sielers : Feature Analysis and Polarity Classification of Expressions from Twitter and SMS Data

Harshit Jain, Aditya Mogadala and Vasudeva Varma
Search and Information Extraction Lab, IIIT-H
Hyderabad
India
harshit.jain@research.iiit.ac.in, aditya.m@research.iiit.ac.in,
vv@iiit.ac.in

Abstract

In this paper, we describe our system for the SemEval-2013 Task 2, Sentiment Analysis in Twitter. We formed features that take into account the context of the expression and take a supervised approach towards subjectivity and polarity classification. Experiments were performed on the features to find out whether they were more suited for subjectivity or polarity Classification. We tested our model for sentiment polarity classification on Twitter as well as SMS chat expressions, analyzed their F-measure scores and drew some interesting conclusions from them.

1 Introduction

In recent years there has been a huge growth in popularity of vaious social media microblogging platforms like Twitter. Users freely share their personal opinions on various events and entities on these platforms. However, while character constraints make sure the opinions are short and to the point, they also contribute to the noisy nature of Twitter data.

The contextual polarity of the phrase in which a particular instance of a word appears may be quite different from the word's prior polarity. Positive words are used in phrases expressing negative sentiments, or vice versa. Also, quite often words that are positive or negative out of context are neutral in context, meaning they are not even being used to express a sentiment. This is evident from the example of underlined phrase in the following tweet:

> *Lana Del Rey at Hammersmith Apollo in May...Very badly want tickets*

In a technique with large lexicon of words marked with their prior polarity, *badly* would have a negative score making the whole sentence with negative sentiment. Even if we perform phrase-level analysis for the phrase "*Very badly*", *Very* only acts as an intensifier for *badly* and the whole sentence is still marked negative. It's only when we look further from the underlined phrase that we realize that "*Very badly*" in the context of wanting something shows positive sentiment.

Early work on sentiment analysis is based on document-level analysis of reviews (Pang, B., and Lee, L., 2004). This approach isn't feasible for microblogging data due to the extremely small size of individual documents. The results on the effectiveness of part-of-speech features are mixed. While most regard POS features helpful in subjectivity classification (Barbosa, L. and Feng, J., 2010), some report very insignificant improvement on using them (Kouloumpis, E., Wilson, T. and Moore, J., 2011). However, most phrase-level approaches began with a large lexicon of words marked with their prior polarity (Kim, S. M., and Hovy, E., 2004; Hu, M., and Liu, B, 2004). Wilson, Wiebe and Hoffman (2005) sought to include contextual polarity in the foray by using various dependency relation based features for subjectivity and polarity classification. Our goal is to perform contextual sentiment polarity classification in the domain of noisy expressions from tweets and SMS messages.

2 Data

We use the annotated Twitter expressions provided by SemEval-2013 Task 2 (Wilson et al., 2013) or-

*Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 525–529, Atlanta, Georgia, June 14-15, 2013. ©2013 Association for Computational Linguistics

ganizers for training our model. Each instance of the data contains an expression and its parent tweet. There are a total of 24939 tweet expressions in the training dataset and they are annotated into four classes:

- **Objective**: Expressions carrying no opinion by themselves or even in the context of their parent tweet.
- **Positive**: Expressions carrying positive sentiment in the context of the parent tweet.
- **Negative**: Expressions carrying negative sentiment in the context of the parent tweet.
- **Neutral**: Expressions carrying prior subjectivity but are rendered objective in the context of their parent tweet.

Two separate lexicons for emoticons and interjections having non-zero prior polarities were created. 47 Subjective emoticons were extracted from training data as well as from various popular chat services. 212 Subjective interjections were extracted from training data as well from Wiktionary[1].

We test our trained model on two separate test datasets provided by SemEval-2013 Task 2 organizers, 1) Twitter expressions and 2) SMS expressions.

2.1 Preprocessing

Data preprocessing consists of three steps: 1) Tokenization, 2) Part-of-Speech (POS) tagging, and 3) Normalization. For the first two steps we use Twitter NLP and Part-of-Speech Tagging system (Gimpel, K., et al., 2011). It is a Tokenizer and POS Tagger made for Twitter dataset and thus contains separate POS tags for hash-tags(*#*), at-mention(*@*), URLs and E-Mail addresses(*U*) and emoticons(*E*). The POS Tagger identifies common abbreviations and tags them accordingly. We use Twitter NLP and Part-of-Speech Tagging system for the SMS expressions too due to similar noisy nature of SMS data. For the normalization process, all upper case letters are converted to lower case, and instances of repeated characters are replaced by a repetition of two characters. This is done so that existing legal words having characters repeating two times aren't harmed. #hash-tags are stripped of the # character and then treated as a normal word/phrase, at-mention(@) denote the name of a person/organization and thus they are treated as proper noun and since URLs don't carry any sentiment, they are ignored in the expression. We expect the normalization process to aid in forming better features and in turn improving the performance of the system as a whole.

3 Features

We use three types of features for our classification experiments,

- Phrase Prior Polarity Features
- POS Tag Pattern Features
- Noisy data specific Features

Both Phrase Prior Polarity and POS Tag features are computed for the expression to be analyzed as well as, if available, two words [2] before and after the expression.

3.1 Phrase Prior Polarity Feature

Every expression in the dataset is represented by its aggregate positive and negative polarity score. Senti-Wordnet (Baccianella, S., Esuli, A., and Sebastiani, F., 2010), Emoticon Lexicon and an Interjection Lexicon are used to calculate these prior polarities. Bigrams and trigrams are identified by their presence in Senti-Wordnet. For each identified unigram, bigram or trigram, we compute the mean of all its subjective wordnet sense scores under the POS tag assigned to it. If a unigram word isn't present in Senti-Wordnet, its stemmed[3] form is searched keeping the original POS Tag. We perform negation detection by enabling a flag whenever a word occurring in negation list appears. The negation list consists of words like *no*, *not*, *never*, etc, as well all words ending with *-n't*. Negation words act as polarity reversers, for e.g., consider the following expression: *"not so sure"*. In a simple bag of words approach, *"not so sure"* wouldn't be classified as negative due to the presence of *sure*. To overcome this,

[1] `http://en.wiktionary.org/wiki/Category:English_interjections`

[2] The figure of two words was reached empirically upon trying various lengths.

[3] The stemmer used is Snowball Stemmer for English.

prior polarities of all words are reversed on the occurrence of a negation word. Some negation words such as *no*, *not*, *never*, also carry their own negative score (-1), in case no subjective word is found in the expression, their individual negative score is added to the aggregate prior polarity of the expression. Adjectives and adverbs are treated as polarity shifters. They either shift the prior polarities of nouns and verbs, or in case of objective nouns and verbs, contribute their own prior polarities to the expression, e.g., *"exceedingly slow"*, *"little truth"*, *"amazing car"*, etc.

On encountering any emoticon or interjection in the expression that is present in our lexicon, its corresponding score is added to the aggregate prior polarity of the expression.

Finally, both positive and negative prior polarities of the expression are normalized by the number of words in the expression after tokenization.

3.2 POS Tag Pattern Feature

Both Tweets and SMS messages are extremely short. Twitter is a social microblogging platform having just 140 character space for a tweet while SMS messages have little word length due to typing constraints on a mobile device. All the above factors contribute to the noisiness of data. Hence, it isn't enough to find prior polarities of n-grams occurring in the expression. We thus formed a heuristic technique of using POS tag patterns as features. POS tag patterns carry information regarding POS tags combined with the location of their occurrence in the expression as a feature. For e.g., the POS tag pattern for the expression *"not so sure"* in the tweet

> *@thehuwdavies you think the Boro will beat Swansea? I'm not so sure, December/January is when we implode*

will be *RRA*, where R = Adverb and A = Adjective.

3.3 Noisy data specific Features

Interjections and emoticons are useful indicators of subjectivity in a sentence. Even if many interjections or emoticons don't carry a defininte sentiment polarity, they do indicate that some sort of opinion from the user is available in the tweet or sms. Some examples of interjections and emoticons with no fixed prior polarity are, *"wow"*, *"oh my god"*, *":-o"*, etc.

4 Experiments and Results

Our goal for these experiments is two-fold. First, we want to evaluate the effectiveness of our features when using them for subjectivity classification as compared to sentiment polarity classification. Second, we want to evaluate and compare the performance of our learnt model when tested upon Twitter and SMS expression data. We use Naive Bayes classifier in Weka (Hall, M., et al., 2009) as the learning algorithm.

Feature Analysis between Subjectivity and Polarity Classification For our first set of experiments, we re-label all positive, negative and neutral expressions as subjective for subjectivity classification in the training dataset. For polarity classification we remove all objective expressions from the training dataset and perform 3-way classification between positive, negative and neutral expressions. In both cases we perform 10-fold cross validation on the training dataset. For subjectivity classification we have 24939 tweet expressions with 15565 objective and 9374 subjective expressions. Subjective expressions contain 5787 positive, 3131 negative and 456 neutral expressions. Table 1 shows the accuracy of subjectivity and sentiment polarity classification results and improvement due to each feature.

It is fairly evident from Table 1 that phrase prior polarity features are equally important for both subjectivity and sentiment polarity classification. The same however, doesn't completely hold true for the other two feature types. While POS Tag pattern features provide an improvement of 1.89% in subjectivity classification accuracy, they only provide a 0.64% increase in accuracy in polarity classification. Many inferences can be drawn from this result and a deeper analysis is required on POS tag patterns to prove that this wasn't a mere aberration. Emoticon and interjection feature too give lower improvement in accuracies during sentiment polarity classification (0.44%) as compared to subjectivity classification (0.83%). This, however, is expected since most common emoticons and interjections with prior polarities are already covered in the total score of the expression. Thus, the noisy data based binary features have significant contribution only when the emoticons and interjections aren't present in the lexicon. This implies that these binary features only

Features	Subjectivity	Polarity
f1	86.58	72.93
f1 + f2	88.47	73.57
f1 + f2 + f3	89.3	74.01
f1 + f2 + f3 - context	84.38	72.25

f1	:	Phrase Prior Polarity Features
f2	:	POS Tag Pattern Features
f3	:	Noisy Data Specific Features
context	:	Phrase Prior Polarity and POS Tag pattern features defined for 2 words before and after the expression

Table 1: Accuracies for all three features used for Subjectivity and Sentiment Polarity Classification.

Class	Precision	Recall	F-measure
positive	0.8120	0.8120	0.8120
negative	0.6477	0.7073	0.6762
neutral	0.3333	0.0375	0.0674

(a) Twitter expression data

Class	Precision	Recall	F-measure
positive	0.6823	0.8263	0.7475
negative	0.7520	0.6947	0.7222
neutral	0.0588	0.0063	0.0114

(b) SMS expression data

Table 2: Precision, Recall and F-measure scores for positive, negative and neutral classes computed on Twitter and SMS expressions data.

hint towards the expression being subjective. The *context* features, i.e., phrase prior polarity and POS tag pattern features defined for 2 words before and after the expression also carry more significance during subjectivity classification than in sentiment polarity classification.

Polarity Classification comparison for Twitter and SMS expression data For the second set of experiments comparing the performance of polarity classification in Twitter expressions and SMS expressions, we use the polarity classification model learnt in the above experiment. Tables 2(a) and 2(b) shows the precision, recall and F-measure scores for both Twitter and SMS expressions.

The polarity classification accuracies for Twitter and SMS expressions are 74.76% and 70.82%, respectively. Closer inspection of test data shows that SMS expressions exhibit more aggressive usage of abbreviations and slangs and are in general noisier than Twitter expressions. This is probably due to the fact that typing on a cellphone is more cumbersome than on a keyboard. The quantitative distribution of positive, negative and neutral classes in both datasets affects the F-measure scores of individual classes. This is evident from the difference in positive and negative F-measures of Twitter and SMS expressions data. In both datasets, neutral class F-measure is extremely low. This is partially expected due to the low quantity of neutral class expressions in Twitter (160/4435) and SMS (159/2334) data. Still, it seems that more fine-grained analysis of neutral expressions is required for better polarity classification accuracy.

Our method ranks 16th (F-measure: 0.7441) out of 28 participating systems for Twitter data and 12th (F-measure: 0.7348) out of 26 participating systems for SMS data. The best performing system have 0.8893(NRC-Canada) and 0.8837(GUMLTLT) averaged(positive, negative) F-measure score for Twitter and SMS data, respectively.

5 Conclusions

Our experiments on features show that phrase prior polarity features give good results for both subjectivity and polarity classification. POS tag pattern features, emoticon and interjection features, on the other hand, are better suited for subjectivity classification. A deeper analysis is required and various relational and dependency features should be identified and used to improve the performance of polarity classification. SMS expressions are noisier in general than Twitter expressions and thus the polarity classifier gives less accurate results for it. However, both of these datasets face problems common to the polarity classifier. More research is needed with a balanced dataset to understand various underlying relational causes for an expression to become neutral and to further confirm the conclusions of this paper.

References

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

Barbosa, Luciano, and Junlan Feng. 2010. Robust sentiment detection on twitter from biased and noisy data. *Proceedings of Coling*. Beijing.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-Speech Tagging for Twitter: Annotation, Features, and Experiments. *Proceedings of ACL 2011*.

Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. 2009. The WEKA data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.

Hu, Minqing, and Bing Liu. 2004. Mining and summarizing customer reviews. *KDD-2004*.

Kim, Soo-Min, and Eduard Hovy. 2004. Determining the sentiment of opinions. *Coling-2004*.

Kouloumpis, Efthymios, Theresa Wilson, and Johanna Moore. 2011. Twitter Sentiment Analysis: The Good the Bad and the OMG!. *Proceedings of ICWSM*. Barcelona.

Pak, Alexander, and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. *Proceedings of LREC*. Malta.

Pang, Bo, and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *Proceedings of the ACL*.

Theresa Wilson and Zornitsa Kozareva and Preslav Nakov and Sara Rosenthal and Veselin Stoyanov and Alan Ritter. *SemEval-2013 Task 2: Sentiment Analysis in Twitter*. Proceedings of the International Workshop on Semantic Evaluation. SemEval '13. June 2013. Atlanta, Georgia.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*. Vancouver.