Lsislif: Feature Extraction and Label Weighting for Sentiment Analysis in Twitter

Hussam Hamdan
Aix-Marseille University
hussam.hamdan@lsis.org

Patrice Bellot
Aix-Marseille University
patrice.bellot@lsis.org

Frederic Bechet
Aix-Marseille University
frederic.bechet@lif.univ-mrs.fr

Abstract

This paper describes our sentiment analysis systems which have been built for SemEval-2015 Task 10 Subtask B and E. For subtask B, a Logistic Regression classifier has been trained after extracting several groups of features including lexical, syntactic, lexicon-based, Z score and semantic features. A weighting schema has been adapted for positive and negative labels in order to take into account the unbalanced distribution of tweets between the positive and negative classes. This system is ranked third over 40 participants, it achieves average F1 64.27 on Twitter data set 2015 just 0.57% less than the first system. We also present our participation in Subtask E in which our system has got the second rank with Kendall metric but the first one with Spearman for ranking twitter terms according to their association with the positive sentiment.

1 Introduction

Twitter is one of the most social media platforms which allows the users to express their opinions and feelings towards different issues. The users have become an important source of content. This content may be interesting to analyze for those who are interested in understanding user's interests such as buyers, sellers and producers.

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 sentence level task, but some limitations should be considered in such sentences. The size of tweet is limited to 140 characters, informal language, emotion icons and non-standard expressions are very used, and many spelling errors can be find 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 a lexicon based which uses 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 lexicons, in the other hand the automatic lexicons needs labeled data. The second approach is machine learning one which uses annotated texts with given labels to learn a classifying model. Both lexicon and machine learning approaches can be combined to achieve a better performance. 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.

In this paper, we present our supervised system which adapts a logistic regression classifier with several groups of features and weighting schema for positive and negative labels. The features are grouped into 5 groups: word n-gram, lexicon-based, negation, Z score and semantic features. We also describe our system used for ranking terms according to their positivity, in which we derive the term polarity score from different lexicons.

The rest of this paper is organized as follows. Section 2 outlines existing work of sentiment analysis in Twitter. Section 3 describes the data and resources. The features we used for training the classifier presented in Section 4. Our experiments are described

Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015), pages 568–573,
Denver, Colorado, June 4-5, 2015.

in section 5, our participation in subtask E is described in section 6 and future work is presented in section 7.

2 Related Work

Three main approaches for sentiment analysis can be identified in Twitter. 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 dictionaries have been created manually such as MPQA Lexicon (Wilson et al., 2005) or automatically such as SentiWordNet (Baccianella et al., 2010).

Machine learning approach adapts different classifiers and features. Naive Bayes, Maximum Entropy MaxEnt and Support Vector Machines (SVM) were adapted in (Go et al., 2009) in which the authors reported that SVM outperforms other classifiers. They tried a unigram and a bi-gram 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 results. Authors in (Hamdan et al., 4 29) used the concepts extracted from DBPedia and the adjectives from WordNet, they reported that the DBpedia concepts are useful with Nave-Bayes classifier but less useful with SVM. Many features were used with SVM including the lexicon-based features in (Mohammad et al., 2013) which seem to get the most gain in performance. Another work has also proved the importance of lexicon-based features with logistic regression classifier (Miura et al., 4 08; Hamdan et al., 2015a; Hamdan et al., 2015b).

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 et al., 2011) authors demonstrated that using label propagation with Twitter follower graph improves the polarity classification. In (Tan et al., 2011) authors employed social relation for user-level sentiment analysis. In (Hu 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 Data and Resources

3.1 Labeled Data

We used the data set provided in SemEval 2013 for subtask B of sentiment analysis in Twitter (Nakov et al., 2013). 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 2015 contains about 2390 tweets (Rosenthal et al., 5 06). Table 1 shows the distribution of each label in each data set.

Twitter	all	neg.	pos.	neut.
train	9684	1458	3640	4586
dev	1654	340	739	575
test-2015	2390	365	1038	987

Table1. Sentiment labels distribution in the training and development, test data sets.

3.2 Sentiment Lexicons

The system exploits two types of sentiment lexicons: manual constructed lexicons and automatic ones. The manual ones are the Bing Lius Opinion Lexicon which is created in (Hu and Liu, 2004) and augmented in many research papers; and MPQA subjectivity lexicons (Wilson et al., 2005). Both lexicons contain English words annotated positive and negative. While the automatic lexicons are NRC Hashtag Sentiment Lexicon (Mohammad, 6 07), Sentiment140 Lexicon (Mohammad et al., 2013), and SentiWordNet (Baccianella et al., 2010). NRC Hashtag Sentiment Lexicon and Sentiment140 Lexicon contain tweet terms with scores, positive score indicates association with positive sentiment, whereas negative score indicates association with negative sentiment. NRC has entries for 54,129 unigrams and 316,531 bigrams; Sentiment140 has entries for 62,468 unigrams, 677,698 bigrams. SentiWordNet is the result of automatically annotating

all WORDNET synsets according to their degrees of positivity, negativity, and neutrality.

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 the these terms. This dictionary maps certain twitter expressions and emotion icons by their meaning or their corresponding sentiment (e.g. gr8 replaced by great, :) replaced by very-happy).

4 Feature Extraction

4.1 Word ngrams

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.

4.2 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 the negation key and a punctuation. All these words have been added to the feature space.

4.3 Twitter dictionary

All terms presented in a 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.

4.4 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.

4.5 Z score

Z score can distinguish the importance of each term in each class, their performances have been proved in (Hamdan et al., 2014). 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 (Hamdan et al., 4 29) we tested different thresholds for choosing the words which have higher Z score.

$$Zscore(ti) = \frac{tfr_{ij} - mean_i}{sd_i} \quad (1)$$

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.

4.6 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.

4.6.1 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.

4.6.2 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.

4.6.3 Semantic Role Labeling Features

Authors in (Ruppenhofer and Rehbein, 2012) 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 trained the L1-regularized Logistic regression classifier implemented in LIBLINEAR (Fan et al., 2008). The classifier is trained on the training data set using the features of Section 4 with the three polarities (positive, negative, and neutral) as labels. 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 development set 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 have given the best F1 score for the development set and therefore it was selected for our submission.

5.2 Results

The evaluation score used by the task organizers was the averaged F1-score of the positive and negative classes. In the SemEval-2015 competition, our submission is ranked third (64.27) over 40 submissions, just 0.57% less than the first system.

Table 2 shows the results of our experiments after removing a feature group at each run for the three test sets 2013, 2014, and 2015. For the test set 2015, we note that using Z score feature provides a gain of 0.45%, n-gram provides a gain of 0.28%, lexicon features gain is about 3.31%, LDA gain is 0.8%, Brown clusters 0.44%, semantic role labeling decreases the F1 score by 0.83%. The most influential features is the sentiment lexicon features; they provided gains of 3.31%.

Because of negative effect of semantic role labeling features, we have done another analysis in order to estimate if these features are useful or not, the fact that the combination of features makes some of them not influential are not sufficient to consider the features not useful. Thus, we repeat the same classification process but add one feature group at a time (Tabel 3). Z score seems to give gain of 1.91%, LDA topics gain is 0.66%, semantic role labeling 0.64%, brown clusters 3.38% and sentiment lexicons 6.58%. The most influential features is also the sentiment lexicon features. Brown cluster features obtains an interesting gain of 3.38%. From the previous two analysis, we find that sentiment lexicon features are the most influential ones as concluded by (S. M. Mohammad et al., 2013). Some features have improved the performance in test set 2015 but not in the other test sets such as Z score, Semantic Role Labeling.

Run	Test-2015	Test-2014	Test-2013
All features	64.27	71.54	71.34
all-zscore	63.82	73.05	69.99
all-lexicons	60.96	67.6	66.63
all-ngram	63.99	69.06	69.67
all-srl	65.1	71.81	70.41
all-topics	63.47	71.49	71
all-brown	63.82	70.74	69.9

Table2. The F1 score for each run, All features run exploits all features while the others remove a feature group at each run Zscore, lexicons, n-gram, srl, topics and brown cluster, respectively.

Run	Test-2015	Test-2014	Test-2013
bl	57.47	66.71	66.25
bl+lexicon	64.05	70.57	69.31
bl+zscore	59.38	63.47	65.28
bl+brown	60.85	66.71	66.25
bl+topics	58.13	-	-
bl+srl	58.13	66.69	63.35

Table3. The F1 score for each run, bl run exploits the n-gram, negation, twitter dictionary features

while the other runs add to bl one feature group at each run, lexicon, Zscore, brown, topics, slr features have been respectively added.

6 SubTask E: determining strength of association of Twitter terms with positive sentiment

This subtask is new in SemEval-2015, the objective is to provide for each Twitter term a score between 0 and 1 that is indicative of its strength of association with positive sentiment. If a word is more positive than another, then it should have a higher score than the other. Participants are provided with 200 terms with their scores as a trail data. The test data includes 1315 terms to rank. The organizers have chosen Kendall's Tau correlation coefficient to compare the ranked lists, they have also provided the scores of Spearman's Rank Correlation, but participating teams will be ranked according to Kendall's Tau.

To rank these terms, we have used six different sentiment lexicons for computing the score for each twitter term. Four of them are described in section 3.2 (manaul lexicons: Bing Liu and MPQA Subjectivity Lexicon , automatic constructed lexicons: NRC Hashtag and Sentiment140) and we have built two other automatic construction lexicons: the first named PMi-Sem from the training tweets provided by SemEval-2013 sub-task B Table 1, the second named PMI-sentiment140 from the sentiment140 corpus (Go et al., 2009), we calculated PMI from the labeled tweets for the two corpus using the following equation:

$$PMI(word, positive) = log\frac{p(positive, word)}{p(positive).p(word)} \quad (2)$$

where p(positive,word): The joint probability of the positive class and the word. p(positive): the probability of positive class. p(word): the probability of the word in whole corpus.

6.1 Score computing

If the word exists in a manual constructed lexicon (two lexicons), a score of 1 is assigned if the word is positive else -1 if negative. If the word exists in an automatic constructed lexicon (four lexicons), the lexicon score of the word is used. For each lexicon which does not have the word a default score is assigned, this default score is chosen to be 1/(number of the words in the test set). the final score is the average score of the previous six scores.

Run	Kendall	Spearman
all	0.621	0.820
all-BingLiu	0.616	0.816
all-MPQA	0.616	0.815
all-NRC Hashtag	0.510	0.689
all-Sentiment140	0.617	0.813
all-PMI-Sem	0.620	0.822
all-PMI-sentiment140	0.621	0.821

Table4. The results of Twitter term ranking, the first run *all* exploits all six lexicons, one lexicon is removed in the following runs.

The test data set contains 1315 twitter terms. Our system is ranked second with Kendall 0.004% less than the first ranked system, but first with Spearman. Table 4 shows our results with the two evaluation metrics. We repeat the experiment after removing one lexicon at each run, we can note that NRC Hashtag is the most influential lexicon.

7 Conclusion and Future Work

In this paper, we tested the impact of combining several groups of features on the sentiment classification of tweets. A logistic regression classifier with weighting schema is used, the sentiment lexicon-based features seem to get the most influential effect with the combination.

We have also exploited four existing lexicons and constructed two other lexicons using PMI metric in order to rank the twitter terms according to their association with positive sentiment.

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 which we think promising in measuring the association between each term and sentiment labels.

References

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *in Proc. of LREC*.

Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., and Lin, C.-J. (2008). LIBLINEAR: A library for large linear classification. 9:1871–1874.

Go, A., Bhayani, R., and Huang, L. (2009). Twitter sentiment classification using distant supervision. pages 1–6.

Hamdan, H., Bechet, F., and Bellot, P. (2013-04-29). Experiments with DBpedia, WordNet and SentiWordNet as resources for sentiment analysis in micro-blogging. In *International Workshop on Semantic Evaluation SemEval-2013 (NAACL Workshop)*.

Hamdan, H., Bellot, P., and Bechet, F. (2014). The impact of z_score on twitter sentiment analysis. In *In Proceedings of the Eighth International Workshop on Semantic Evaluation (SemEval 2014)*, page 636.

Hamdan, H., Bellot, P., and Bechet, F. (2015a). lsislif: Feature extraction and label weighting for sentiment analysis in twitter. In *In Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*.

Hamdan, H., Bellot, P., and Bechet, F. (2015b). Sentiment lexicon-based features for sentiment analysis in short text. In *In Proceeding of the 16th International Conference on Intelligent Text Processing and Computational Linguistics*.

Hu, M. and Liu, B. (2004). 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.

Hu, X., Tang, L., Tang, J., and Liu, H. (2013). Exploiting social relations for sentiment analysis in microblogging. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, WSDM '13, pages 537–546. ACM.

Miura, Y., Sakaki, S., Hattori, K., and Ohkuma, T. (2014-08). 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.

Mohammad, S. (2012-06-07). #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.

Mohammad, S. M., Kiritchenko, S., and Zhu, X. (2013). NRCCanada: Building the state-of-the-art in sentiment analysis of tweets. In *In Proceedings of the International Workshop on Semantic Evaluation, SemEval 13*.

Nakov, P., Rosenthal, S., Kozareva, Z., Stoyanov, V., Ritter, A., and Wilson, T. (2013). 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.

Rosenthal, S., Nakov, P., Kiritchenko, S., Mohammad, S. M., Ritter, A., and Stoyanov, V. (2015-06). SemEval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015. Association for Computational Linguistics.

Ruppenhofer, J. and Rehbein, I. (2012). 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.

Speriosu, M., Sudan, N., Upadhyay, S., and Baldridge, J. (2011). Twitter polarity classification with label propagation over lexical links and the follower graph. In *Proceedings of the First Workshop on Unsupervised Learning in NLP*, EMNLP '11, pages 53–63. Association for Computational Linguistics.

Tan, C., Lee, L., Tang, J., Jiang, L., Zhou, M., and Li, P. (2011). User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pages 1397–1405. ACM.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., and Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations*, HLT-Demo '05, pages 34–35. Association for Computational Linguistics.