Spam Review Detection Using Topic Modeling

ABSTRACT

Online spam reviews are deceptive evaluations of products and services. They are often carried out as a deliberate manipulation strategy to deceive the readers. Recognizing such reviews is an important but challenging problem. Most of the existing research initiatives on spam review detection have focused on either finding deceptive positive reviews or deceptive negative reviews. In this paper, we present our findings showing how keywords generated using topic modeling perform as a deciding factor in recognizing the truthful versus deceptive reviews. Our experimental results show that using only the keywords generated from topic modeling as features for a Support Vector Machines (SVM) classifier results in 87% classification accuracy compared to the 63% accuracy of human judges. We also compare our results with other approaches that consider all the words as features.

Author Keywords

Keywords- Deceptive Review; Opinion Spam; Topic Modeling; SVM; LDA; Data Mining;

ACM Classification Keywords

CCS → Computing methodologies → Artificial intelligence → Natural language processing → Information extraction.

INTRODUCTION

Online reviews help customers to make a choice and/or to assess the quality before buying a product or availing a service. However, reviews are only useful if they are not deceptive. Truthful reviews serve as a collective intelligence making it much easier to save customer's time and money. Reviews are also instrumental for companies to boost their sales. Therefore, to improve their chances of success, some businesses solicit and/or manufacture fictitious reviews that are deliberately written to appear authentic but to deceive the readers [1]. Such reviews are known as spam reviews or opinion spam reviews.

There are two kinds of spam reviews - *deceptive positive reviews* and *deceptive negative reviews*. *Deceptive positive reviews* refer to fake reviews that are usually given to promote a company's own products. For example, the authors in [2] have estimated that up to 6% of positive hotel reviews are deceptive, suggesting that some hotels might be posting fake positive reviews to hype their own offerings. On the other hand, *deceptive negative reviews* are fake reviews that are usually written to demote the competitors. For example, the authors in [3] reported that restaurants are more likely to receive deceptive negative reviews when they face increased competition from other restaurants that serve similar types of food. Therefore, it is important to identify both positive and negative deceptive reviews.

In this paper, we present an approach that uses Topic Modeling and Support Vector Machines (SVM) to detect both deceptive positive and deceptive negative reviews. Our approach uses only the topic words that are generated by topic modeling, compared to the existing approaches, e.g. [1 and 4], that use all the words in the dataset.

RELATED BACKGROUND

In the real world, it is very difficult for a human judge to be able to classify an online review as spam or authentic. In the study reported in [5], human judges achieved only 54% correct lie-truth judgments to discriminate lies from truths in documents. The authors in [1, 4] used three human judges for evaluating human performance in detecting deceptive reviews. The highest accuracy human judges achieved was 61% and 65% for positive and negative reviews, respectively. According to [6], this has led to the scarcity of labeled datasets that can be used to train a classifier to detect spam reviews. This motivated us to use topic modeling since it does not rely on labeled data for learning.

In [5], the authors made the first attempt to investigate opinion spam reviews. They considered duplicate and near-duplicate reviews to be indicative of spam reviews. Since then, many supervised approaches have been proposed for spam detection, e.g. [9, 10, 11]. However, as discussed earlier, due to the unreliability of manually labelled data, these approaches have limitations.

While earlier works, e.g. [1, 2, **Error! Reference source not found.**], explored different characteristics of reviews and reviewers, such as total number of reviews left by a reviewer, date of a review relative to when a product first became available, etc., the usefulness of applying topic modeling on spam review detection has been fully investigated.

Topic modeling is an approach for discovering topics from a large corpus of text documents. The most common output of a topic model is a set of word clusters and a topic distribution for each document. Each word cluster is called a topic and is a probability distribution over words in the corpus. Topics are aspects that refer to the attributes and components of an entity. For example, in the sentence "*The picture and sound* *quality of Samsung-HDTV are great.*", *picture* and *sound* are aspects of the Samsung-HDTV. As per [6], the advantage of topic modeling is that it can automatically extract aspects and put them into separate groups. For example, it can extract and group *organization*, *cleanliness*, and *comfort* under one aspect in a hotel review dataset.

One of the techniques for topic modeling is Latent Dirichlet Allocation (LDA) [13], which is an unsupervised method. LDA is a generative probabilistic model for collections of discrete data such as text corpora. In [14], the authors proposed a generative LDA-based topic modeling approach for fake review detection. Their approach is a variation of LDA that aims to detect subtle differences between the topic-word distributions of deceptive reviews versus truthful reviews. However, in our research, we use LDA-based semi-supervised learning to build our model for detecting deceptive and truthful reviews.

The researchers in [1, 5, 8] applied n-gram–based text categorization techniques for extracting all the words from the corpus to use them as features. Then they used these features in different machine learning techniques, such as SVM [1], logistic regression [5], and naïve Bayes [8], to identify deceptive reviews. We used SVM because it has been found that SVM comparatively performs better than other data mining algorithms in text classification [**Error! Reference source not found.**]. In our proposed approach, we also follow a two-step process, but we first extract the features using LDA topic modeling, and then use those extracted features in SVM for spam review detection.

PROPOSED APPROACH & SYSTEM IMPLEMENTATION

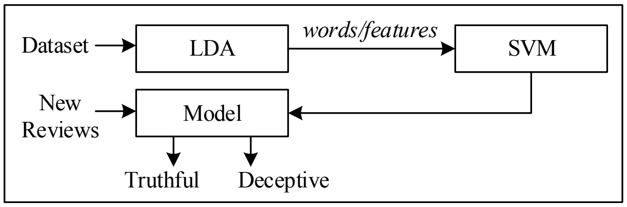

Figure 1. System Overview

In Figure 1, we summarize our approach. Input to the LDA is a dataset consisting of all the reviews. The LDA procedure requires the user to set two parameters specifying the number of components (*n_component*) and the number of top words (*n_top_words*). Number of components represents the total number of topics to be generated by LDA and number of top words represents the total number of top words that will be generated for each topic. Output from LDA is topic words (a.k.a. top words).

In the second step, we use the topic words as features for SVM. To be more specific, a linear SVM with stochastic gradient descent is used to build our model. The new queries are then classified by the model as truthful or deceptive.

To conduct our experiment, we used Ott et al.'s publicly available dataset of opinion spam [1]. The dataset contains 800 positive reviews (400 truthful and 400 deceptive) and 800 negative reviews (400 truthful and 400 deceptive) of 20 popular Chicago hotels. We use the phrase *positive reviews dataset* to refer to the part of the dataset that contains truthful and deceptive reviews that promote a product. Likewise, we use the phrase *negative reviews dataset* to refer to the part of the original dataset that contains truthful and deceptive reviews that demote a product.

We use Python's scikit-learn package to extract the features from the review dataset with LDA. We then use the extracted features to build a model using SVM. The generated model is then used for measuring performance against the dataset. We used 5-fold cross-validation for evaluating the model's performance.

For the LDA procedure, we set the number of components to 2, 3, or 4 because a review can be "truthful or deceptive", "truthful, deceptive, or neutral", or "truthful positive, truthful negative, deceptive positive, or deceptive negative". We started from 50 top words for each topic. Then, for each new experiment we increased the number of top words by 50 and continued this process for up to 1000 top words per topic. A sample output from LDA with 2 components and 20 top words is as follows:

Topic #0: *hotel room great chicago stay location nice staff stayed service rooms good clean night bed just got comfortable michigan bathroom*

Topic #1: *hotel chicago stay room staff great rooms service stayed time place like friendly recommend business city definitely comfortable experience beautiful*

Below are samples of two reviews with the words that appear in either *Topic #0* or *Topic #1* shown in bold. The first review is a sample of a positive truthful review while the second is a sample of a positive deceptive review.

1. I was completely blown away by this **hotel**. It was magnificent. I got a **great** deal and I am so happy that I stayed here. Before arriving I was nervous as I had read a few bad reviews about the impact the renovation was having on peoples **stay**, for example very noisy. However, whilst the renovation was still going on and the gym was not open nor the restaurant, it made no difference to me. My **room** was huge, **bathroom** was spacious with excellent water pressure, **bed** was perfect and the view was amazing. Hotel is so close to the **great** shops of Magnificent Mile, plus a comfortable walking distance to Hancock tower and Millennium Park.
2. After recent week **stay** at the Affinia Hotels, I can definitely say i will be coming back. They offer so many in **room** amenities and services, Just a very comfortable and relaxed place to be. My most enjoyable experience at the Affinia **Hotel** was the amazing customization they offered, I would recommend Affinia hotels to anyone looking for a nice place to **stay**.

PERFORMANCE RESULTS

In this section, we report the deception review detection performance with our proposed approach. In **Figure *2***, we present the effect of the number of components (i.e. topics) on the accuracy. We used the whole dataset for this experiment. We observed similar accuracy for deception detection with number of topics set to 2, 3, or 4. Now, if we select the number of top words in a topic to be 50, then for 2, 3, and 4 topics, we get 100, 150, and 200 top words (i.e. features for SVM), respectively. However, 2 topics provides higher accuracy than 3 or 4 topics for the same number of top words, as per **Figure *2***. Therefore, we decided to use n_component = 2 (i.e. 2 topics), because this means fewer features and reduced dimensionality for SVM in addition to better accuracy.

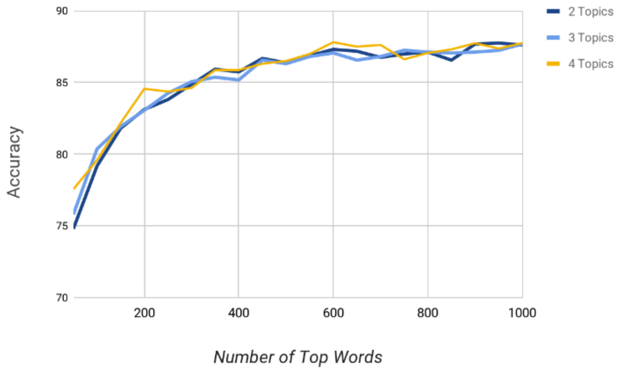

Figure 2. Accuracy of deception detection.

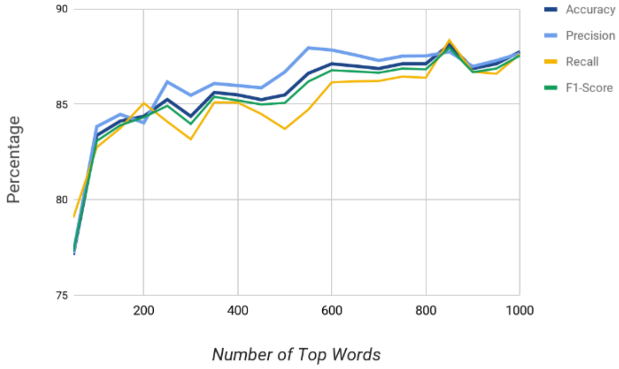

Figure 3. Accuracy, Precision, Recall & F1-Score of *truthful reviews* of spam review detection using topic modeling and SVM on the positive reviews dataset.

Figure 3 and **Figure *4*** show the performance of our model on the positive reviews datasets. The accuracy varies from 84% to 87% for top words between 200 and 1000 words. In comparison, Ott et al. [1] achieved 89% but they considered all the available words of the dataset as features with a dimensionality of approximately 5476. Whereas, our model's accuracy is based on a maximum dimensionality of 2000. Our model also performs similarly with the negative reviews dataset as can be seen in **Figure *5*** and **Figure *6***. With regards to precision, recall, and F1-score for truthful reviews and deceptive reviews, Figure 3 and Figure 4 show that increasing the number of top words contribute positively to higher precision, recall, and F1-score for our model. This validates that our proposed model is working properly since accuracy, precision, recall, and F1-score results are consistent. Similarly, in Figure 5 and Figure 6, we show the effectiveness of our model on negative reviews.

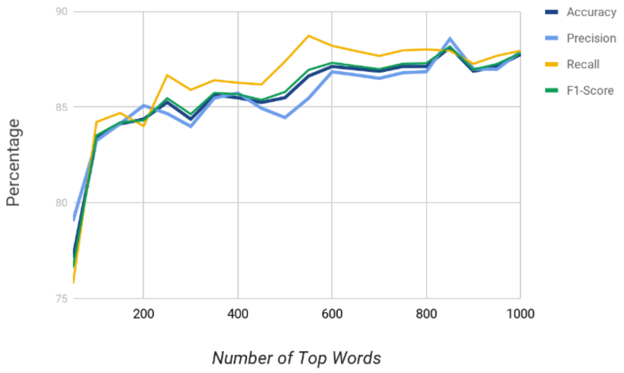

Figure 4. Accuracy, Precision, Recall & F1-Score of *deceptive reviews* of spam review detection using topic modeling and SVM with positive reviews dataset.

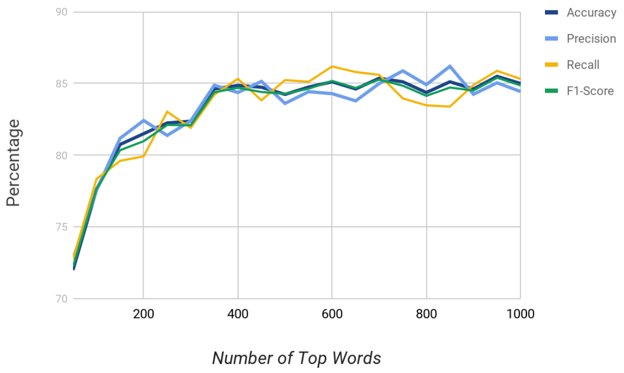

Figure 5. Accuracy, Precision, Recall & F1-Score of *truthful reviews* of spam review detection using Topic Modeling and SVM with negative reviews data set.

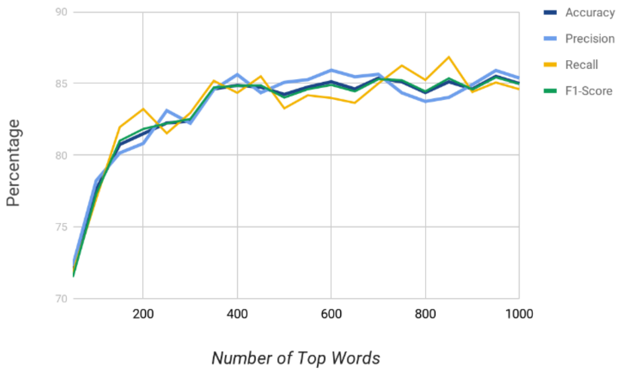

Figure 6. Accuracy, Precision, Recall & F1-Score of *deceptive reviews* of spam review detection using Topic Modeling and SVM with negative reviews data set.

CONCLUSION

In this paper, we presented an approach for spam review detection using topic modeling. Our approach reduces the number of features and hence dimensionality for SVM. Our proposed model detects deceptive positive and deceptive negative reviews with over 84% accuracy. The accuracy we achieved is comparable with other approaches that use all the words in a dataset as features. The research results show that a fraction of all words in a document is sufficient to tag a review as truthful or deceptive.

REFERENCES

1. Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, (HLT '11), Vol. 1, 309-319.
2. Myle Ott, Claire Cardie, and Jeff Hancock. 2012. Estimating the prevalence of deception in online review communities. In *Proceedings of the 21st international conference on World Wide Web* (WWW '12), 201-210.
3. Michael Luca and Georgios Zervas. 2015. Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud. *Management Science*, 62(12): 3412-3427.
4. Myle Ott, Claire Cardie, and Jeffrey T. Hancock. 2013. Negative Deceptive Opinion Spam. In *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics: human language technologies*, 497-501.
5. Charles F. Bond, Jr. and Bella M. DePaulo. 2006. Accuracy of Deception Judgments. *Personality and Social Psychology Review*, 10(3): 214-234.
6. Bing Liu. 2015. *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*. Cambridge: Cambridge University Press.
7. Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining* (WSDM '08), 219-230.
8. Fangtao Li, Minlie Huang, Yi Yang, and Xiaoyan Zhu. 2011. Learning to identify review spam. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence*, 2488-2493.
9. Jiwei Li, Myle Ott, Claire Cardie, Eduard Hovy. 2014. Towards a general rule for identifying deceptive opinion spam. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, 1566–1576.
10. Somayeh Shojaee, et al. 2013. Detecting deceptive reviews using lexical and syntactic features. In *Proceedings of Intelligent Systems Design and Applications (ISDA)*, 53–58.
11. Ahmad S. J. Abu Hammad. 2013. An Approach for Detecting Spam in Arabic Opinion Reviews. Master's Thesis. Islamic University of Gaza.
12. Huayi Li, et al. 2017. Bimodal Distribution and Co-Bursting in Review Spam Detection. In *Proceedings of the 26th International Conference on World Wide Web* (WWW '17), 1063-1072.
13. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning Research*, 3(Jan): 993-1022
14. Jiwei Li, Claire Cardie, and Sujian Li. 2013. Topicspam: a topic-model based approach for spam detection. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2: 217-221.
15. Thorsten Joachims. 1998. Text categorization with Support Vector Machines: Learning with many relevant features. In *ECML-98. Lecture Notes in Computer Science*, vol 1398.