Using Feature Selection Metrics for Polarity Analysis in RepLab 2012

Hogyeong Jeong and Hyunjong Lee

Seoul, Republic of Korea
{hogyeong.jeong,hyunjong.lee.s}@gmail.com

Abstract. In this workings notes paper for RepLab 2012, we describe our method of using feature selection metrics for polarity analysis. We use the correlation coefficient, a one-sided metric, to assign polarity scores to relevant words within a tweet; we then use aggregate of these scores to determine polarity of the tweet. Our results show a reasonable level of performance compared to other methods.

Keywords: Correlation coefficient, feature selection, polarity analysis

1 Introduction

This paper describes a correlation-coefficient based procedure for determining polarity of a tweet. Correlation coefficients have been used successfully for text categorization, and also for sentiment analysis [1][2][3] . In this paper, we describe how the correlation coefficient can be used to perform polarity analysis, and compare our results with other participants in RepLab 2012 [4].

1.1 Tasks Performed

Among the many tasks for RepLab 2012, we concentrated on polarity analysis of the profiling task. Also, while there were tweets in both English and Spanish, we chose to focus on only the English tweets.

1.2 Main Objectives of Experiments

The main objective of the experiment was to determine the polarity (positive, neutral, or negative) of a single tweet. Irrelevant tweets were excluded from the polarity analysis. Although this task is closely related with sentiment analysis [1], it is somewhat different as it focuses on reputation instead of sentiment.

1.3 Related Work

Our method uses some of the feature selection metrics that are described in [2] to perform polarity analysis. In particular, we use the one-sided correlation coefficient, as there are three polarity classes to consider: positive, neutral, and negative. Although how it is actually used can vary, correlation coefficient has been used in a closely related task of sentiment analysis [3].

2 Approaches Used

We submitted four runs for the task: one using the basic method, two using the basic method with modified thresholds, and the fourth that incorporated human input for borderline cases.

2.1 Preprocessing

First, we had to preprocess the tweets to extract relevant keywords that we would use to determine polarity. To do this, we used the Stanford part-of-speech tagger, and extracted nouns and adjectives [5].

2.2 Correlation Coefficients

As part of the RepLab 2012 task, we were given a small set of labeled files that we could use for training. Given the terms that we extracted above in the preprocessing phase, and the three polarity categories (positive, neutral, and negative), we calculated the correlation coefficient for a term t on class c_i as

$$CC(t, c_i) = \frac{\sqrt{N}[P(t, c_i)P(\bar{t}, \bar{c_i}) - P(t, \bar{c_i})P(\bar{t}, c_i)]}{\sqrt{P(t)P(\bar{t})P(c_i)P(\bar{c_i})}} \tag{1}$$

where $\bar{t}$ denotes other terms and $\bar{c_i}$ denotes other classes.

2.3 Basic Method

Using the correlation coefficients that we calculated above (for each term and each polarity class), we can sum the correlation coefficients of a class for all the relevant terms within a tweet. After the summation of coefficients is done for each polarity class (positive, negative, and neutral), we then assign polarity of the tweet to be that of the class corresponding to the largest summation.

While this represents the most basic usage of the correlation coefficients, we can modify the thresholds somewhat to try to achieve better performance.

2.4 Modified Threshold 1 - Same Class Proportions as the Training Set

One approach is to try to set the thresholds so that the resulting class proportions on the test set is equal to the class proportions on the training set. This method usually works best when the training set is bigger than the test set, and the test set is similar to the training set. Unfortunately for RepLab 2012, the test set was much larger than the training set and it was also quite different from the training set [4].

2.5 Modified Threshold 2 - Best Performance on the Training Set

Another approach is to set thresholds that corresponded to the ones that achieve best performance in the training set (via a 5-fold testing within the training set). Again, we would expect better performance if the training set is large and similar to the test set - neither of which were met for RepLab 2012.

2.6 Basic Method with Human Input

One nice result of the correlation coefficient approach is that we can get a measure of confidence on our classifications. For example, our confidence of a tweet being positive on classification whose values are (positive=3.8, neutral=0.2, negative=-2.5) is much larger than our confidence if the values were (positive=0.3, neutral=0.2, negative=-0.7) instead.

We can take advantage of this additional information by introducing human input for the borderline cases where the difference between the top two classes is small (we used 0.5 as a threshold). These are the cases where the automatically generated categorizations would have a high risk of being incorrect.

3 Results

The evaluation on classifications was performed using reliability, sensitivity, and the corresponding F measure, which are modified recall and precision measures [6]. There were a total of 42 runs submitted for the task, of which 3 served as baselines. Evaluations were done separately for the English and the Spanish tweets, and the results that we provide below correspond to the English results [4]:

Method (rank of 42)	**Reliability**	**Sensitivity**	**F(R,S)**
Best (ranked 1st)	.369	.350	.348
Human Input (ranked 10th)	.364	.275	.285
Basic (ranked 15th)	.265	.280	.260
Modified Threshold 1 (ranked 26th)	.230	.194	.198
Modified Threshold 2 (ranked 28th)	.241	.184	.194

As we feared, modifying the thresholds resulted in much worse performance, because the training set was small, and unlike the test set. Meanwhile, the basic correlation coefficient based method performed okay, ranking 15th of 42 submitted runs. As expected, our run integrating human input on just the borderline cases led to a marked improvement compared to the other methods.

4 Conclusion and Future Directions

We were able to achieve reasonable results using relatively simple approach using correlation coefficients. Further, we showed that we can markedly improve our results by incorporating human input on cases deemed to be borderline by the correlation coefficients.

As a future direction, we can try exploiting the massive background data that we did not use for our current results. Because the training set in this case was so small, we can expect better results if we can exploit the background data to help expand our training set. Once we have expanded the training set in such a way, we may be able to expect better results from the modified threshold approaches that were not able to perform well with a small training set.

References

1. Pang, B., Lee, L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval **2**(1-2) (2007) 1–135
2. Zheng, Z.: Feature selection for text categorization on imbalanced data. ACM SIGKDD Explorations Newsletter **6** (2004) 2004
3. Marchetti-Bowick, M., Chambers, N.: Learning for microblogs with distant supervision: Political forecasting with twitter. In: EACL. (2012) 603–612
4. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., Rijke, M.d.: Overview of replab 2012: Evaluating online reputation management systems. In: CLEF 2012 Labs and Workshop Notebook Papers. (2012)
5. Toutanova, K., Manning, C.D.: Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. In: In EMNLP/VLC 2000. (2000) 63–70
6. Gonzalo, J., Peters, C.: The impact of evaluation on multilingual text retrieval. In: SIGIR. (2005) 603–604