RLOSD: Representation Learning Based on Opinion Spam Detection

Zeinab Sedighi
Department of Computer Engineering, Faculty of Computer and Electrical Engineering, University of Kashan, Kashan, I.R.Iran
sedighi.63@gmail.com, sedighi.63@grad.kashanu.ac.ir

Hossein Ebrahimpour-Komleh
Department of Computer Engineering, Faculty of Computer and Electrical Engineering, University of Kashan, Kashan, I.R.Iran
ebrahimpour.kashanu@gmail.com

Ayoub Bagheri
Department of Computer Engineering, Faculty of Computer and Electrical Engineering, University of Kashan, Kashan, I.R.Iran
ayoub.bagheri@gmail.com
a.bagheri@kashanu.ac.ir

ABSTRACT

Nowadays, by vastly increasing in online reviews, harmful influence of spam reviews on decision making causes irrecoverable outcomes for both customers and organizations. Existing methods investigate ~~for~~ a way to contradistinction between spam and non-spam reviews. Most algorithms focus on feature engineering approaches to expose an accommodation of data representation. In this paper we propose a decision tree-based method to reveal deceptive reviews from trustworthy ones. We use unsupervised representation learning along with traditional feature selection methods to extract appropriate features and evaluate them with a decision tree. Our model takes data correlation into consideration to opt suitable features. The result shows the better performance in detecting opinion spam, comparing most common methods in this area.

Keywords

Opinion spam detection, Representation learning, Natural language processing, Review mining, PCA.

1. INTRODUCTION

The more use of the Internet and social media increases, the greater impacts of users and customer's opinion will be. Hence, so many companies supply the possibility of survey of their products for customers. These comments could contain very important information and thus can easily affect people's decisions which have significant impacts in the e-commerce, social and political areas. Organizations and different companies can use the user's feedback comments in estimating a community belief on a subject or a particular product, in marketing, designing and sailing their product. On the other hand, other customers and users also can benefit from these comments to decide purchasing a product. This value and importance has caused some people or organizations abuse this feature and create fake reviews to entice other users in order to achieve their goals. In this way, they make users decisions compatible with their own goals. These fake reviews called spam. Accordingly, blind trust on comments may lead to make wrong decision, thus this issue is very important for organizations and users. Spam reviews have a negative impact on decisions and will follow declining confidence of users and customers in companies and organizations which in many cases such business could cause irreparable consequences. From political and social points of view, opinion spam reviews also lead to opinion deviation of groups of users. Since all comments on the web and social networks are not reliable, the need to develop and to use techniques identifying spam comments is indispensable.

In this study, in two successive phases of feature engineering and feature reduction, relevant features are elected and excessive and redundant terms are eliminated from the feature space. Application of these machine learning approaches enhances the overall model performance. At first step, a pre-processing level is accomplished in which features extracted from the document using bigram, POS tagging and TF_IDF methods. Therefore, terms with high frequency are chosen from the whole document. Then, Modified Mutual Information (MMI) along with PCA are used to opt the significant and important terms within former features to diminish the data dimension, in the second step. Consequently, the model complexity and execution time for detecting spam reviews are reduced. To filter out spam reviews from non-spam ones, a decision tree classifier is used which ranks features using Information Gain measure. The results illustrate that the RLOSD model is able to separate spam reviews from the whole corpus effectively. This claim is proven by precision, recall and F-measure which compare the RLOSD with SVM, naïve Bayes and Log Regression results.

This paper structure is organized as follows. Section 2 overviews related works on representation learning and detecting spam reviews. Sections 3 explains the methodology we used in this study. Section 4 represents the discussion and finally, section 5 concludes the paper.

2. RELATED WORKS

Here we mention to related works in two subsections in short. First we explain representation learning tasks and then review opinion spam detection literature.

2.1. Representation learning

The efficiency of machine learning models highly depends on which feature they use. Representation learning, feature learning or even feature engineering techniques tries to learn features which can fairly represent raw data. Discovering useful representations automatically is the aim of representation learning. These methods can be divided in two categories: supervised and unsupervised feature learning[1].

Supervised feature learning approaches are train using labeled data like artificial neural network models. Unsupervised ones are learned by unlabeled data such as clustering techniques, Independent Component Analysis and Principal Component Analysis.

2.2. Opinion spam detection

Today's, Internet users can express their opinions on the web about various topics including their purchase experiences. These comments contain useful information for other users, developers and organizations [۲]. Increasing use of these reviews caused some people and organizations abuse them and provide fake reviews to their profits. Due to changing people's opinions by review spam, identifying and presenting strategies to prevent opinion spam is considered more, recently. This article seeks to provide a solution to improve detection of spam reviews.

After the initial preliminary solution presented by Jindal [۲], many techniques were raised to detect spam reviews and drew much attention. Dixit et al. [۳] have classified reviews into three groups: 1) spam comments, 2) commercial comments and 3) non-spam comments. The first category, review spam, has attracted a lot of attention in the sense that detecting these comments even manually, is difficult.

There are many challenges in detecting spam in which different methods investigate each of them and combat it[4]. All these methods are looking for a way to distinguish between spam and non-spam data. Different algorithms are used for this separation divided into supervised, unsupervised and semi-supervised methods. In order to discern review spam, supervised learning methods use classification techniques. Jindal et al. [۲] compared Naïve Bayes, logistic regression and SVM methods after feature extraction and opinion summarization. Ott et al. [۵] tested SVM and Naïve Bayes on their self-generated artificial data sets and originated features using POS tags and LIWC. Li et al. [۶] applied a data set contains three cross domain reviews to avoid algorithms data-dependency to a specific domain. They performed SVM and SAGE for its classification.

Although labeled data sets for spam reviews doesn't exist in the web, however some, like Ott's data set, are built artificially and used by many [۷-۹]. Since there are many unlabeled data around the world, these data occasionally are used by some people. Reymond et al. [۱۰] presented an unsupervised based model which bring in semantic language techniques. Even few studies were done on semi-supervised methods to gain little labeled data for clustering all the data sets. PU-learning was introduced by liv et al. [۱۱] is an approach which was used by Mentes et al. [۱۲] to detect spam reviews.

Utilizing unused representation learning techniques and significant feature engineering methods to extract effective features of spam data with high impact is the contribution of our article. We propose a model which learns representations of input data and affords detecting spam reviews using them. The model consists of two steps; at first we extract beneficial representations of review documents. To this end we employ statistic characteristic of data and compute correlation among words and phrases inside documents. Then at the second step we utilize these representations and analyze reviews. Using this procedure, we outperform traditional models and achieve better accuracy.

3. RLOSD MODEL

The proposed model flowchart is shown in Figure 1. To classify as spam or non-spam, a review should be coded into a representation form which could be processed by a learning algorithm. This is done by applying feature learning algorithms. One of the most popular representation in text processing is producing a vector of words as features. A collection of TF-IDF, POS tagging, n-grams, MMI and PCA is used to select suitable features, in this article. We detail its procedure in the following sections.

3.1. Pre-processing

3.1.1. Remove stop-words

Words contain low importance such as *the, as, a and an* in text processing are called stop-words. The aim is to filtered out them from the document corpus since they are immense with useless information. Removing these frequent words is important as they may mislead the classification procedure.

3.1.2.Stemming

The process of reducing the derived words to their word root is called stemming in information retrieval. It usually maps related words to the same root, even if they were different words. To this end, an algorithm should apply for suffix stripping which can significantly treat by complex suffix. In this study, stemming is accomplished based on Porter algorithm [۱۳]. It handles complex suffix composed of modest suffix.

3.1.3. Part-of-speech tagging

The process of assigning description to a token is called tagging. It denominates POS tagging when the description refers to one of the part of speech. Part of speech are categories consisting nouns, adjectives, verbs, adverbs and etc. since suppling useful knowledge about a word and its vicinity, these categories are effective. Knowing the POS of a word gains a lot information about its neighbors and

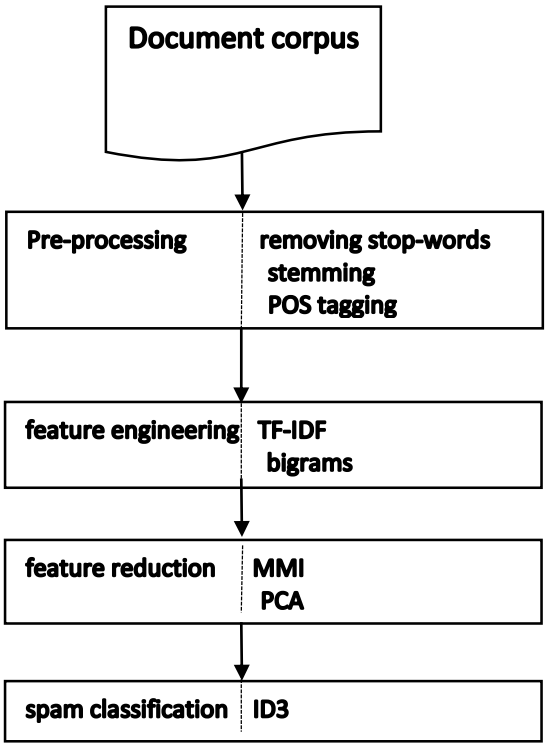

Figure 1. The RLOSD model steps

its syntactic structure. These aspects make POS tagging an important feature selection technique in text processing.

3.2. Feature learning

3.2.1. Term Frequency–Inverse Document Frequency

The term frequency inverse document frequency known as TF-IDF for short, is a statistic that demonstrates how effective a word is in a document [14]. It is a weighting factor which is applied in text mining. This can compute by comparing the frequency of application of a term inside a particular document versus the whole data set or corpus. Since it is possible that numerous documents include an identical word various times, TF-IDF contracts this value by the frequency of the term in the entire corpus called Inverse Document Frequency.

The TF-IDF is composed of two terms; the former calculates normalized term frequency ~~by~~ computing the number of times a term occurs in a document divided by the whole number of the document terms. The latter is inverse document frequency which measures whether the term is common or scarce in the entire corpus documents. It calculates the logarithm of the number of entire documents divided by the number of the ones containing the term. There are several ways to define both statistics. One them illustrates below:

$$tf(t,d) = \frac{f_{t,d}}{\max\{f_{t',d} : t' \in d\}}$$

Where $f_{t,d}$ is the frequency of term t in document d.

$$idf(t,d) = \log(1 + \frac{N}{n_t})$$

Where N and n_t are number of whole documents in the corpus D and number of documents containing term t, respectively.

Words appear repeatedly in specific documents can be obtained by multiplication of these values which leads high values. Instead, terms appear frequently in all documents lead low value.

$$tf\_idf(t,d,D) = tf(t,d) \times idf(t,d)$$

3.2.2. n-grams

Allocating probability to a sequence of words are known Language Models [15]. Their simplest model which produce a sequence of n most probable items in text or speech processing is called n-gram. Predicting the sequence is based on statistical property of a given document. A sequence of one word is referred to as unigram, of two words is known as bigram and of three words is trigram. One of the n-grams benefits is that it is independent of any language. However, using n-gram for feature extraction might lead to numerous numbers of possible sequences on a large dataset. In such circumstances just some of n-grams could have the ability of discriminating.

3.3. Feature reduction

Because of huge amount of features generated in the preceding step, it's necessary to extract informative features which plays an important role in detecting deceptive reviews. As there exist few irrelevant features in text processing, PCA is a remarkable technique to extract suitable features. RLOSD employs MMI along with PCA to cull better representation from document.

3.3.1. Modified Mutual Information

Since Mutual Information measures the probability of a feature with a class and doesn't measure the co-occurrence of a feature and other features and classes, Bagheri and Saraee [16] introduce a Modified version of Mutual Information as MMI which consider all possible combinations of co-occurrences of a feature and class label.

3.3.2. Principal Component Analysis

PCA is a feature representation learning method which can be used to reduce data dimensions. It illustrates similarities between input data and tries to find patterns in data. To represent a smaller set of variables PCA preserves the data variance as much as possible. It produces eigenvectors correspond to largest eigenvalues of the input data covariance matrix. These eigenvectors are features of

input data in which the data has the largest variations and called principal components. Actually the number of computed components in PCA is the same as the number of data variables, however, in most situations only the first few eigenvectors are considered to maintain the variance of whole data.

The first component in PCA considers maximum amount of variance in variables and demonstrate that it correlates with further amount of data. Second component considers maximum amount of variance that not considered by the preceding component. It shows that it correlates with some variables which don't have solid correlation with the first component. It indicates that obtained components in PCA are uncorrelated with each other and the correlation between them is zero. The procedure carries on consequently [17].

3.4. Spam classification

After employing feature reduction stage, to distinguish between spam and non-spam reviews, it's essential to apply a classifier for the discrimination. In RLOSD decision tree is use to this end.

3.4.1. Decision tree methods

A decision tree model has a predictive structure that maps an input item to a target value. Decision tree creates a hierarchical division of the given document corpus using different text features. It determines the partition to which ~~is~~ most likely an input data belongs. To this end, a condition is used on the data attributes to divide the data space hierarchically. It can be used to visually represent decision and decision making.

These algorithms make a tree in a top-down manner which select a variable to split the tree at each step. Different algorithms use different measures to select an appropriate split variable such as Gini index, Information gain, Gain ratio ~~and~~ etc.

A decision tree procedure usually has two general steps. At the first step, the tree grows greedily until it reaches the leafs. To avoid overfitting in training data, the tall and un necessary branches are eliminated from the tree in the second step.

We apply C4.5 Decision tree in this study with information gain as split measure.

4.EXPERIMENTS AND RESULTS

Experiments are carried out by RLOSD model first on Yelp data set then on the dataset of Li et al. [7] which consists of reviews in three domains. In this section we introduce the data sets and due to the accuracy comparison in text categorization, RLOSD is compared with logistic regression, naïve Bayes and Support Vector Machine methods. These methods are employed separately to the datasets in which the feature dimension reduced by PCA.

4.1. Yelp data set

Yelp's website presents a source of trade reviews which has web pages assigned to some places such as restaurants or schools, where its users can give commands or reviews on their products and services and rate them from one to five stars. It has surveyed that more than 20% of Yelp's reviews are deceptive [18] like other reviews on the Internet. Yelp website does not delete suspicious reviews but locates them in a list, which is public and available, however, avoids presenting them on the business~~es'~~ pages.

This challenge is still open problem. In this study, we use a supervised approach to utilize Yelp's filtered reviews. These are restaurant reviews which gathered by [19].

We examine the proposed model for text categorization on yelp dataset to examine its performance, dimensionality reduction and classifier techniques.

4.2 Three domain dataset

This dataset consists of reviews in three domains named Hotel, Restaurant and Doctor. Data in all domains is separated in two groups deceptive and truthful reviews. In each of these domain, truthful reviews ~~are~~ gathered from customer reviews and spam reviews are obtained from Turker and Employee who have information about domains. The Specifications of the dataset is demonstrated in table 1.

Table 1. specifications of dataset

Domain	**Turker**	**Employee**	**Customer**
Hotel	800	280	800
Restaurant	200	120	400
Doctor	200	32	200

4.3. Preprocessing step

The procedure of pre-processing starts by removing the stop words and diminishes the amount of useless words. Then, stemming is applied by the Porter algorithm [20]. After all, to distinguish role of the words, POS tagging is applied.

4.4. Feature engineering and feature reduction step

To extract helpful features for classifying reviews, feature engineering techniques are required. In this article, bigrams and TF_IDF are applied to extract more repetitive words in the document.

After feature extraction phase, combating huge amount of features is needed. Utilizing numerous amount of features in training step leads to overfitting. To avoid this obstacle, we propose a feature reduction approach to remove unnecessary features. Since in text categorization issue, there are only very few irrelevant features, we employ MMI along with PCA approach to deduce feature space, prevent overfitting and reduce time complexity.

One of the most notable steps in the PCA procedure is estimating the number of principal component. It should be select among all the principal components in order to represent the data as well. Various criteria are used to estimate the optimal number of principal components. We address cross validation technique to deal with it.

4.5. Classification step

After feature reduction via PCA, to separate spam reviews from non-spam ones, a classifier is needed. We apply four classifiers and compare their results with each other. Some authors [21] use Information Gain metric to reduce feature space which results in enhancing the performance of model. Here, we propose to apply ID3 which evaluates variables and features using the mentioned metric for classification. To proving our claim, we compare this model against other common methods in text categorization.

4.6. Evaluating RLOSD performance

To evaluate the performance of model we employ precision, recall and F-measure. These are most common criteria used for evaluate the efficiency of text classifier modes which we use for C4.5, SVM, Log Regression and Naïve Bayes classifiers. Precision is the deduction retrieve documents which are relevant to the retrieve documents.

$$precision = \frac{tp}{tp + fp}$$

, where tp and fp are true positive and false positive respectively.

Recall is the deduction relevant retrieve documents to the relevant documents.

$$recall = \frac{tp}{tp + fn}$$

, where fn represent false negative value.

The F-measure is a kind of harmonic mixture of the precision and recall.

$$F = 2 \times \frac{precision\ .\ recall}{precision + recall}$$

4.6.1 Evaluation on Yelp dataset

Table 2. results for review spam detection

models	precision	recall	F-measure
SVM	71.72	73.43	72.56
Naïve Bayes	64.32	65.83	65.06
Log Regression	63.74	65.56	64.64
RLOSD	76.21	77.63	76.91

Table 2 illustrates the comparison between the proposed model and other common models used in review spam detection on Yelp dataset. As seen in the table, the highest accuracy is obtained when we use Information Gain inside the classifier.

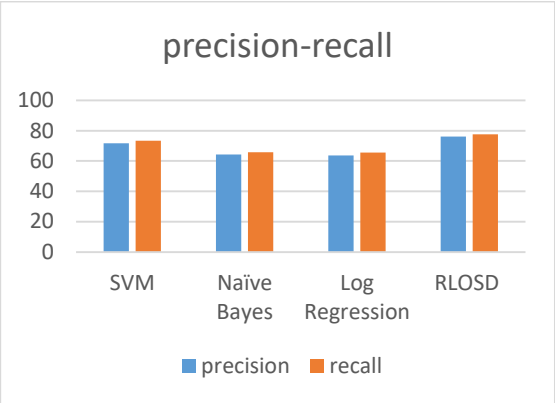

Figure 2. precision-recall in review spam detection

A precision-recall chart is shown in figure 2. As mentioned before, precision is the fraction of retrieved terms which are relevant while recall is the fraction on relevant terms which are retrieved. Figure 3 displays a comparison between four models F-measure on Yelp dataset. Here, we depict the state of matching by precision and this matching convergence by recall. High precision and high recall show the effectiveness of RLOSD model.

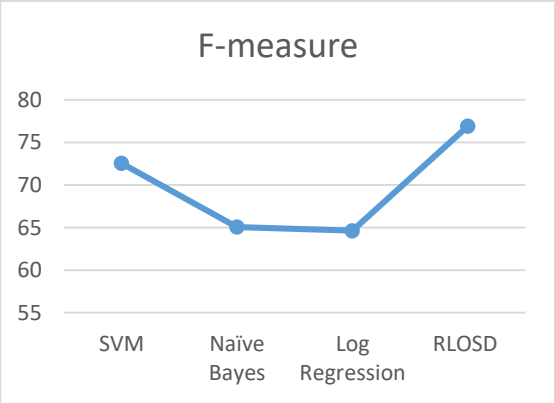

Figure 3. F-measure comparison on Yelp dataset

SVM sometimes works well, but in our experiences in yelp dataset, it doesn't earn enough information for review classification, although it has high demands for dimension independency.

Despite Naïve Bayes popularity in text categorization, assuming independency among features in classification doesn't result well.

We try other decision tree techniques like random forests but in order to exist sparse dimensions in text processing, it doesn't appear suitable for review detection.

4.6.1 Evaluation on three-domain dataset

Table 3. F-measure comparison among review spam detection models on three-domain dataset

models	Hotel	restaurant	Doctor
SVM	66.9	80.1	73.4
Naïve Bayes	65.9	77.2	70.3
Log Regression	66.4	76.5	71.7
RLOSD	78.3	81.8	75.0

The comparison among various spam classifications is represented in table 3. It demonstrates that the result of RLOSD model learn the best patterns of data and gain better result in deceptive spam classification compared other existing models. The scores of F-measure are all far above the other mentioned methods. The results show the effectiveness of incorporating representation based techniques in deceptive review detection.

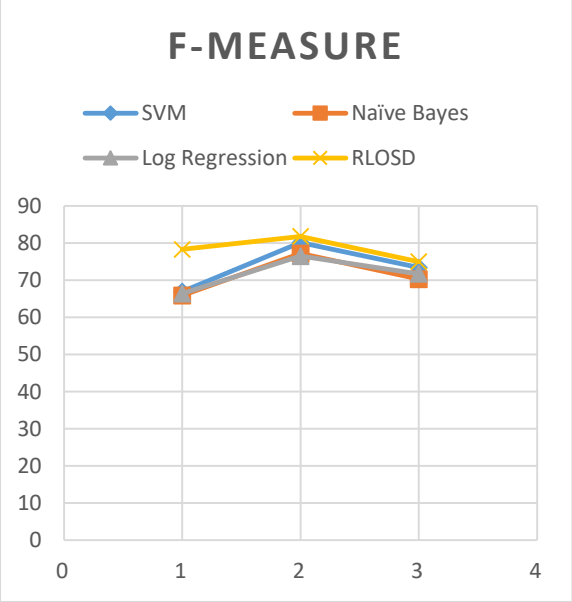

Figure 4. F-measure comparison on three-domain dataset

Figure 4 illustrates F-measure score of all models. It is obvious that the RLOSD has higher performance compared other methods.

The results of the proposed method are significant compared to the others. In addition, the fact that we have achieved this result using machine learning instead of using linguistic methods, is of particular importance.

5.CONCLUSION

In this paper, a combinational model composed of feature engineering and feature reduction phases is built to reduce the dimensionality of the feature space and eliminate redundant and irrelevant features which improve the performance of review spam detection. In the first step, a preprocessing process is done due to remove ineffective words from the document and prepared the data for analysis. This procedure is done by removing stop-words, stemming and POS tagging. In the next level, feature engineering techniques, e.g. TF_IDF and bigrams are applied to compose features. Since this level numerous features generates, PCA is used to remove unnecessary features and reduce the feature space. Finally, for distinguish between spam and non-spam reviews a classifier is employed in which using Information Gain to rank features and detect deceptive reviews. To prove this claim, RLOSD results is compared by other classifiers and it reveals that our model obtains higher performance.

REFERENCES

[1] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence,* vol. 35, pp. 1798-1828, 2013.

[2] R. Y. Lau, S. Liao, R. C. W. Kwok, K. Xu, Y. Xia, and Y. Li, "Text mining and probabilistic language modeling for online review spam detecting," *ACM Transactions on Management Information Systems,* vol. 2, pp. 1-30, 2011.

[3] M. Crawford, T. M. Khoshgoftaar, J. D. Prusa, A. N. Richter, and H. Al Najada, "Survey of review spam detection using machine learning techniques," *Journal of Big Data,* vol. 2, p. 1, 2015.

[4] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Systems with Applications,* vol. 42, pp. 3634-3642, 2015.

[5] B. Liu, *Web data mining: exploring hyperlinks, contents, and usage data*: Springer Science & Business Media, 2007.

[6] R. V. Bandakkanavar, M. Ramesh, and H. Geeta, "A survey on detection of reviews using sentiment classification of methods," *IJRITCC,* vol. 2, pp. 310-314, 2014.

[7] J. Li, M. Ott, C. Cardie, and E. H. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," in *ACL (1)*, 2014, pp. 1566-1576.

[8] J. Eisenstein, A. Ahmed, and E. P. Xing, "Sparse additive generative models of text," 2011.

[9] A. Mukherjee, V. Venkataraman, B. Liu, and N. Glance, "Fake review detection: Classification and analysis of real and pseudo reviews," UIC-CS-03-2013. Technical Report2013.

[10] A. Abbasi, Z. Zhang, D. Zimbra, H. Chen, and J. F. Nunamaker Jr, "Detecting fake websites: the contribution of statistical learning theory," *Mis Quarterly,* pp. 435-461, 2010.

[11] N. Jindal, B. Liu, and E.-P. Lim, "Finding unusual review patterns using unexpected rules," in *Proceedings of the 19th ACM international conference on Information and knowledge management*, 2010, pp. 1549-1552.

[12] F. Li, M. Huang, Y. Yang, and X. Zhu, "Learning to identify review spam," in *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, 2011, p. 2488.

[13] G. Fei, A. Mukherjee, B. Liu, M. Hsu, M. Castellanos, and R. Ghosh, "Exploiting Burstiness in Reviews for Review Spammer Detection," *ICWSM,* vol. 13, pp. 175-184, 2013.

[14] S. Robertson, "Understanding inverse document frequency: on theoretical arguments for IDF," *Journal of documentation,* vol. 60, pp. 503-520, 2004.

[15] F. Song and W. B. Croft, "A general language model for information retrieval," in *Proceedings of the eighth international conference on Information and knowledge management*, 1999, pp. 316-321.

[16] A. Bagheri and M. Saraee, "Persian Sentiment Analyzer: A Framework based on a Novel Feature Selection Method," *arXiv preprint arXiv:1412.8079,* 2014.

[17] N. O'Rourke and L. Hatcher, *A step-by-step approach to using SAS for factor analysis and structural equation modeling*: Sas Institute, 2013.

[18] S. Rayana and L. Akoglu, "Collective opinion spam detection: Bridging review networks and metadata," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 985-994.

[19] D. Kotzias, M. Denil, N. De Freitas, and P. Smyth, "From group to individual labels using deep features," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 597-606.

[20] P. Willett, "The Porter stemming algorithm: then and now," *Program,* vol. 40, pp. 219-223, 2006.

[21] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *Journal of machine learning research,* vol. 3, pp. 1289-1305, 2003.