Detecting Deceptive Reviews Using Lexical and Syntactic Features

Somayeh Shojaee*, Masrah Azrifah Azmi Murad†, Azreen Bin Azman‡, Nurfadhlina Mohd Sharef§ and Samaneh Nadali¶

Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
Selangor, Malaysia

Email: somayeh.shojaee@gmail.com, {masrah†, azreenazman‡, nurfadhlina§}@upm.edu.my, sm.nadeali@gmail.com¶*

***Abstract*—Deceptive opinion classification has attracted a lot of research interest due to the rapid growth of social media users. Despite the availability of a vast number of opinion features and classification techniques, review classification still remains a challenging task. In this work we applied stylometric features, i.e. lexical and syntactic, using supervised machine learning classifiers, i.e. Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) and Naive Bayes, to detect deceptive opinion. Detecting deceptive opinion by a human reader is a difficult task because spammers try to write wise reviews, therefore it causes changes in writing style and verbal usage. Hence, considering the stylometric features help to distinguish the spammer writing style to find deceptive reviews. Experiments on an existing hotel review corpus suggest that using stylometric features is a promising approach for detecting deceptive opinions.**

***Keywords*-Deceptive; Opinion; Lexical; Syntactic; Classification**

I. Introduction

Social media is source of information about users' behaviors and interests. There are obvious benefits for different parties such as companies or governments in understanding what the public think about their products and services. User opinion can have impact on sales of products, change of government policy, and etc. However, the widespread sharing and employing of user-contributed reviews have also increased worries about the reliability of them. Manually detecting and examining fake reviews and reviewers are not workable because of the problem of information overload.

As the number of reviews increases, different kinds of methods are established to improve the opinion mining tasks e.g. [3], [4], [5], [6], [7], [11], [15], [17] and [20]. Most of the existing works on review spam detection have focused on spotting review spam that can be detected by a human but currently detecting deceptive reviews is the concern of companies, governments and researchers. These kinds of reviews are not easily determined by a human reader, for instance one of the following two reviews is spam and the other is ham (The comments are selected from the corpus [9]).

- Spam: While traveling for business I had my family join me and we stayed at the Fairmont Chicago, Millennium Park. It is breathtakingly beautiful and plush. The rates for rooms, though not the cheapest choice are quite reasonable for such fabulous amenities. The room decor is elegant modern and my wife says the spa is divine! Guests of this hotel are able to live in luxury - every room is equipped with complementary coffee, access to high speed internet, bathrobes, and 42 inch flat screen televisions. Enjoy the gorgeous city views and live the good life at the Fairmont Chicago, Millennium Park.
- Ham: My husband and I decided to take a trip to Chicago at the last minute and quickly chose the Conrad, not really knowing it's location. We were pleasantly surprised to find it was almost right on Michigan Ave. (It's connect to Nordstroms right in the heart of downtown.) Great for shopping. The hotel itself was fabulous. The staff was extremely helpful in giving suggestions for shows and restaurants. I recommend going to Joe's that's right across the street from the hotel. Everyone was raving about it . . .but we couldn't get a reservation. The room was wonderful and the bed and pillows were very comfortable. It had a very warm and inviting atmosphere. My husband loved the workout room. I highly recommend the Conrad!!

The authors in [3] used syntactic stylometry features for deception detection by considering unigram, bigram, and the combination of them as features. In another work [11] the authors detected deceptive opinion spam by the help of machine learning techniques using lexico-syntactic patterns, such as n-grams (unigrams, bigrams, trigrams) and part-of-speech (POS) tags and features are derived from the LIWC (Linguistic Inquiry and Word Count) output.

Using writing-style features to identify spammers have got the researchers' attention. Spammers try to write wise reviews and to hide their writing style; therefore it causes changes in writing style and verbal usage. To distinguish truthful reviews from fake reviews, there are several linguistics hints. For instance, spammer in compare with real

978-1-4799-3516-1/13/$31.00

reviewer use simpler, short, and fewer average syllables per word [2].

In this study we focus on the writing styles of the reviewers to classify deceptive and trustful reviews. Therefore, the main goal of this paper is to show that using stylometric (writing-style) features is a promising approach to detect fake reviews. The generated features for this study are categorized into two types, i.e. lexical and syntactic features. To verify the relative distinctive power of stylometric features, classification technique is applied separately to each type and combination of them.

The paper is organized as follows: in Section II, we summarize related works; in Section III, we introduce the deceptive opinion spam corpus we use in our study; in Section IV, we explain the features; in Sections V and VI, we describe the experimental part and in Section VII, we conclude the article.

II. Related Works

There are different types of spam such as e-mail spam, web spam and opinion spam [7]. Among these three types of spam, more concern is raised about opinion spam. The authors in [7] find that opinion spam is different from the other two types of spam. Detecting deceptive reviews are more complicated than other fake reviews. Therefore, different kinds of features and machine learning methods are used to recognize ham from spam.

Jindal and Liu in [7] categorized review spam in three types:

- Type 1 (untruthful opinions): undeserving positive reviews to promote, or malicious negative reviews to damage the reputation.
- Type 2 (reviews on brands only): review on brand, the manufacturers, services, and the sellers not on the product.
- Type 3 (non-reviews): non-reviews such as advertisements and irrelevant reviews comprising no opinions (e.g., questions).

Authors in [6], [7] defined three categories of features in the context of product review sites. The first category is review centric features such as number of feedbacks and length of review's body, the second category is reviewer centric features e.g. average rating given by reviewer and ratio of the number of cases in which he/she was the only reviewer and the last category is product centric features such as standard deviation in ratings and sale rank.

In [7], they generated 35 features based on the three feature categories mentioned above. The logistic regression, SVM, and Naive Bayes classifiers are applied and the logistic regression outperformed the others by the average AUC of 98.7% for spam types 2 and 3. The reason for high AUC value is detecting these types of reviews are easy because of little efforts of spammer to write these types of reviews. However, detecting the first type of spam is difficult and complicated. To detect first type of spam, they annotated the dataset based on the duplicate reviews with similarity score of at least 90% labeled as spam. They performed 10-fold cross validation on the data. It obtained the average AUC value of 78% using all the features.

In [15] the authors analyzed the behavioral characteristics between normal users and malicious users. Three characteristics of spammer are proposed: first, spammer probably have a high-frequency commenting time series than normal users. Second, they post similar and unrelated comments to their commented objects. Third, fake users send spam reviews without considering the domains of the objects. They designed two testing strategies and carried out experiments to evaluate the three above mentioned methods: A) A user is a spammer if each of his three behavioral characteristics presents that he is a spammer. B) A user is a spammer if any one of his three behavioral characteristics represents that he is a spammer. Experimental results indicate the accuracy of their detection strategy I (strategy for high accuracy) and strategy II (strategy for wide coverage) are 100% and 92.6%, respectively.

Some works focus on the singleton review spam detection such as [17]. They observed that trustful reviewers' arrival pattern is steady and uncorrelated to their rating pattern, but spammmers have opposite behaviors. These behaviors are observable through unusually correlated temporal patterns. The singleton review spam detection problem is considered as abnormally correlated pattern detection problem. They got the recall 75.86% and the precision 61.11%.

Nowadays, deceivers are becoming wiser, for instance they try to hide information by having additional imaginary effort, which often causes wise changes in human behavior [4]. Often, these changes influence writing style of spammers, therefore the stylometric features are considered as a method to find deceptive reviews.

In [5] and [20] four categories of features based on the writing style are extracted. The categories are lexical, syntactic, structural and content-specific features. In [5], three different clustering algorithms are used. Among them k-means and bisecting k-means outperformed Expectation-Maximization (EM). The best F-measure 90% achieved by applying k-means over a combination of all four feature types when e-mails per user is limited to 40.

In [20], the comparative learning algorithms are used for building feature-based classification models for on-line messages authorship identification. Neural networks outperformed Support Vector Machine and C4.5 significantly for detecting deceivers. When all features were used, the three classifiers achieved accuracy of 90 to 97% and 72 to 88%.

In a recent work [11], lexico-syntactic patterns are applied to detect deceptive opinion reviews. Integrating work from psychology and computational linguistics (n-gram), developed to detecting deceptive opinion spam, and ultimately developed a SVM light classifier to train their linear SVM

models that is nearly 90% cross-validated accurate.

III. Deceptive Opinion Spam Corpus

We applied the opinion spam corpus with gold-standard deceptive opinions which is introduced in [11] and [10]. The corpus is publicly available on-line [9]. The trustful reviews are collected from 20 most popular Chicago hotels from TripAdvisor and deceptive reviews gathered using Amazon Mechanical Turk (AMT) from those same hotels. This corpus includes: 400 truthful positive reviews from TripAdvisor and 400 deceptive positive reviews from Mechanical Turk are explained in [11], and 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp and 400 deceptive negative reviews from Mechanical Turk which are explained in [10]. These yield the final corpus of 1600 reviews.

IV. Feature Generation

The existing works have shown that the stylometric features are effective in detecting deception in different context such as e-mail, web and opinion [3], [5], [12] and [20]. The stylometric features reflect the writing style of reviewers. Although deceivers have some specific writing styles to hide the information such as simpler and shorter words, using stylometric features are very helpful to detect untruthful reviewer and consequently fake reviews. In this paper, we adopted 234 stylometric features; the features are categorized into lexical (character-based and word-based) and syntactic features. A brief description of features are given below.

Using the lexical features help to learn about the preferred use of each characters and words of a reviewer. Lexical features can be divided into two categories, character-based and word-based features. Some of the lexical character-based features used are based on [5], [12] and [20] as they are explained in Table I. These contain character count (N), ratio of digits to N, ratio of letters to N, ratio of uppercase letters to N, ratio of spaces to N, ratio of tabs to N, occurrences of alphabets (A-Z) (26 features), and occurrences of special characters: $<>\hat{}\ \%\,|\,\{\}\,[]\,\backslash\,/\,\#\ \ \sim\ +\ -\ \div\ *\ \&\,@\,\$$ (20 features).

Lexical word-based features are: token count (T), average sentence length in terms of characters, average token length, ratio of characters in words to N, ratio of short words (1-3 characters) to T, ratio of word length frequency distribution to T (16 features), ratio of types to T, vocabulary richness (Yule's K measure [18] and [19]) to parametrize vocabulary richness in each of reviews. Vocabulary richness measures the distribution of word frequencies. Yule proposed a measure, so called the Yule's K value. It is defined as follows:

$$\text{Yule's } K = 10000 - (M_2 - M_1)/(M_1 \times M_1) \qquad (1)$$

Where M_1 is the number of all word forms a text consists of and M_2 is the sum of the products of each observed frequency to the power of two and the number of word types observed with that frequency. The larger Yule's K can result the smaller diversity of the vocabulary.

The last two lexical word-based features are Hapax legomena and Hapax dislegomena. Hapax Legomena and Hapax dislegomena are the terms used for once-occurring and twice-occurring words. In total, we choose 77 lexical features (see Table I).

Syntactic features represent the writing style of reviewers at the sentence level. Syntactic features include: occurrences of punctuations that include . ? ! : ; ' " (7 features) [1]. Mosteller [8] for the first time showed the effectiveness of occurrences of punctuations to clarify the disputed work. Occurrences of function words (150 features) which are selected based on the English function words are listed in [20]. In total, we select 157 syntactic features (see Table I).

V. Experimental Design

Our goal in this section is to evaluate the adopted features to analyze whether it can precisely classify deceptive reviews. The set of experiments are required to be designed in such a way that can answer the following questions. Which of the classification algorithms perform better than others for a given dataset? What is the relative strength of each two different types of stylometric features? In the experiments part, these questions will be answered.

We have performed three set of experiments. More precisely, in order to evaluate lexical features, syntactic features, and both lexical and syntactic features together we applied classification method.

To test the validity of our experiment, F-measure as a performance measurement is applied. In fact, F-measure is derived from precision and recall, which are the basic measures used in evaluating F-measure commonly employed in the field of Information Retrieval. The default is to equally weight precision and recall, giving a balanced F-measure. Therefore, F_1 formula is used which is presented by the following equation:

$$F_{\beta=1} = 2 * \frac{recall \times precision}{recall + precision} \qquad (2)$$

The first feature set includes lexical features and the second one contains syntactic features and the last one is the combination of lexical and syntactic features. To examine the impact of each feature set we conduct three different experiments. We used the WEKA [16] machine learning platform as a tool for conducting our experiments. In all the three set of experiments two different WEKA's classification algorithms, namely Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) was implemented in WEKA with a polynomial kernel, which is an algorithm for efficiently solving the optimization problem which arises during the training of support vector machines, and Naive Bayes (NB) were applied. For all classification tasks, we use

Table I
ADOPTED FEATURES

Feature type	Description	Implementation
Lexical Features		
Character-based features		
1. Total number of characters (N)	All letters Aa-Zz, all digits, all punctuation marks	Total number of character tokens
2. Ratio of total number of digits to N	All digits 0-9	Total number of digit characters/N
3. Ratio of total number of letters to N	All letters Aa-Zz	Total number of alphabetic characters/N
4. Ratio of total number of upper-case letters to N	All upper-case letters A-Z	Total number of upper-case alphabetic characters/N
5. Ratio of total spaces to N	All spaces	Total number of white-space characters/N
6. Ratio of total tabs to N	All tabs	Total number of tab spaces/N
7-32 Occurrences of alphabets (26 features)	All letters Aa-Zz	Frequency of letters
33-52 Occurrences of special characters (20 features)	<>ˆ % | { } [ ] \ / # ∼ + − ÷ ∗ & @ $	Frequency of special characters
Word-based features		
53. Total number of tokens (T)	All words	Total number of words in review
54. Average sentence length in terms of character	Average sentence length based on word tokens in sentence	Number of words in review/Number of sentences
55. Average token length	Average number of characters in a word review, based on alphabetic word tokens	Number of letters in review/number of words in review
56. Ratio of characters in words to N		Total number of characters in review/N
57. Ratio of short words (1-3 characters) to T	e.g. the, if	Total number of short tokens in review/T
58-73 Ratio of word length frequency distribution to T (16 features)	Length 1-16 based on the whole word length in whole reviews	Word length frequency in a review/T
74. Ratio of types to T	Words come in both types and tokens. For example, there is only one word type 'the' but there are numerous tokens of it on a documents	Total number of types in a review/T
75. Vocabulary richness (Yule's K measure)	A vocabulary richness measure defined by Yule	Using Yule's K equation
76. Hapax legomena	The definition of measure can be found in [14]	
77. Hapax dislegomena	The definition of measure can be found in [14]	
Syntactic Features		
78-84 Occurrences of punctuations (7 features)	. ? ! : ; ' "	Frequency of punctuation marks
85-234 Occurrences of function words (150 features)	Using list of words that presented in [20]	Frequency of each function word

80% of data for training and 20% for testing, with 10-fold cross validation.

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The results for the comparison of three different feature sets and two classifiers are summarized in Table II and Figure 1. We found that the F-measure value kept increasing by using syntactic features, and combining syntactic features to lexical features. For all three feature sets, SMO outperformed Naive Bayes. The best F-measure value was accomplished when using SMO and the combination of both lexical and syntactic features. In the following, we explain the results based on feature sets and classifiers.

Lexical Features: looking at the individual feature set, lexical, SMO and Naive Bayes attain 81% and 70% in terms of F-measure, respectively. Although a few features such as ratio of tabs to N may not be useful for short reviews, the results indicate that lexical features using SMO are applicable for deception detection.

Syntactic Features: In case of using syntactic features alone, SMO and Naive Bayes attain 76% and 69% accuracy, respectively. These results show that the F-measure value is improved using SMO classifier comparing to the Naive Bayes classifier. The number of function words has impact

Table II
F-MEASURE FOR THREE FEATURE SETS AND TWO CLASSIFIERS

	Naive Bayes	SMO
Lexical	70%	81%
Syntactic	69%	76%
Lexical + Syntactic	74%	**84%**

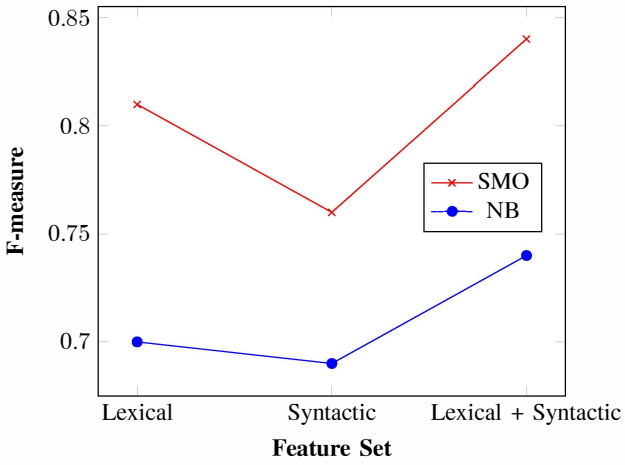

Figure 1. F-measure vs. Feature Sets and Classification Algorithms

on the accuracy, for example in [13], the authors found that the best accuracy for their experiment achieved by using 60 function words. Therefore, changing the number of function words may improve the results of experiment.

Lexical and Syntactic Features: The F-measure value of SMO classifier is 84% and for Naive Bayes is 74%. The best result for both classifiers is obtained by applying SMO classifier on lexical and syntactic combination, i.e. lexical and syntactic features.

The best F-measure value, 84%, is obtained by applying SMO over a combination of both lexical and syntactic features. F-measure values of SMO in all three experiments are higher in comparison with Naive Bayes and the results are comparable to most previous studies using writing-style features [3], [5], [20], and using other spam detection techniques [17].

VII. CONCLUSION

In this paper we examined the effects of the stylometric features on the review classification. We tested 77 lexical and 157 syntactic features and two different classifiers, i.e. Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) and Naive Bayes for deceptive review classification. Since some linguistic features change when people hide their writing style hence by identifying those features fake reviews can be recognized.

We categorized our experiments based on the three types of features: lexical and syntactic separately and combination of them. The best F-measure value achieved by applying SMO over a combination of both lexical and syntactic features. The promising results obtained through our experiments validated usefulness of stylometric features for deception detection.

However, better alternative techniques do exist and they may further improve performance, which require more investigations. We have identified several future research directions based on the current study. First, to extract content-specific features and second, to find the effects of combination of the content-specific features with syntactic and lexical features to find optimal set of features with high accuracy.

ACKNOWLEDGMENT

This work is partially supported by the Malaysia Ministry of Science, Technology and Innovation, Sciencefund, number 5450721 and the Universiti Putra Malaysia, Research University Grant Scheme (RUGS) number 05-02-12-2153RU.

REFERENCES

[1] H. Baayen, H. V. Halteren, and F. Tweedie. Outside the cave of shadows: using syntactic annotation to enhance authorship attribution. *Literary and Linguistic Computing*, 11(3):121–132, 1996.

[2] J. Burgoon, J. P. Blair, T. Qin, and J. F. Nunamaker Jr. Detecting deception through linguistic analysis. In Hsinchun Chen, Richard Miranda, DanielD. Zeng, Chris Demchak, Jenny Schroeder, and Therani Madhusudan, editors, *Intelligence and Security Informatics*, volume 2665 of *Lecture Notes in Computer Science*, pages 91–101. Springer Berlin Heidelberg, 2003.

[3] S. Feng, R. Banerjee, and Y. Choi. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 171–175, Stroudsburg, PA, USA, 2012. Association for Computational Linguistics.

[4] M. G. Frank, M. O'Sullivan, and M. A. Menasco. Human behavior and deception detection. J. G. Voeller (Ed.), Handbook of Science and Technology for Homeland Security. New York: John Wiley & Sons., 2009.

[5] F. Iqbal, H. Binsalleeh, B. C. M. Fung, and M. Debbabi. Mining writeprints from anonymous e-mails for forensic investigation. *Digit. Investig.*, 7(1-2):56–64, October 2010.

[6] N. Jindal and B. Liu. Analyzing and detecting review spam. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 547 –552, oct. 2007.

[7] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the international conference on Web search and web data mining*, WSDM '08, pages 219–230, New York, NY, USA, 2008. ACM.

[8] F. Mosteller and D. L. Wallace. Inference and disputed authorship: The federalist. In *behavioral science:quantitative methods edition.* Addison-Wesley, 1964.

[9] M. Ott. Deceptive Opinion Spam Corpus v1.4. [Online]. Available: http://myleott.com/op_spam/

[10] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Short Papers*, Atlanta, Georgia, USA, June 2013. Association for Computational Linguistics.

[11] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. 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*, HLT '11, pages 309–319, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[12] L. Pearl and M. Steyvers. Detecting authorship deception: a supervised machine learning approach using author writeprints. *Literary and Linguistic Computing*, 27(2):183–196, 2012.

[13] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Computer-based authorship attribution without lexical measures. *Computers and the Humanities*, 35(2):193–214, 2001.

[14] F. J. Tweedie and R. H. Baayen. How variable may a constant be? measures of lexical richness in perspective. *Computers and the Humanities*, 32(5):323–352, 1998.

[15] Q. Wang, B. Liang, W. Shi, Z. Liang, and W. Sun. Detecting spam comments with malicious users' behavioral characteristics. In *Information Theory and Information Security (ICITIS), 2010 IEEE International Conference on*, pages 563 –567, dec. 2010.

[16] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems).* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.

[17] S. Xie, G. Wang, S. Lin, and P. S. Yu. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 823–831, New York, NY, USA, 2012. ACM.

[18] G.U. Yule. On sentence- length as a statistical characteristic of style in prose: With application to two case of disputed authorship. *Biometrika*, 30:363–390, 1939.

[19] G.U. Yule. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944.

[20] R. Zheng, J. Li, H. Chen, and Z. Huang. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Am. Soc. Inf. Sci. Technol.*, 57(3):378–393, February 2005.