Sentibase: Sentiment Analysis in Twitter on a Budget

Satarupa Guha [*], **Aditya Joshi** [*], **Vasudeva Varma**
Search and Information Extraction Lab
International Institute of Information Technology, Hyderabad
Gachibowli, Hyderabad, Telengana, India
`{satarupa.guha,aditya.joshi}@research.iiit.ac.in`
`vv@iiit.ac.in`

Abstract

Like SemEval 2013 and 2014, the task Sentiment Analysis in Twitter found a place in this year's SemEval too and attracted an unprecedented number of participations. This task comprises of four sub-tasks. We participated in subtask 2 — Message polarity classification. Although we lie a few notches down from the top system, we present a very simple yet effective approach to handle this problem that can be implemented in a single day!

1 Introduction

Social media not only acts as a proxy for the real world society, it also offers a treasure trove of data for different types of analyses like Trend Analysis, Event Detection and Sentiment Analysis, to name a few. SemEval 2015 Task 10 subtask B (Rosenthal et al., 2015) specifically deals with the task of Sentiment Analysis in Twitter. Sentiment Analysis in social media in general and Twitter in particular has a wide range of applications — Companies/services can gauge the public sentiment towards the new product or service they launched, political parties can estimate their chances of winning the upcoming elections by monitoring what people are saying on Twitter about them, and so on. In spite of the availability of huge amount of data and the huge promises they entail, working with social media data is far more challenging than regular text data. Being user-generated, the data is noisy; there are misspellings, unreliable capitalization, widespread use

of creative acronyms, lack of grammar, and a style of writing that is very typical of its own which makes the problem of Sentiment Analysis on Twitter more challenging. Also, the cues for positive or negative sentiment in social media text are starkly different, thereby generating a whole new domain for exploration.

2 Related Work

SemEval 2013 (Nakov et al., 2013) and 2014 tasks (Rosenthal et al., 2014) on Sentiment Analysis in Twitter not only contributed to this field by making huge amounts of annotated datasets available for research, but also encouraged researchers to come up with better solutions for this challenging problem. There has been numerous initiatives outside SemEval too. (Pak and Paroubek, 2010) is one of the early attempts at using Twitter as a corpus for Sentiment Analysis, which shows how to automatically collect a corpus for the same and performs linguistic analysis of the collected corpus. (Bakliwal et al., 2012) presents a simple sentiment scoring function which uses prior information to classify and weight various sentiment bearing words/phrases in tweets. (Wilson et al., 2005) demontrates an efficient technique for automatically identifying the contextual polarity for a large subset of sentiment expressions. (Mohammad et al., 2013) and (Kiritchenko et al., 2014) establishes benchmark in Sentiment Analyis in Twitter as well as in the field of Aspect Based Sentiment Analysis by incorporating various innovative linguistic features. (Agarwal et al., 2011) introduced POS-specific prior polarity features and (Kouloumpis et al., 2011) explored the use of a tree

---

The first two authors made equal contribution to this work

kernel to obviate the need for tedious feature engineering. (Kouloumpis et al., 2011) evaluate the usefulness of existing lexical resources as well as features that capture information about the informal and creative language used in microblogging. Recent publication from (Socher et al., 2013) has further raised the bar for Sentiment Analysis in general, but it is not specifically designed to tackle tweets data.

3 Approach

3.1 Preprocessing

We acquire a list of acronyms and their expanded forms [1]. We use this list as a look-up table and replace all occurrences of acronyms in our data by their expanded forms. We normalize all numbers that find a place in our data by replacing them with the string '0'. We do not remove stop words because they often contribute heavily towards expressing sentiment/emotion. We do not also stem the words because stemming leads to the loss of the parts of speech information of the word and makes the use of lexicons unnecessarily complicated.

3.2 Vocabulary Generation

We assign a unique ID to all words occurring in our data. All the hashtags we encounter in the data are hashed to a single string place holder and a single unique ID is assigned to it, as opposed to different Ids for different hashtags. Hashtags are mostly formed by concatenation of multiple words without any space in between, and therefore, unless hashtags are segmented into meaningful chunks, raw hashtags seldom add any semantics to the sentence. Hence, we do not distinguish between the different hashtags and consider them as a single unit. Similarly, we hash all mentions of the kind @user1 and @user2 to a single string placeholder and assign a single ID to it. This is because these words prefixed by '@' are all named entities and do not contribute anything to the semantic meaning or towards the polarity of a tweet.

3.3 Feature Engineering

The task required us to classify a tweet into positive, negative and neutral polarity categories. This can essentially be treated as a 2-step process

- Classify each tweet into subjective (positive/negative) and objective(neutral) classes.

- Classify subjective tweets into positive and negative ones.

We keep this philosophy in mind, but do not explicitly model the problem as two sub-problems. We treat them as a single step, but we select features such that some of the features are best suited for distinguishing between subjective and objective classes, while some others are engineered to be able to tell a positive tweet from a negative one. The problem with treating the problem as a pipeline of two steps is that we would have to deal with the propagation of errors from one step to the other. If a subjective tweet is mis-classified as an objective one, we rob that tweet of its opportunity of being classified any further in the next step and immediately label them as neutral. This might be detrimental in cases where certain features lead us to believe a tweet is objective, while a combination of all features might rightly lean them towards positive or negative polarities. We take aid from both extrinsic features like emoticons and grapheme stretching as well as intrinsic ones like unigrams and so on. Following is the list of features we employ and also their underlying motivation:

- *Unigrams* — For each word in a tweet, we look up the vocabulary we generated in the previous step. If the word is present in the vocabulary, we determine its position in the feature vector from the unique ID assigned to it and put 1 in its position. All other positions are 0 by default. Unigram features contribute to understanding both the distinction between subjective and objective tweets as well as between positive and negative tweets.

- *Number of hashtags* — Inspection of the data lead us to believe that the more the number of hashtags in the tweets, the more the author's involvement with it and hence more the subjectivity.

- *Presence of URLs* — A factual tweet is often accompanied by a URL as a proof of its validity or for more enthusiastic of the author's followers to go and explore the news/fact further. Hence, presence of URLs is likely to indicate that the sentence is objective/neutral.

  For example- "Jose Iglesias / Igleisas started at shortstop Wednesday night for the second http://t.co/Gkpx9Blu" and "Today In History November 02, 1958 Elvis gave a party at his hotel before going out on maneuvers. He sang and... http://t.co/Za9bLTcE"

- *Presence of exclamation marks* — From our observation, a subjective tweet is much more likely to be ended by exclamation marks than a purely factual or objective tweet. Further, positive tweets are more prone to contain exclamation than negative ones.

- *Presence of question marks and wh-words* — Tweets containing question marks or wh-words like "why" and "where" are seldom objective. Statistics tells us that this feature can act as a strong cue for not only subjectivity, but also of negativity.

- *Number of positive/negative emoticons* — An emoticon is a representation of a facial expression such as a smile or frown, formed by various combinations of keyboard characters and used heavily in tweets to convey the writer's feelings or intended tone. Quite intuitively, positive emoticons accompany positive tweets and negative emotions juxtapose negative tweets. More the number of emoticons, it leans with more confidence towards the corresponding polarity. However, a lot of sarcastic tweets also contain positive emoticons, but we do not explicitly handle sarcasm and hence ignore such possibility.

- *Number of named entities* — Just like URLs, named entities act as another indication for factual and hence neutral sentences. For example, "Remember this? Santorum: Romney, Obama healthcare mandates one and the same http://t.co/sIoG48TO #TheRealRomney @userX @userY". We extract named entities

using a python wrapper for the Stanford NER tool (Finkel et al., 2005). However, ablation experiments done after the submission of system in the competition revealed that this feature actually ended up degrading performance by more than 2% of F1 score.

- *Grapheme Stretching* — Words with characters repeated multiple times (at least twice) herald strong subjectivity, most often positive. For example, "Not only is @userZ home from China, she's in LA...I called her and screamed Mandyyyyyyyyyyyyyy...I'm gonna hug her for 2 hrs tomorrow!" and "daniel radcliffe was sooo attractive in the 3rd and 5th films omg im in love". We used the number of grapheme stretched words as a feature for our sentiment classifier.

- *Number of words with unusual capitalization* — Words with characters made upper-case or lower-case out of turn might potentially convey subjectivity. This feature also proved to slightly degrade performance, during post-competition ablation experiments.

- *Number of words with all the characters capitalized* — Strongly positive or strongly negative tweets often have words in all caps in order to convey the excitement that normally the loudness of a voice intones.

- *Presence of numbers* — Numbers are used profusely in factual tweets. For example- "13:58 Steven Pourier, Jr. (OLC) MADE the 2nd of the 2 shot Free Throw. DaSU leads 90 - 36 in the 2nd Half. #NAIAMBB". Hence the presence of numbers can be used as an useful feature to distinguish between subjective and objective tweets.

- *Lexicon features*- We use 15 lexicon features extracted from publicly available lexicons, which prove to be one of the most powerful features in our features list. Social media data, specially tweets, have a style of language use that is quite different from other text data. We included lexicons which are specially tailored to handle social media data, like Senti-

ment140 and NRC Hashtag Lexicon. We elaborate on the lexicon features as the following:

From Sentiwordnet (Baccianella et al., 2010), we extract

- Number of positive tokens
- Number of negative tokens
- Total positive sentiment score
- Total negative sentiment score
- Maximum sentiment score

From Bing Liu's opinion lexicon (Hu and Liu, 2004), we extract

- Number of positive tokens
- Number of negative tokens

From MPQA subjectivity lexicon (Wilson et al., 2005), we extract

- Number of positive tokens
- Number of negative tokens

From NRC Emotion Association lexicon (Mohammad and Turney, 2013), we extract

- Number of positive tokens
- Number of negative tokens

From Sentiment140 lexicon (Go et al., 2009), we extract

- Sum of sentiment score
- Maximum sentiment score

From NRC Hashtag Lexicon (Mohammad and Kiritchenko, 2014), we extract

- Sum of sentiment score
- Maximum sentiment score

3.4 Training Classifier

Once we have extracted all the features, we train a linear SVM using Python based Scikit Learn library (Pedregosa et al., 2011) for the purpose of classification. We experimentally ascertained the optimal value of the parameter C to be 0.025. In order to cope with the slight class imbalance in the data, we automatically adjust weights inversely proportional to class frequencies.

Feature	F1
all	56.67
all - number of entities	58.68 [2]
all - grapheme	56.52
all - exclamation	55.91
all - emoticons	56.61
all - number of hastags	56.69
all - unigrams	53.52
all - lexicons	48.40
all - wh words	56.62
all - illegal capitalization	56.90 [2]

Table 1: Ablation Experiment on Twitter 2015 dataset.

Dataset	Our Score	Best Score
Twitter 2015	56.67	64.84
Twitter 2015 Sarcasm	62.96	65.77
Twitter 2014	63.29	74.42
Twitter 2014 Sarcasm	47.07	59.11
Twitter 2013	61.56	72.80
Live Journal 2014	67.55	75.34
SMS 2013	59.26	68.49

Table 2: Official Results for SemEval 2015.

4 Experiments and Results

We used the official training and test sets provided for the SemEval 2015 task to train and evaluate our system. Tweets in the training data that were not available any more through the Twitter API were removed from the training set. For the evaluation, we compute precision, recall and F1 measures as computed by the scorer package provided for the task. Table 1 shows the ablation experiment we carried out, thereby highlighting the usefulness of the various features used. Table 2 records the F1 score obtained by our submission on different datasets. Our performance on Twitter 2015 Sarcasm data set is encouraging - we stand 4th on the data set.

5 Conclusion

This papers details the description of the system submitted by team Sentibase for SemEval 2015 Task 10. As the title of the paper suggests, the goal

---

[2]The fact that this feature degrades performance became clear during post-competition experiments

of this work was more to, put together a complete Sentiment Analyzer for Twitter in a day's time that achieves competitive performance without going through complex modeling techniques, than to up the ante in the state-of-the-work picture of Sentiment Analysis.

References

Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment analysis of twitter data. In *Proceedings of the Workshop on Languages in Social Media*, LSM '11, pages 30–38, Stroudsburg, PA, USA.

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *in Proc. of LREC*.

Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, and Vasudeva Varma. 2012. Mining sentiments from tweets. In *Proceedings of the 3rd Workshop in Computational Approaches to Subjectivity and Sentiment Analysis*, WASSA '12, pages 11–18, Stroudsburg, PA, USA.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *In ACL*, pages 363–370.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

Minqing Hu and Bing Liu. 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, New York, NY, USA.

Svetlana Kiritchenko, Xiaodan Zhu, Colin Cherry, and Saif Mohammad. 2014. Nrc-canada-2014: Detecting aspects and sentiment in customer reviews. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 437–442.

Efthymios Kouloumpis, Theresa Wilson, and Johanna Moore. 2011. Twitter sentiment analysis: The good the bad and the omg!

Edward Loper and Steven Bird. 2002. Nltk: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Saif M. Mohammad and Svetlana Kiritchenko. 2014. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, pages n/a–n/a.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. 29(3):436–465.

Saif M. Mohammad, Svetlana Kiritchenko, and Xiaodan Zhu. 2013. Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets. *CoRR*, abs/1308.6242.

Preslav Nakov, Sara Rosenthal, Zornitsa Kozareva, Veselin Stoyanov, Alan Ritter, and Theresa Wilson. 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, Atlanta, Georgia, USA, June.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*.

Fabian Pedregosa, Gal Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and douard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August.

Sara Rosenthal, Alan Ritter, Preslav Nakov, and Veselin Stoyanov. 2014. Semeval-2014 task 9: Sentiment analysis in twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 73–80, Dublin, Ireland, August.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif M Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation*, SemEval '2015, Denver, Colorado, June.

Richard Socher, Alex Perelygin, Jean Y. Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, HLT '05, pages 347–354, Stroudsburg, PA, USA.