Identification of Fake Reviews Using New Set of Lexical and Syntactic Features

Rupesh Kumar Dewang
CSED MNNIT Allahabad,
Allahabad (U.P) INDIA
rupeshdewang@mnnit.ac.in

A.K. Singh
CSED MNNIT Allahabad,
Allahabad (U.P) INDIA
ak@mnnit.ac.in

ABSTRACT

The services and products of E-Commerce portals in this digital age are heavily reviewed by the users. These reviews provide useful insights on the quality/usage of these products. Due to such importance of reviews, they can be faked to give false opinions about products and subsequently mislead the users. In this paper we are proposing new set of lexical and syntactic features set and applying supervised algorithms for performing classification on fake reviews dataset *(gold standard)*. We focus on the writing style, that include type of punctuation mark, Part-of- Speech (POS) etc., that are helpful for detection of reviews spam. The final results give promising accuracy 91.51% for detecting fake reviews.

Keywords

Opinion mining, reviews spam, lexical and syntactic features, supervised learning algorithms.

1. INTRODUCTION

On-line shopping is increasing day by day; reported in the survey conducted by *"Google and Forrester Consulting"*[1]. They conveyed that the on-line shopping business in India will reach $15 billion in 2016. One another similar study by *"Assocham-PwC"*[2] say that, right now Indian customers spend Rs. 6000 on average for on-line shopping; this will increase up to 67% to Rs. 10,000 in 2016.

In fig.1. it is shown that on-line retails market size and its expectedly growth in India [16] from 2007-08 to 2015-16 will increase by 2015-16 to $504 billion.The growing trend of on-line shopping is highly influenced by the product reviews written by user who have purchased those goods. Looking at the power of reviews in influencing buyers, merchant often ask them to write reviews.

Figure 1: On-line Retails Market Size and Growth by CRISIL Research [16]

As the number of customers are increasing day by day, reviews received by product are also increasing in the same manner. In order to promote (written positive review) or demote some target product (written negative review), fake reviews can be provided by the competitors. *Liu* [9, 10] was first to report this issue in year 2007-08. They had shown that there are three types of fake review: *untruthful opinion, review on brands and non-review.*

The writing styles are defined in the way of making sentences. The reviews generated by users could be biased or unbiased. The issue of finding fake review based on writing style of reviews is given by [14]. They mainly focus on helpful products review by the readability and some structural features test. By this readability test they suggested that we can warn those users who used some complex word and long sentences by compression between helpful review and non-helpful review. But author did not focus on any other features like uni-gram, bi-gram etc. and not any spam and non-spam reviews were identified.

Ott et al. [18] used these issues and generated gold standard data set (which is used in this paper). For constructing this dataset they used AMT (American Mechanical Turk) and human sources. They classified deceptive reviews by using POS features (which was called genre identification) and LIWC (Linguistic Inquiry and Word Count) software to detect psychologically text in four categories. Many authors worked on readability, writing style, genre identification, linguistic features etc.,[17, 1, 8, 15, 11].The example of

[1] $http://www.livemint.com/Industry/nWCyKyN5flefBDTQPU5AwM/100-million-on-line-shoppers-in-India-by-2016-Report.html$.
[2] http://tech.firstpost.com/news-analysis/Indian-online-shopping-to-increase-to-67-percent-in-2015-report-247318.html

ICCCT '15, September 25-27, 2015, Allahabad, India

DOI: http://dx.doi.org/10.1145/2818567.2818589

one fake review and one non-fake review is given below, of *"Sofitel Chicago"* hotel form gold standard dataset[18].

Fake review*: "On our first trip to Chicago, my fiance and I stayed 2 nights on our anniversary and were pleasantly surprised by the Hotel and the service we received from the personnel. In the past from other "luxury hotels" we have paid much more and received much less. Here the room were very spacious, comfortable and sleek. We spent some time in the hotel bar, LE BAR, and enjoyed a few of their specialty martinis that were second to none! We also enjoyed dinner at the hotel's Cafe des Architects and received service some of the best service in quite some time. If traveling with a lover we recommend the "Air of Romance" package. It was the best! 6 stars!"*

Non fake review*: "On a recent trip to Chicago to attend a major trade shows I had the pleasure of staying with the Sofitel Chicago Water Tower. I say stay with as that is how they make you feel, from check in to check out they go above and beyond to ensure your stay is perfect and you want to return. The rooms are clean, chic and roomy, the beds - the best I have slept in, bathrooms large and super clean. What can I say, my new home when in Chicago."*

On the basis reading these reviews, user cannot judge which one is fake or non-fake. The human expert are able to judge the reviews which are fake and non-fake, but they take lot of time and as per [16] on-line user are increasing day by day, so we require machine learning techniques which gives the results very fast, due to this reason we need some techniques to identify. There are "30 ways to identify" [3] the fake reviews. But reviews dataset are very big and increasing day by day. The machine learning techniques, patterns discovery, psychology behavioral method and some relational modeling etc. could be used to identify the fake reviews. The Review content, like lexical features, content and style similarities and semantic inconsistency are also helpful for reviews spam detection[4].

In this paper, he have applied cognitive language learning concept in review spam detection. Our hypothesis is that, reviewers are sometimes extending their own knowledge by copying text from other review with some modifications for creating fake reviews. Spammer knows the psychology of users. They use cognitive process that is heavily influenced by previous reviews posted by other users. We mainly focused on the lexical feature and syntactic feature which is acquired from the cognitive language learning for reviews spam detection. In this work, we have generated new set of 16 lexical and 25 syntactic features for spam detection. The lexical feature helps us to learn about each character used and different words used by the writer. The syntactic features are used for syntactic development in language learners. We have used same gold standard dataset as proposed in [18]. First we have extracted the features then applied four supervised classification algorithms. The results obtained are promising in detecting fake reviews and non-fake reviews. Rest of this paper is organized as follows: In section 2, we have discussed related works; in Section 3, we have discussed about the dataset used for spam detection; in Section 4, we explained the new features constructed; in Section 5, we have discussed experimentation and analysis ; Section 6, we explained results discussion and in last Section 7 conclusion is provided.

[3]http://consumerist.com/2010/04/14/how-you-spot-fake-online-reviews/

[4]http://www.cs.uic.edu/ liub/FBS/fake-reviews.html

2. RELATED WORK

Spam basically defines, false or fake message sent using electronic media to many recipients. Message may also contain some hidden script for doing other types of cyber attack. Electronic spamming is defined as sending spam messages using the electronic media system.

The research in spam detection started in email spam domain. Later this term is applied in many other media such as chat spam, newsgroup spam, web spam, blog spam, SMS spam, Internet forum spam, social spam, file sharing spam and review spam [5]. In the year 2008 the existence of review spam is defined by [9, 10], in which review spam is taken as new type of spam which lie on e-commerce web site in the form of review. They have defined three types of review spam:

Type 1 (untruthful opinions): when review is not given as a product alike, means review given for promoting and demoting the product reputation by positive and negative posting reviews.

Type 2 (reviews on brands only) : When the review is written in consideration of only brands name instance of product of that brands, spammers want to damage the reputation of brands.

Type 3 (non-reviews): when the review only contains advertisement and other irrelevant opinion of the products like question and answering and random text.

Based on this categorization [9, 10] they detected the reviews are spam or non spam. The readability of the reviews is major issue for detection of reviews spam. When the reviews are easily readable means chances of spam is more. The issue of finding fake review based on writing style of reviews is given by [14]. They mainly focused on helpful products review by the readability and some structural features test. By this readability test they suggested that we can warn those users who used some complex word and long sentences by comparison between helpful review and non-helpful review.

N.Hu. et. al also addressed issues related to writing style. They used three measures to check writing style of reviewers and manipulation in reviews by ratings, sentiments and readability and found around 10.3% products are subject to on-line manipulation. They have shown the manipulation applicable in on-line ratings and this is not possible in sentiment [8]. The first publicly available gold standard dataset [17] with positive and negative sentiment, contain 400 reviews of 20 Chicago hotels. *M.Ott et. al.* has proposed n-gram based text categorization technique to detect the deceptive reviews [17]. They have achieved 86% accuracy on the same gold standard dataset.

Linguistics analysis is also useful for detection of shill reviews (spam reviews) as reported in [15]. They had performed comparison between normal reviews and shill reviews as a main aim of paper. They created shill reviews dataset. They used available measures of readability and subjectivity of other area and performed it to find products related features. They changed the existing perception that, shill reviews are more readable and on the contrary maintained that the shill reviews are less readable in terms of content which is repetitive and long. It tried to describe

[5]$http://en.wikipedia.org/wiki/Spammingcite_note-4$

many features. They also tried to focus on official feature with products details. The main limitation of this paper is, the sample size used is very small. On the other hand they only collected positive reviews, because negative reviews are harmful for organization. In [11] the authors also worked on linguistics features, psychological features, current concerns, spoken features and punctuation feature and shown 93% accuracy when they run classification algorithms on these features.

Banerjee et. al, has addressed the problem of fake review in [1, 3, 2, 4] terms of linguistics analysis and writing style. They used genre identification, readability and writing style. In [3] they have shown the way of writing the fake reviews by spammer mind set, so user can know by reading the review which is original or fake [1]. *Banerjee et. al* proposed a framework for fake reviews detection using readability, genre and writing style in [2]. They had applied same concept on manipulative and authentic negative reviews using supervised algorithms to detect the fake and non-fake reviews based on the linguistic features which mainly understands ability, level of details, writing style and cognition indicators [4].

Another work on linguistic features has proposed in [22]. They proposed novel model on deep linguistic features, which is derived from syntactic dependency parsing tree. They used both English and Chinese dataset. Results showed that proposed method gave 2% better results when compared to previous work.

Shojaee et. al., also worked on linguistic features in [20], they used stylometric features and applied supervised learning algorithms, Sequential Minimal Optimization (SMO) and Naive Bayes, to detect deceptive opinion and results showed that, when they had applied SMO on combined set of lexical and syntactic features, they achieved F-measure 84%.

Goswami et. al., also used linguistic features approach in hospitality sector, they have done one thing to combine the genre identification and text categorization[6] with merging POS and achieved 92% accuracy.

Li et. al, had proposed new gold standard dataset, which consists of data from three different domain Hotel, Restaurant, Doctor [12]. The dataset belong to three types of reviews, i.e. customer generated truthful reviews, Turker generated deceptive reviews and employee (domain-expert) generated deceptive reviews. For features generation they have applied SAGA model, which is adaptive in nature and easily applied in cross domain classification task.

This work shows that the new research direction is going to merge new set of features in linguistic domain with some more attributes like product details, location of login and ratings etc. This will be helpful in increasing the accuracy of spam detection.

3. DATASET USED FOR SPAM DETECTION

We used publicly available gold standard dataset, [18, 17]. The Dataset contains 1600 reviews of 20 numbers of *Chicago hotels*. The reviews are labeled into truthful or deceptive reviews. The details of the dataset is describe as [18], in which 400 truthful positive reviews gathered from *TripAdvisor*[18], 400 deceptive positive reviews gathered from *Mechanical Turk*[18], 400 truthful negative reviews from *Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp*[17]and 400 deceptive negative reviews from *Mechanical Turk* [17]. Dataset explained in terms of positive and negative sentiment is calculated in paper [18, 17]. There are only two attribute given one is hotel name and another is class of review whether it is truthful or deceptive. For experimentation we have used 400 positive deceptive reviews and 400 positive truthful reviews for classification purpose.

4. NEW FEATURE CONSTRUCTION

Early researchers worked on various features like linguistic feature, readability and informativeness etc., for detecting reviews spam. The sentiment analysis techniques were also used by many authors. We have created new set of lexical and syntactic features. The lexical and syntactic features like POS, TTR, lexical diversity etc., are more useful for checking reviews is spam or non spam. We have analyzed the three dimensions lexical richness through the lexical diversity, sophistication and variation). We have explained the features below:-

4.1 Lexical Feature

Lexical feature help us to learn about each characters used and different words used by the writer. Lexical features mostly used are n-gram, POS etc. Spammer used many different words for writing down the spam reviews. Lexical diversity is defined as distribution of content words such as nouns, verbs, adjectives and adverb. The longer reviews have minimum number of tokens as compared to the shorter reviews. When any reviews are having larger number of content words then reviews contain more information about the products. To calculate the lexical diversity (Ld) in particular review the following formula can be used.

$$Ld = (NLex/N) * 100 \quad (1)$$

Where Nlex= Number of lexical word tokens N= Number of all tokens.

Type-Token ratio (TTR) is the ratio of number of word types (T) in the review to total number word token (N) in a review (R). We also calculate the variation of TTR: like, Corrected $TTR(CTTR) = T/\sqrt{2N}$, $RootTTR = T/\sqrt{N}$, Billogarithmic $TTR = LogT/LogN$ and Uber Index= Log2T /Log(N/T)[21].We have calculated the lexical variation in terms of ratio of the number of lexical types to lexical tokens used. We have also used various variation of verb, these are verb variation-1(Tverb/Nverb), squared verb variation-1 (T2verb/Nverb), and corrected verb variation-1 ($Tverb/\sqrt{2}Nverb$) [21], then we have calculated lexical density of the review by ratio of the number of lexical items in relation to the total number of words in review. When lexical density of review is high, this shows review contains the large amount of information-carrying words and low lexical density shows comparatively less information-carrying words. We have also included the average number of character per word (NumChar) and average numbers of syllables per word (NumSyll) [21]. The main focus of these two indicator on word level is to calculate the text complexity of reviews.

4.2 Syntactic Features

We have used 25 syntactic features from [13], which was used for syntactic development in language learners. We measured syntactic complexity by sentential clause and T-unit length of unit. The review (sentence) is contemplated with a group of words delineated with punctuation mark. Clause has some construction with a subject and a finite verb. T-unit is distinguished as one main clause together

Table 1: Set of Lexical Features

	Lexical Density(LD)
	Type-token ratio(TTR)
	Corrected TTR(CTTR)
	Root TTR
	Billogarithmic TTR
	Uber Index
	Lexical Word Variation
	Verb variation
	Squared VV1
	Corrected VV
	Noun variation
	Adjective Variation
	Adverb Variation
	Modifier Variation
	Mean textual lexical density
	Average Number Character per word(Num Char)
	Average Number Syllables per word(Num Syll)

with some subservient clause or non-clausal construction that is joined in it. We have taken review (sentence) complexity as another feature which is based on clauses per sentence. Another measure is fabricated of three ratios that measure the amount of content coordination in reviews text, which are clauses per T-unit, deepened clauses per clauses, complex T-units per T-unit and dependent clauses per T-unit.

5. EXPERIMENTS AND ANALYSIS

We have used the *R-package koRpus*[6] to calculate the lexical and syntactic features set (which are shown in Table 1 and 2 respectively). *koRpus* is used to measure the similarities and differences between texts, as well as it is now used for scientific researches and calculating lexical and readability diversity features [5].We have used 400 positive deceptive reviews and 400 positive truthful reviews dataset, (total 800 reviews) then we have generated the features (lexical and syntactic one by one) value using *koRpus*. In *R-programming*, we use *tree-tag()* function for tagging the input file. Then we have used *tokenize()* for generating the tokens, then next we have calculated the lexical diversity of text using *lex.div()*, which gives the TTR value and variants value .

We have used *WEKA toolkit* for classification[7], in which we have mainly used Sequential minimal optimization(SMO), Naive Bayes, Decision tree and Bayesian Logistic algorithms. We first selected features set and train the classifier algorithms using *WEKA*. We used 5- fold cross validation. We used 80% data to train the model and 20% data to test the model. We performed testing on reviews, review is true positive (TP) when the fake review is classified correctly, false negative (FP) when incorrectly classified and same way true negative (TN) when truthful review correctly classified, otherwise false positive (FP) when incorrectly classified. F-measure also knows by F1 score. It is applied in statistical analysis of classification task. F-measure is used to test accuracy. It is combined Precision and recall. The traditional formula of F- measure is:

$$F1 = 2 * (Precision * Recall)/(Recall + Precision) \quad (2)$$

[6]http://rpackages.ianhowson.com/cran/koRpus/

Table 2: Set of Syntactic Feature

	Mean length of clause
	Mean length of a sentence
	Mean length of T-unit
	Num. Of Clauses per sentence
	Num. of T-Units per sentence
	Num. of Clauses per T-unit
	Num. of Complex-T-Units per T-unit
	Dependent Clause to Clause Ratio
	Dependent Clause to T-unit Ratio
	Co-ordinate Phrases per Clause
	Co-ordinate Phrases per T-unit
	Complex Nominals per Clause
	Complex Nominals per T-unit
	Verb phrases per T-unit
	Num. NPs per sentence (NumNP)
	Num. VPs per sentence (NumVP)
	Num. PPs per sentence (NumPP))
	Avg. length of a NP (NPSize)
	Avg. length of a VP (VPSize)
	Avg. length of a PP (PPSize)
	Num. Dependent Clauses per sentence (NumDC)
	Num. Complex-T units per sentence (NumCT)
	Num. Co-ordinate Phrases per sentence (CoOrd)
	Num. SBARs per sentence (NumSBAR)
	Avg. Parse Tree Height (TreeHeight)

SMO is used for finding the solution of quadratic programming problem, which comes when we trained SVM [19]. The amount of memory used in SMO is linear because the quality of SMO is easily applicable on large dataset. SMO gives very good results when compared with linear SVM method. We used decision tree algorithm which is in *WEKA* named as J-48. Decision tree learning uses a decision tree as a predictive model which maps observations about item conclusions regarding its target value. We used Naive bayes algorithm which consists with probabilistic classifier. The proposed method had combined lexical and syntactic feature set which correctly identify 373 out of 400 truthful reviews, and 358 out of 400 fake reviews.

6. RESULTS DISCUSSION

The main aim of this paper is to correctly classify reviews as spam or not spam on using given set of new lexical and syntactic features. We have used four algorithms on three features set; first feature is lexical, second is syntactic and third is lexical combined with syntactic.

In table 3 we have shown the F-measure percentage calculated from four supervised algorithm with three features set. We have calculated the lexical feature and syntactic feature on dataset define above in section 3, and then applied SMO, Decision tree, Naive Bayes and Bayesian logistic, algorithms one by one in Weka. When we combined lexical and syntactic features SMO achieved highest accuracy 91.51 % .Using purely lexical features the decision tree classifier gives accuracy value 79.77%, and Naive bayes gives 77.54%, Bayesian Logistic gives 74.67% and SMO gives best result 85.91%.

When we used syntactic features, in which decision tree algorithm gives 74.29%, Naive bayes gives 72.79%, Bayesian Logistic gives 69.90% and SMO gives 80.77%. We combined lexical and syntactic feature then all four algorithms gave

Table 3: F- Measure for various Supervised Algorithms using different features set

	SMO	DT	NB	BayL
Lexical	85.91%	79.77%	77.54%	74.67%
Syntactic	80.77%	74.29%	72.79%	69.90%
Lexical+Syntactic	**91.51%**	84.63%	81.24%	79.44%

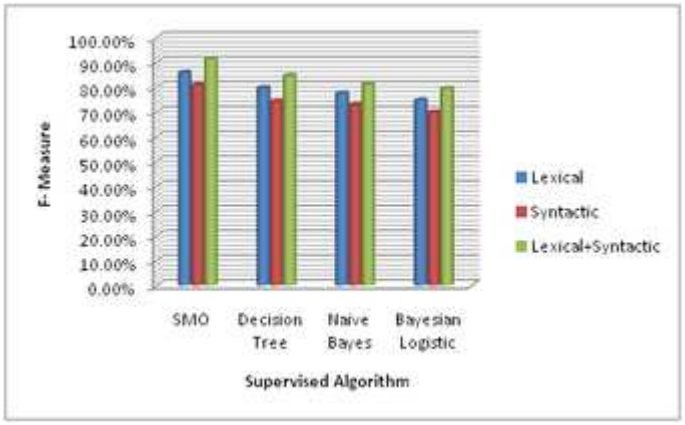

Figure 2: F-measure for Various Supervised Algorithm Using Different Features Set

best results with respect to papers [17, 1, 20, 6], decision tree gives the 84.63%, Naive bayes gives 81.24%, Bayesian Logistic gives 79.44% and SMO gives 91.51%. In fig 2. We showed the accuracy percentage of SMO, Decision tree, Naive bayes, Bayesian logistic algorithm with respect to lexical, syntactic and combine lexical and syntactic together, in which SMO gave 91.51% accuracy in combining two features set.

7. CONCLUSIONS

This paper presents the analysis based on new set of lexical and syntactic features for review spam detection using supervised algorithms. In our proposed method when we added some new features set (lexical and syntactic), accuracy is increasing in promising way. In future, some more new features set could be used such as combining the sentiments with n-gram, lexical and syntactic features.

8. REFERENCES

[1] S. Banerjee and A. Y. Chua. A linguistic framework to distinguish between genuine and deceptive online reviews. In *Proceedings of the International Conference on ICWS*, 2014.

[2] S. Banerjee and A. Y. Chua. A study of manipulative and authentic negative reviews. In *Proceedings of the 8th International Conference on UIMC*, page 76. ACM, 2014.

[3] S. Banerjee and A. Y. Chua. Understanding the process of writing fake online reviews. In *Digital Information Management (ICDIM), 2014 Ninth International Conference on*, pages 68–73. IEEE, 2014.

[4] S. Banerjee, A. Y. Chua, and J.-J. Kim. Using supervised learning to classify authentic and fake online reviews. In *Proceedings of the 9th International Conference on UIMC*, page 88. ACM, 2015.

[5] C. Cristancho and E. Anduiza. Connective action in european mass protest. In *workshop on Activist Social Media Communication*, 2013.

[6] M. Goswami and S. Gupta. Determination of fake reviews in hospitality sector. In *Proceedings of the International Conference on Data Mining (DMIN)*, page 1, 2014.

[7] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten. The weka data mining software: an update. *ACM SIGKDD explorations newsletter*, 11(1):10–18, 2009.

[8] N. Hu, I. Bose, N. S. Koh, and L. Liu. Manipulation of online reviews: An analysis of ratings, readability, and sentiments. *DSS*, 52(3):674–684, 2012.

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

[10] N. Jindal and B. Liu. Opinion spam and analysis. In *Proceedings of the 2008 International Conference on Web Search and Data Mining*, pages 219–230. ACM, 2008.

[11] A. Karami and B. Zhou. Online review spam detection by new linguistic features. *iConference 2015 Proceedings*, 2015.

[12] J. Li, M. Ott, C. Cardie, and E. Hovy. Towards a general rule for identifying deceptive opinion spam. ACL, 2014.

[13] X. Lu. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, 15(4):474–496, 2010.

[14] M. P. O'Mahony and B. Smyth. Using readability tests to predict helpful product reviews. In *Adaptivity, Personalization and Fusion of Heterogeneous Information*, pages 164–167, 2010.

[15] T. Ong, M. Mannino, and D. Gregg. Linguistic characteristics of shill reviews. *Electronic Commerce Research and Applications*, 13(2):69–78, 2014.

[16] C. Opinion. *published by CRISIL RESEARCH*. Feb. 2014.

[17] M. Ott, C. Cardie, and J. T. Hancock. Negative deceptive opinion spam. In *HLT-NAACL*, pages 497–501, 2013.

[18] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock. Finding deceptive opinion spam by any stretch of the imagination. pages 309–319. ACL, 2011.

[19] J. Platt et al. Fast training of support vector machines using sequential minimal optimization. *Advances in kernel methodsâĂŤsupport vector learning*, 3, 1999.

[20] S. Shojaee, M. A. A. Murad, A. Bin Azman, N. M. Sharef, and S. Nadali. Detecting deceptive reviews using lexical and syntactic features. In *(ISDA), 2013 13th International Conference*, pages 53–58. IEEE, 2013.

[21] S. Vajjala and D. Meurers. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. ACL, 2012.

[22] Q. Xu and H. Zhao. Using deep linguistic features for finding deceptive opinion spam. In *COLING (Posters)*, pages 1341–1350. Citeseer, 2012.