A Non-syntactic Approach for Text Sentiment Classification with Stopwords

V. Suresh
Dept. of Computer Science and Automation
Indian Institute of Science
Bangalore, India
suresh.venkatasubramanian@gmail.com

Ashok Veilumuthu
Dept. of Management Studies
Indian Institute of Science
Bangalore, India
ashok@mgmt.iisc.ernet.in

Avanthi Krishnamurthy
Dept. of Computer Science and Automation
Indian Institute of Science
Bangalore, India
avanthi.krishnamurthy@gmail.com

C. E. Veni Madhavan
Dept. of Computer Science and Automation
Indian Institute of Science
Bangalore, India
cevm@csa.iisc.ernet.in

Kaushik Nath
SAP Research
Bangalore, India
kaushik.nath@sap.com

Sunil Arvindam
SAP Research
Bangalore, India
sunil.arvindam@sap.com

ABSTRACT

The present approach uses stopwords and the gaps that occur between successive stopwords –formed by contentwords– as features for sentiment classification.

Categories and Subject Descriptors

I.2.7 [**Natural Language Processing**]: Text Analysis—*Sentiment Analysis*

General Terms

Sentiment Analysis, Text Mining

Keywords

Text and Language Applications, Sentiment Mining, Stopwords, Topic Models, Latent Dirichlet Allocation

1. INTRODUCTION

Automating the understanding of sentiments expressed in online reviews could provide useful handles for designing marketing strategies, campaigns and promotions and be a vital component for the commercial success of products. A typical approach to binary sentiment analysis (yes/no, good/bad, positive/negative) is as follows. Turney[3] presented a rule based approach comprising three steps. First, word bigrams (or phrases in general) that are capable of holding sentiment expressions are extracted from the text based on the underlying POS (part of speech) patterns that are usually associated with such semantic orientations. In the second step, the orientation of the bigrams are estimated by computing their PMI's (pointwise mutual information – a standard in IR applications– is indicative of the cooccurrence of a pair of words in a large corpus) with strong polarity indicator words like *excellent* and *poor*. Finally, a phrase is deemed positive or negative based on the PMI score. A review is categorized as positive or negative based on the overall score attained by the phrases in the review. The approach attains an accuracy of 66% for movie reviews. In this work we explore the possibility of performing sentiment classification which does not require syntactic preprocessing like POS tagging and is essentially independent from the effects of incorrect language usage. In this context we consider stopwords as features for such classification. The remainder of the work is organized as follows. In section 2 we present the motivation for the present approach. Description of the experiments carried out and the results obtained are given in section 3. We present our conclusions in section 4.

2. STOPWORDS AND SENTIMENTS

Stopwords are usually the non-semantic words like articles, prepositions, conjunctions, pronouns etc., they are numbered in hundreds whereas the size of English lexicon is of the order of hundreds of thousands of words. Also, stopwords are inevitable in usage and despite their small vocabulary they constitute a significant part of text thereby enabling a better estimation of their occurrence probabilities even with a small text corpus (such computations for contentwords usually require orders of magnitude larger corpus). This provides the basic appeal to use stopwords as features to study text. Previously, stopwords have been used as features for stylometric purposes[2] and studying the distribution of stopwords is an established approach in author identification. Given that stopwords are endowed with enough discriminatory powers to enable author identification, we consider it a natural extension to view sentiment classification as a variation of the author identification problem: the *authors* here are sentiments!

3. EXPERIMENTS AND RESULTS

We demonstrate the efficacy of our approach by applying it on a standard movie review dataset: *Polarity Dataset*

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

V2.0, available from Cornell University[1]. The dataset contains positive and negative reviews, 1000 each. All non-English characters are filtered out from these reviews as a basic cleanup operation. The overall methodology is as follows: the reviews –after suitable pre-processing as explained below– are converted to topic vectors using Latent Dirichlet Allocation (LDA)[1]. LDA is a semantic clustering approach which views each document in a corpus as a mixture of topics generated by an underlying topic model; words are probabilistically associated with topics (number of topics is determined by the user) and words relevant to a particular topic have a higher probability of getting generated under that topic. After computing topic vectors for each review in the movie review dataset, we train an SVM classifier with topic vectors from 800 reviews; the accuracy of the classifier is tested with the remainder of the reviews (200 of them). In our experiments we used a Gibbs sampler based LDA implementation *GibbsLDA++*[2] for computing the topic vectors for the review documents. The Support Vector Machine implementation *LIBSVM*[3] was used for classification. We use a standard list[4] of 571 stopwords in our experiments. Note that one could extract stopwords based on word frequencies alone –choosing the top one percent of the most frequently occurring words in a corpus is one such possibility. Before presenting the results obtained with LDA and SVM, we first show the results of using stopwords and *gaps* –number of contentwords that occur between successive stopwords, this will be explained in more detail shortly– with a naive Bayes classifier in order to establish a baseline. We use the popular spam filter *CRM114*[5] for performing naive Bayes classification on positive and negative reviews.

category	accuracy
positive	68% (136/200)
negative	60% (120/200)

Table 1: Naive Bayes classifier: Stopwords & Gaps

Table 1 shows results obtained for 200 test documents in each category (800 in each were used for training). On the average, a baseline of 64% is obtained with a straightforward classifier. This is very close to the results presented by Turney [3]. In the following, we show that our approach results in a marked improvement in accuracy over this baseline. To begin with, we show in Table 2 the results obtained by considering the entire review corpus as it is. With the observation that contentwords do not contribute to sentiment classification with LDA and SVM, we explore the efficacy of stopwords and *gaps* for the purpose of sentiment classification. As can be seen in any text, successive pairs of stopwords are separated by a sequence of contentwords. The length of these separations depend on the number of contentwords that occur between successive stopword pairs.

In our experiments, we removed all the contentwords from the reviews and retained only the stopwords and gaps in the documents (such document structures are facilitated by the bag-of-words model used by LDA). The documents contain stopwords and gaps –represented by their integer values– that occur between immediate stopword pairs. For example, the previous sentence becomes: *the* 3 *and* 2 *by* 0 *their* 2 *that* 1 *between* 3; wherein {the, and, by, their, that, between} are considered as stopwords present in that sentence. Note that 0 captures bigram stopword occurrences without any contentwords in-between (as in *by their*).

#	accuracy (SVM)	
topics (LDA)	positive reviews	negative reviews
10	25%	23%
15	50%	35%
20	54%	43.5%
25	44%	30%

Table 2: Effect of content words on accuracy

#	accuracy (SVM)	
topics (LDA)	positive reviews	negative reviews
10	60.5%	41%
15	62.5%	53%
20	46.5%	35%
25	78.5%	72%

Table 3: Effect of stop words and gaps on accuracy

The results obtained with this approach is shown in Table 3. Significant improvement is observed in the accuracy of both positive and negative reviews. Peak accuracy is observed for both the review classes for 25 topics. This is a 10% improvement over the baseline. We also note that experiments wherein the documents were formed only with stopwords (without the inclusion of gaps) failed to yield results better than the baseline (results not shown). We also note that the dataset considered has almost similar distribution of gaps in both the positive and negative review categories indicating that the accuracy of the approach is a result of the combined usage of both these properties –stopwords and gaps. Overall, our results seem to suggest that stopwords along with the gaps, dictate the expression of review sentiments.

4. CONCLUSIONS

The present approach is complementary to the conventional syntactic approaches for sentiment and opinion mining and hence a useful addition to the repertoire of opinion mining methods. Extending the approach from the present binary review classification (positive or negative) to perform more graded classifications (star ratings for example) would widen its applicability in sentiment analysis.

5. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[2] F. Mosteller and D. L. Wallace. *Inference and Disputed Authorship : The Federalist / [by] Frederick Mosteller [and] David L. Wallace*. Addison-Wesley, Reading, Mass. :, 1964.

[3] P. D. Turney. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. In *Procs.*, ACL '02, pages 417–424, 2002.

[1] `http://www.cs.cornell.edu/People/pabo/movie-review-data/`

[2] `http://gibbslda.sourceforge.net`

[3] `http://www.csie.ntu.edu.tw/~cjlin/libsvm`

[4] `http://www.ai.mit.edu/projects/jmlr/papers/volume5/lewis04a/lyrl2004_rcv1v2_README.htm`

[5] `http://crm114.sourceforge.net/`