Online Review Spam Detection by New Linguistic Features

Amir Karam, University of Maryland Baltimore County
Bin Zhou, University of Maryland Baltimore County

Abstract
With the fast growing and importance of online reviews, malicious users start to abuse the online review websites and deliberately post low quality, untrustworthy, or even fraudulent reviews, which are typically referred to as "spam reviews". Many existing studies on review spam detection are based on classification models. Features such as the number of verbs used in the reviews are commonly used to construct the spam review classification model. Surprisingly, many linguistic features of users' reviews have not been thoroughly considered for review spam detection. In this paper, we focus on different types of linguistic features and evaluate their performance on detecting spam reviews. Our empirical evaluation conducted on a spam review benchmark dataset validated the proposed features significantly improve the performance of online review spam detection, reaching more than 93% accuracy.
Keywords: Spam Detection; Online Review; Classification; Content Features; Linguistic
Citation: Karami, A., Zhou, B. (2015). Online Review Spam Detection by New Linguistic Features. In *iConference 2015 Proceedings*.

Contact: amir.karami@umbc.edu

1 Introduction

The recent development of Web 2.0 has generated plentiful of user-created content. Among various types of user-generated data on the web, reviews about businesses, products, or services written by the users are becoming more and more important due to the word-of-mouth effect and their impact on influencing consumer's purchase decisions. However, with the increasing popularity of online review websites such as TripAdvisor[1] and Yelp[2], malicious users start to abuse the convenience of publishing online reviews and deliberately post low quality, untrustworthy, or even fraudulent reviews. Such "spam reviews" can result in significant financial gains for organizations and individuals, and meanwhile lead to negative impact on their competitors. For example, a few recent studies have reported a new category of business which hires people to write positive reviews for some companies to attract users' awareness and increase the profits[3].

Spam reviews undoubtedly reduce the quality of reviews. They may even mislead users to make wrong purchase decisions. Therefore, there is a great demand to detect spam reviews thoroughly on the web. Recently, several existing studies investigated various machine learning techniques to automatically construct spam classification models based on specific features (Ott, Choi, Cardie, & Hancock, 2011; Mihalcea & Strapparava, 2009). For example, Ott *et al*. (2011) investigated the lexical features such as the frequency of verbs used in the reviews. Their results indicated that those lexical features are useful for building a spam classification models for spam reviews.

Despite the usefulness of many lexical features, the classification performance of existing models for spam review detection is still far from satisfactory. An interesting direction to explore is whether some other features, such as users' sentiments and feelings, as well as many other linguistic features of reviews would be incorporated into the classification model. In this paper, we focus on various types of linguistic features in users' reviews such as the number of pronouns, psychological features such as the affective processes, current concerns such as degree of leisure, spoken features such as degree of assent, and punctuation such as number of colons. We evaluated the spam classification performance by considering more than 40 different classification algorithms on a spam review benchmark dataset. Our experimental results verified that the combination of linguistic features with some others (e.g., the frequency of words) could improve the detection performance over the state-of-the-art method, reaching more than 93% accuracy.

[1] http://tripadvisor.com
[2] http://www.yelp.com
[3] http://www.cnet.com/news/fake-reviews-prompt-belkin-apology/

The remainder of the paper is organized as follows: in Section 2, we briefly review some relevant studies. In Section 3, we present the set of features generated from Linguistic Inquiry and Word Count tool in detail. An empirical study was conducted to verify the effectiveness of our method and the results are provided in Section 4. Finally, we present a summary, limitations, and future directions in Section 5.

2 Related Work

The spam detection has different domains including Web (Castillo et al., 2006), Email (Chirita, Diederich, & Nejdl, 2005), and SMS (Karami & Zhou, 2014a, 2014b). The problem of opinion spam was introduced by investigating supervised learning techniques to detect fake reviews (Jindal & Liu, 2008). Lim et al. (2010) tracked the behavior of review spammers and found some specific behaviors such as targeting specific products or product groups in order to maximize their impact (Lim, Nguyen, Jindal, Liu, & Lauw, 2010). Using machine learning techniques is one of the popular approaches in online review spam detection. Some examples are employing standard word and part-of-speech (POS) n-gram features for supervised learning (Ott et al., 2011), using a graph-based method to find fake store reviewers (Wang, Xie, Liu, & Yu, 2011), and using frequent pattern mining to find groups of reviewers who frequently write reviews together (Mukherjee, Liu, Wang, Glance, & Jindal, 2011; Mukherjee, Liu, & Glance, 2012).

3 Feature Encoding

In websites with online reviews, there are a set of k online reviews $R = \{r_1, ..., r_k\}$. Each message consists of words, numbers, etc. R can be about any topic. The online review spam detection problem can be described as the prediction of whether r_i is a deceptive positive review using classifier c. A classifier predicts whether or not r_i is a deceptive positive reviews.

$$c : r_i \rightarrow \{deceptive - reviews, truthful - reviews\}$$

For classification, we need to extract a set of n features $F = \{f_1, ..., f_n\}$ from R. We outline two groups of features to detect deceptive opinion spam trained on the dataset.

3.1 Developed-LIWC Features

Linguistic Inquiry and Word Count (LIWC)[4] is a tool that analyzes 80 different features of text including linguistic processes such as number of pronouns, psychological processes such as affective processes, current concerns such as degree of leisure, spoken features such as degree of assent, and punctuation such as number of colons. While the function of LIWC is not text classification, we can see each output of LIWC as a feature. Although all or some of LIWC features have been used in other research for spam detection (Ott et al., 2011; Zhou, Burgoon, Twitchell, Qin, & Nunamaker Jr, 2004; Karami & Zhou, 2014b), we also incorporated an additional 224 features, *LIWC*$^+$, based on various combinations of the raw features collected from LIWC. For instance, we derived the relative polarity by examining the difference between the positive and negative feelings scores. Table 1 shows a sample of our new features. To the best of our knowledge, these kinds of linguistic and sentiment features have not yet been explored for online review spam detection. We call our extension of LIWC *"Developed-LIWC (D-LIWC)"* with 304 features including both LIWC and LIWC$^+$ features.

The rate of "Verb" score to "All Words" score
The rate of "Parentheses" score to the "All Punctuation" score
The rate of "Negative Feelings" score to the all "Affective Feelings" score
The difference between the score of words related to "Family" and the score of words related to "Humans"
The difference between the score of words related to "Leisure" and the score of words related to "Money"

Table 1: A Sample of LIWC$^+$ Features

[4]http://www.liwc.net/

3.2 Unigram Features

In text categorization, one of popular techniques for feature extraction is n-gram including Unigrams, Bigrams, and Trigrams. The second strategy for feature extraction in this research is bag-of-words or Unigrams. We consider this category of n-grams as one of popular features in text categorization. By using this strategy, we can find frequency of each word for reviews. We will investigate bigrams and trigrams in our future work.

4 Experimental Results

In this section, we discuss our empirical evaluation of our approach against the best result in the literature using more than 40 classification algorithms from Random Forest to Naive Bayes. In the experiment, we use Weka[1] for classification evaluation and leverage one available dataset in this research. This dataset[2] is a labeled corpus with 800 positive reviews including 400 truthful positive reviews and 400 deceptive positive reviews. For all classification algorithms, we use 80% of data for training and 20% for testing, with 5-fold cross validation.

4.1 Classifier Performance

In order to determine whether a review is a deceptive-review or truthful-review, we adopt supervised machine learning algorithms. The goal of classification is to classify a review into deceptive-reviews or truthful-reviews. For evaluating the performance of spam detection, we measure F-measure and Accuracy. Features from the approaches in section 3 are used to train classification algorithms using the Weka machine learning toolkit. Among the classification algorithms Support Vector Machine (SVM) shows the best performance. This technique finds a high dimensional separating hyperplane between two groups of data. We use SVM to train two approaches with different combination of features, namely D-LIWC, LIWC, and Unigrams features.

One of the contributions of this paper is to use a feature selection approach to improve the performance of classifiers and also reduce the number of data features to avoid negative effects of the high dimension problem in analyzing large number of online reviews. We chose Chi-Square as the feature selection method, which is among the most effective methods (Yang & Pedersen, 1997). Table 2 presents the classification performance using different sets of the features in section 3.

	Number of Features	**Accuracy**	**F-Measure**	
			Deceptive Reviews	**Truthful Reviews**
LIWC	All Features (80)	76.8%	76.9	76.6
	Selected Features (60)	78.6%	79.1	78.1
D-LIWC	All Features (304)	78.72%	79.3	78.1
	Selected Features (125)	79.22%	79.8	78.6
Unigrams	All Features (4965)	88.4%	88.6	88.2
	Selected Features (2000)	89.11%	89.4	88.8
Unigrams+LIWC	All Features (5044)	85.98%	86.5	85.4
	Selected Features (2000)	88.86%	89.2	88.5
Unigrams+D-LIWC	All Features (5268)	90.62%	91.3	89.8
	Selected Features (2500)	**93.42%**	**93.8**	**93.1**
Ott et al. (2011)	All Features (N/A)	89.8%	89.8	89.8

Table 2: Classifier Performance

We observe that the SVM classifier outperforms the performance of Ott et. al (2011) using Bigrams and LIWC as the best performance in the literature (Table 2). In addition, we explore different number of features to reduce the number of features for handling high sparse dimension and complexity problems in a large number of reviews. In table 2, there are two sets of features: the first one is all features in the categories

[1] http://www.cs.waikato.ac.nz/ml/weka/
[2] http://myleott.com/op_spam/

such as LIWC and the second one is the best performance over different sets of top selected features. For example, the accuracy of all 80 LIWC features is 76.8% and the accuracy of selected 60 features by Chi-Square is 78.6%. We also track the proportion of the features in top n features from $n = 10$ to $n = 100$ features. The result shows that most of top features belong to LIWC$^+$ (Figure 1). For example, the proportion of features including LIWC$^+$, Unigrams, and LIWC in top 100 features are 56%, 26%, and 18%.

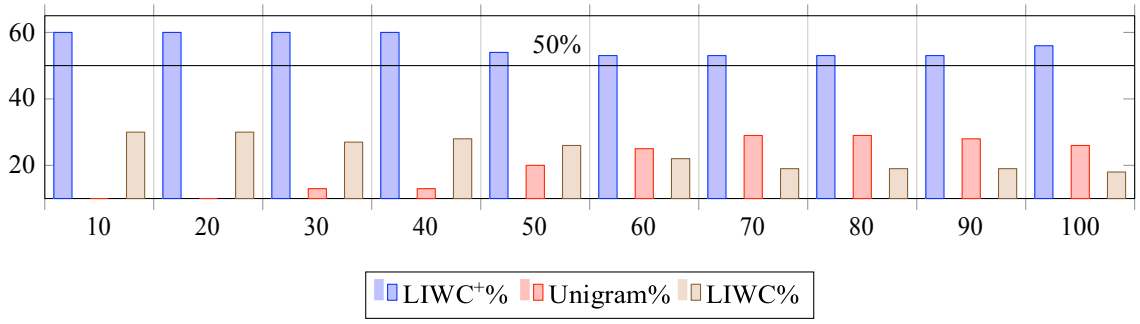

Figure 1: Features' Proportion in Top n Features

5 Conclusion

The recent surge of Web 2.0 makes the emerging communication media such as online review websites particularly attractive for malicious users. The challenge of detecting spam reviews in those websites is mainly due to the huge size of review data and its difficulty in detection. Existing research on online review spam detection has focused on word statistics. Surprisingly, not much work has been conducted to examine deeper semantic categories of text expressions. In this paper, we proposed to employ categories of lexical semantic and linguistic features in the detection of online spam reviews. Our experiment results showed that by incorporating many linguistic features of reviews, the detection performance of spam reviews can be greatly improved, comparing with the state-of-the-art methods.

There are several interesting directions to explore in the future including exploring our approach on negative opinions, as well as opinions coming from other domains such as product reviews. In addition, we are interested in exploring latent semantic features such as grouping the words with similar meaning using topic modeling.

References

Castillo, C., Donato, D., Becchetti, L., Boldi, P., Leonardi, S., Santini, M., & Vigna, S. (2006). A reference collection for web spam. In *Acm sigir forum* (Vol. 40, pp. 11–24).

Chirita, P.-A., Diederich, J., & Nejdl, W. (2005). Mailrank: using ranking for spam detection. In *Proceedings of the 14th acm international conference on information and knowledge management* (pp. 373–380).

Jindal, N., & Liu, B. (2008). Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining* (pp. 219–230).

Karami, A., & Zhou, L. (2014a). Exploiting latent content based features for the detection of static sms spams. In *Proceedings of the 77th annual meeting of the association for information science and technology (ASIST)*.

Karami, A., & Zhou, L. (2014b). Improving static SMS spam detection by using new content-based features. In *20th americas conference on information systems (AMCIS)*.

Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., & Lauw, H. W. (2010). Detecting product review spammers using rating behaviors. In *Proceedings of the 19th acm international conference on information and knowledge management* (pp. 939–948).

Mihalcea, R., & Strapparava, C. (2009). The lie detector: Explorations in the automatic recognition of deceptive language. In *Proceedings of the acl-ijcnlp 2009 conference short papers* (pp. 309–312).

Mukherjee, A., Liu, B., & Glance, N. (2012). Spotting fake reviewer groups in consumer reviews. In *Proceedings of the 21st international conference on world wide web* (pp. 191–200).

Mukherjee, A., Liu, B., Wang, J., Glance, N., & Jindal, N. (2011). Detecting group review spam. In *Proceedings of the 20th international conference companion on world wide web* (pp. 93–94).

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (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-volume 1* (pp. 309–319).

Wang, G., Xie, S., Liu, B., & Yu, P. S. (2011). Review graph based online store review spammer detection. In *Data mining (icdm), 2011 ieee 11th international conference on* (pp. 1242–1247).

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. In *Icml* (Vol. 97, pp. 412–420).

Zhou, L., Burgoon, J. K., Twitchell, D. P., Qin, T., & Nunamaker Jr, J. F. (2004). A comparison of classification methods for predicting deception in computer-mediated communication. *Journal of Management Information Systems*, *20*(4), 139–166.

Table of Figures

Table of Tables