neg}=4.2 has given the best F1 score on the development set of Twitter data set and the combination C=0.3, w_{pos}=1.2, w_{neg}=1.9 has been chosen for Restaurant set.

5.2 Results

The evaluation score used for twitter data set is the averaged F1-score of the positive and negative classes as proposed by the task organizers [17] but the averaged F1-score of all classes for restaurant review. Table 2 shows the results of our experiments after removing one lexicon features at a time for the two test sets besides to the experiment which evaluates the effect of using our sentiment lexicon which based on (ne) metric instead of sentiment140 lexicon which based on PMI metric. Note that using the lexicon features provides a gain of 3.31%, 4.50% for the twitter and restaurant test sets respectively. The manual lexicon features provide a gain of 0.98%, 0.47%, the automatic lexicon ones provide 1.13%, 1.06% which seem to be more influential than the manual ones. The results after removing each lexicon features shown in tabel 2. Note that some lexicon features decrease the performance in restaurant review such as Big-Liu, sentiment 140 and SentiwordNet but all lexicon features are influential in twitter set, the MPQA and SentiwordNet are the less influential. The last line of table 2 shows the results after removing the sentiment140 lexicon features but adding our lexicon features instead. Our features improve the performance by 0.62%, 0.48% on the two data set. This extrinsic evaluation is an indicator that using (ne) metric can be more efficient than using PMI for building a sentiment lexicon.

Table 2. The F-scores obtained on the Twitter and Restaurant test sets, Allfeatures run exploits all proposed features (see 5.1), all-lexicons run removes the lexicons features from the first run (the whole feature space), all-automatic run removes the automatic lexicon features, all-manual, all-MPQA, all-BingLiu, NRC-Hashtag, all-Sentiment140, all-SentiWordNet remove the manual lexicons, MPQA lexicon, Big LIU, NRC Sentiment140, and SentiWordNet respectively from the whole feature space, the last run removes the features extracted from Sentiment140 but adds those extracted from our new lexicon instead.

	Tweet Test	**Rest Test**
Allfeatures	64.27	75.50
all-lexicons	60.96	71.00
all-automatic	63.14	74.44
all-manual	63.29	75.03
all-MPQA	64.11	75.27
all-Bing Liu	63.69	76.33
all-NRC Hashtag	63.90	74.67
all-Sentiment140	63.91	75.62
all-SentiWordNet	64.08	75.60
all-Sentiment140+Our lexicon	64.89	75.98

6 Conclusion and future work

We built two state-of-the-art classifiers for sentiment analysis in short text. One for Twitter data and other for restaurant reviews. We study the impact of lexicon-based features on the performance. We also constructed our own sentiment lexicon using new metric called natural entropy (ne) which boosts the terms that unevenly distributed among the classes. This new lexicon features seem to improve the results more than the features extracted from the same

lexicon but using PMI metric.
As the sentiment lexicon-based features have proved their performance, future work will focus on the automatic lexicon construction on testing several metrics like Z score and KL-Divergence which we think promising in measuring the association between terms and sentiment labels.

7 Acknowledgment

This work was supported by Région PACA, Agoraweb projet.

References

1. Bing Liu. *Sentiment Analysis and Opinion Mining.* Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
2. Peter D. Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 417–424. Association for Computational Linguistics, 2002.
3. Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing - Volume 10*, EMNLP '02, pages 79–86. Association for Computational Linguistics, 2002.
4. Haibing Wu and Xiaodong Gu. Reducing over-weighting in supervised term weighting for sentiment analysis. In *COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland*, pages 1322–1330, 2014.
5. Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. pages 1–6, 2009.
6. Theresa Wilson, Paul Hoffmann, Swapna Somasundaran, Jason Kessler, Janyce Wiebe, Yejin Choi, Claire Cardie, Ellen Riloff, and Siddharth Patwardhan. OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35. Association for Computational Linguistics, 2005.
7. Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *in Proc. of LREC*, 2010.
8. Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. Twitter sentiment analysis: The good the bad and the OMG! In Lada A. Adamic, Ricardo A. Baeza-Yates, and Scott Counts, editors, *ICWSM*. The AAAI Press, 2011.
9. Er Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *In Proceedings of the Seventh Conference on International Language Resources and Evaluation*, 2010.
10. Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, COLING '10, pages 36–44. Association for Computational Linguistics, 2010.

11. Hussam Hamdan, FrÃ©dÃ©ric Bechet, and Patrice Bellot. Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. In *International Workshop on Semantic Evaluation SemEval-2013 (NAACL Workshop)*, 2013-04-29.
12. Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. NRCCanada: Building the state-of-the-art in sentiment analysis of tweets. In *In Proceedings of the International Workshop on Semantic Evaluation, SemEval '13*, 2013.
13. Yasuhide Miura, Shigeyuki Sakaki, Keigo Hattori, and Tomoko Ohkuma. TeamX: A sentiment analyzer with enhanced lexicon mapping and weighting scheme for unbalanced data. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 628–632. Association for Computational Linguistics and Dublin City University, 2014-08.
14. Hussam Hamdan, Patrice Bellot, and Frederic Bechet. lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015.
15. Peter D. Turney and Michael L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. 21(4):315–346, 2003-10.
16. Saif Mohammad. #emotional tweets. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 246–255. Association for Computational Linguistics, 2012-06-07.
17. Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. SemEval-2013 task 2: Sentiment analysis in twitter. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. Association for Computational Linguistics, 2013.
18. Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M. Mohammad, Alan Ritter, and Veselin Stoyanov. SemEval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015. Association for Computational Linguistics, 2015-06.
19. Maria Pontiki, Dimitrios Galanis, Haris Papageogiou, Suresh Manandhar, and Ion Androutsopoulos. SemEval-2015 task 12: Aspect based sentiment analysis. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015.
20. Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '04, pages 168–177. ACM, 2004.
21. Hussam Hamdan, Patrice Bellot, and Frederic Bechet. The impact of z_score on twitter sentiment analysis. In *In Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, page 636, 2014.
22. Josef Ruppenhofer and Ines Rehbein. Semantic frames as an anchor representation for sentiment analysis. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, pages 104–109. Association for Computational Linguistics, 2012.
23. Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. 9:1871–1874, 2008.
24. Hussam Hamdan, Patrice Bellot, and Frederic Bechet. Lsislif: CRF and logistic regression for opinion target extraction and sentiment polarity analysis. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, 2015.