Ensemble Learning to Find Deceptive Reviews using Personality Traits and Reviews Specific Features

Mayank Saini, Aditi Sharan
Jawaharlal Nehru University, New Delhi
India
mayanksaini1986@gmail.com
aditisharan@mail.jnu.ac.in

JDIM
Journal of Digital Information Management

ABSTRACT: *In the current era of Internet, people are increasingly using the e-commerce websites for purchasing goods and services. Reviews and blogs have become the prime source of information for making purchasing decisions. As reviews and blogs directly affect sales and revenue, many e-commerce companies hire people for writing reviews to promote or demote target products and services. These fictitious opinions that are written to sound authentic are known as deceptive reviews. In this paper, we tried to establish a link between personality traits and deceptive/fake reviews. We analyzed personality recognition techniques and deceptive review detection from a psycholinguistic point of view. We tried to capture stable individual characteristics to predict behavioral differences between deceptive and truthful reviewer/review. This study shows that personality clues along with other review specific features can be quite successful to build automatic deceptive review classifiers. We have used various ensemble learning techniques to ensure effective use of the features and achieve good classification accuracy. Our experiments on restaurant and hotel domain have achieved up to 93 and 94 percent accuracy respectively with the final classifier.*

Subject Categories and Descriptors
K.4.4 [Electronic Commerce] H. [Information Systems]: Software psychology

General Terms
E-Commerce, Online shoping, E-Marketing, Opinion Mining

Keywords: Opinion mining, Opinion spamming, Ensemble learning, Personality traits, Readability, Lexical diversity

Received: 11 September 2017, Revised 19 December 2016, Accepted 18 January 2017

1. Introduction

Analyzing people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations are known as opinion mining or Sentiment analysis [1], [2]. Recently many researchers have been evaluating online sources of texts such as product reviews, forum posts, and blogs to help end users in making purchase decisions. Most of the research carried out was based on the assumption that all reviews are authentic. But, not all online reviews are produced by genuine users of the products; so the outcome of the opinion mining research may drift from the reality. Writing these fake reviews that try to mislead human readers deliberately by giving undeserving positive opinions or false negative opinions is known as opinion spamming. People with malicious intentions post fake opinions without disclosing their true identity, and they are also known as opinion spammers. Positive reviews make a significant impact on the sale of the products and reputation which also bring financial gain [3], while the

negative reviews do the opposite. Companies are often indulged in writing deceptive reviews (spam) to promote/demote the product and services. It is important to identify such deceptive spammers due to serious social and financial consequences attached to it. Manually, it is hard for an average person to detect any type of deception [4].

Various studies have been conducted to explore the correlation between personality traits and deception [5][6]. The benefit of using personality traits over other linguistic clues is that it helps us to work on 'reviewer level' rather than the 'review level'. It basically identifies not only a deceptive review but also the deceptive reviewer. Therefore, we tried to examine how personality traits may correlate with deceptive communication style. To further improve the accuracy of the deceptive reviews detection, we have used readability, lexical diversity, subjectivity, extreme emotion and so on, as the features along with personality clues.

Features play a key role in detecting deceptive reviews with the machine learning algorithms. We tried various ensemble learning methods for classification of deceptive and truthful reviews. Ensemble learning principle is based on the assumption that each learning method has some limitations and strengths. So it tries to exploit its strengths and weakness to make a better and informed decision. In this paper, we:

• Have tried to associate personality traits and deceptive review/reviewer across different domains (restaurants, hotels). We are also able to build a deeper picture of how personality behavior is actually realized linguistically.

• Have introduced novelty regarding the feature sets such as readability, lexical diversity, extreme emotions, personality clues etc. and given a very systematic comparison of the feature sets in finding deception by using diverse machine learning models.

• Have explored various homogeneous and heterogeneous ensemble learning models and also shown the comparative analysis with individual classifiers using different feature sets.

The rest of the chapter is organized as follows. The second section describes various works related to the opinion spamming considering different approaches. Section 3 explains feature identification, and construction along with ensemble learning methods used to build the classifiers. Section 4 contains experimental details along with the statistical analysis of the results. The last section comprises of the conclusion as well as the future work.

2. Related work

In recent years, opinion spam detection gained momentum, both in industries as well as in academia. To build a deceptive opinion detection classifier, researchers have to rely on review content [7][8][9], reviewer information [10] such as id, age, geo-location and product characteristics. Initially, the opinion spam detection problem was treated as duplicate review finding. Duplicate-finding attempts to detect frequently iterated reviews posted from the same or different reviewers for one or more products. These approaches tried to find the either the text- based or the concept-based similarities[11]. But in the case of deceptive spam review, the approach fails to work.

Review spamming can be carried out either individually or may involve a group. Reviews of the products arrive randomly. But when the reviews appear in a burst, it might be either due to the sudden popularity of product or spam attack. In one approach author used Kernel Density Estimation to detect review burst and several other review-based features to detect group spammer [12]. In another approach, frequent pattern mining is used to find the potential candidate spammer group. Then they evaluate each group to find a strange and unusual group based on certain characteristics measures such as time window, group deviation, the size of the group etc. Using SVM, they ranked the groups according to the calculated measures. These experiments were conducted on Amazon reviews and results are highly effective.

As earlier studies suggest, ratings have a high influence on revenue. Unfair and biased rating pattern has been studied in several previous works [12][15]. In one of the approaches, the author has identified several characteristic behaviors of the review spammers and model this behavior to detect the spammer [13]. They derived an aggregated behavior scoring method for ranking the reviews according to the degree they demonstrate the spamming behavior. Their study shows that by removing reviewers with very high spam sources, the highly spammed products and product group has experienced significant changes in aggregate rating compared with removing randomly scored or unrelated reviewers.

Some other promising review spam detection methods included duplicate finding methods [11], concept similarity based method [16], content-based method [17][18][19], and review and reviewer-oriented features-base method [20] etc. Spammer detection techniques included graph-based method [21], temporal activity-based method [22] etc.

One of the finest works in the field of deceptive opinion spam identification has been carried out by integrating psychology and computational linguistics [7]. The author claimed that the best performance was achieved by using psychological features and bigrams along with the support vector machine (SVM) and got up to 89 percent accuracy in the study of hotels domain. They have also contributed a large-scale publicly available gold standard data set for deceptive opinion spam research. We have used the same kind of dataset to perform our experiments and used the highest accuracy as a baseline.

3. Features sets and Ensemble Learners

Feature engineering and classifiers play a vital role to detect deception in reviews. This section describes the how personality traits and other linguistic clues are converted as features to train classifiers. A general framework for deceptive opinion detection is shown in figure 1.

3.1 Features identification and construction

Identifying the deception is important as it has serious consequences in terms of credibility as well as monetary loss. Personality traits as features along with other behavioral and linguistic features can play a major role in detecting deceptive reviews.

3.1.1 Personality traits (PT)

In many studies, it has been hypothesized that deception and personality are strongly correlated [23], [24].There is evidence that personality interacts with, and affects the aspects of linguistic production. Personality traits can predict consequential outcomes for individuals such as relationship quality, criminality, deception, volunteerism etc. Some personalities are better in deceiving than others. Examine personality traits and correlate them with deceptive is better than other approaches as it works on the whole narrative rather than a single statement. Identifying personalities based on text have been widely explored and studied using a number of different approaches[5].

Over the last 50 years, the Big Five model [25] has become a standard in psychology and experiments. Using the Big Five model, many studies have shown that different personality traits influence many aspects of task-related individual behavior such as leadership ability, teaching skills, and deception ability etc. Based on various previous studies, we have created a set of linguistic and behavioral clues for each personality. Due to space constraint, only a few of them are shown in table 1.

3.1.2 Readability (READ)

Readability can be defined as: "the degree to which a given class of people find certain reading matter compelling and comprehensible [26]." As per the National Center for Educational Statistics (1993) report average US citizen reads at the 7th-grade level and when it comes to writing it degrades even further. It has been observed from blogs and datasets of reviews that truthful reviewers generally use fewer jargon, simple and familiar words as compared to the deceptive reviewers. To use the readability measure, we have calculated Automated Readability Index (ARI), Coleman Liau Index (CLI), SMOG, Flesch-Kincaid Grade Level (FKGL), Chall Grade(CG) and Linsear (LIN) matrices [27]. All these readability metrics are used as the features.

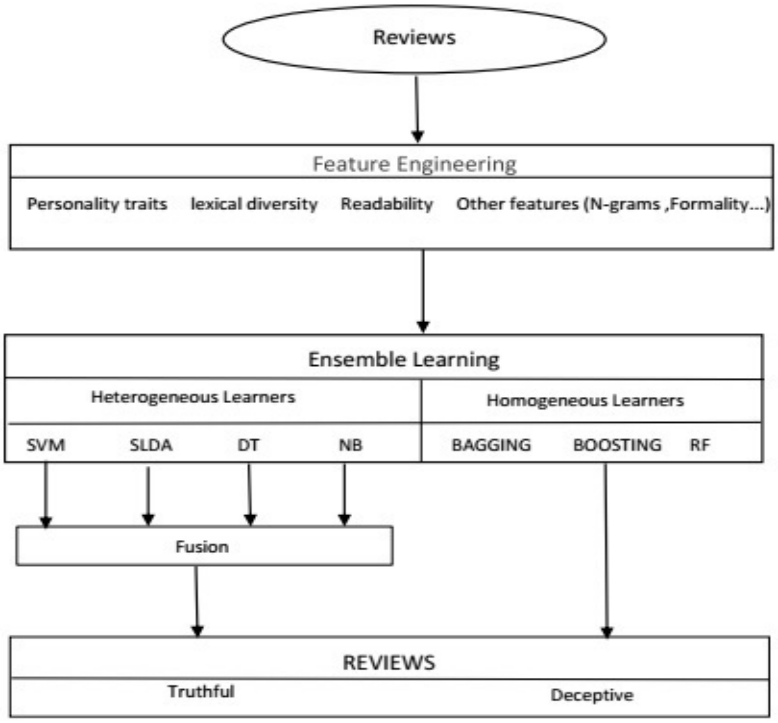

Figure 1. General Framework of Ensemble Learning Based Deceptive Opinion Classifiers

3.1.3 Lexical diversity (LEX)

Lexical diversity is another text characteristic that can be used to distinguish between deceptive and truthful opinions. The more varied vocabulary a text possesses, the higher is the lexical diversity of that text. For a text to be highly lexically diverse, the word choice of the writer needs to be different and diversified with less repetition of the vocabulary. Moreover, previous researchers have shown that lexical diversity is significantly higher in writing than in speaking [28].

Trait	Extraversion		
Type		Introvert	Extravert
Behavior	General Behavior	aloof, reserved, deliberate, shy	Warm, sociable, assertive, playful
	Linguistic and behavioral clues	Use of first person singular negative emotions words use of multiple punctuation expressions high lexical diversity exclusive and inclusive words	Use of references to other people more and adverbs fewer pause,more positive word emotionless formalmore compliments fewer words per statement use of more verb, adverb, pronoun nouns more social words
Trait	Agreeableness		
Type	Agreeableness		Disagreeableness
Behavior	General Behavior	Compassionate, friendly, considerate, cooperative	Suspicious, faultfinding, unfriendly, antagonist
	Linguistic and behavioral clues	Use of first singular nounspositive emotion words more positive and fewer negative wordsfewer articles	More use of articles, negative emotion words discrepancies
Trait	Neuroticism		
Type	Emotional Stability		Neuroticism
Behavior	General Behavior	Calm, unemotional	Insecure, anxious, emotional stability
	Linguistic and behavioral clues	More positive wordsLess negative and emotional words	Use of first person and singular pronoun Singular personMore negative and emotional words
Traits	Openness		
Type	Openness to experience		Closeness to experience
Behavior	General Behavior	Intellectual, appreciate art and ideas, aware of feelings	Conservative, resist change, unimaginative
	Linguistic and behavioral clues	Longer words, words expressing tentativeness (maybe, perhaps), avoidance of past sentences, avoidance of first person singular pronouns	More use of swearing words
Trait	Conscientiousness		
Type	Conscientiousness		Unconscientiousness
Behavior	General Behavior	Disciplined; organized, dutiful; persistent; compulsive; perfectionist.	Spontaneous; careless, impulsive; inefficient.
	Linguistic and behavioral clues	Avoid words which reflect discrepancies (should, would)Avoid negative words Avoid negative emotion words	Use of negations, negative emotion, causation, exclusive words, discrepancies, topics concerned with death

Table 1. Personality traits and Linguistic and behavioral clues

There are various reasons that make the lexical diversity as the most important feature for detecting the deceptive reviews. First, it is highly correlated to extra-version personality [29]. So it can help us to determine the personality trait. Second, it can be used individually as a deceptive clue. We found the high diversity in truthful reviews when comparing it with the deceptive ones. The reason might be that when done individually or as a group, an employee writes more than one review to make a significant impact. So such reviews have higher similarity and less lexical diversity. Not only this issue, but also when they have to write reviews of those products or services of which they are not aware of, then they tend to borrow the vocabulary from the previously written reviews. This phenomenon also leads to low lexical diversity.

We have used Type-Token Ratio, Guiraud's Root TTR,

Dugast's Uber Index, Maas' Indices, Measure of Textual Lexical Diversity, Moving-Average Type-Token Ratio, Carroll's Corrected TTR, Mean Segmental Type-Token Ratio, Summer's index, and Moving-Average Measure of Textual Lexical Diversity.

3.1.4 Other features

Apart from other features, we have used N-grams, Parts of speech to get the context of review and subjectivity, objectivity scores, formality, extreme emotions etc. to accurately judge the personality. *Unigrams* is referred as UG and *unigrams* along with *bigrams* are referred as NG in this paper.

3.2 Ensemble learning methods

The idea of ensemble methodology is to build a predictive model by integrating multiple models. It is well-known that ensemble methods can be used for improving prediction performance. Ensemble learning is a two steps process. Firstly, a number of base learners are trained and then their results are combined to obtain the final output. Ensemble methods differ from each other due to various factors such as:

- Based on inter-classifiers relationship
- Based on type of classifiers used as base learners
- Based on how the output is combined

Firstly, the base learners can be generated independently (in parallel) or in such a way that it may influence the learning of subsequent base learner (in sequential). For example, in boosting, generation of each consecutive classifier depends on its predecessor. Secondly, ensemble learning approach can either use single machine learning algorithm to create a set of homogeneous base learners or multiple algorithms to produce heterogeneous base learners. Finally, ensemble methods can differ on the strategy of combining the classifiers generated by an induction algorithm. The simplest combiner determines the output solely from the outputs of the individual inducers. Majority voting, algebraic combiners, borda count etc. are the most commonly used output combining methods. In this paper, we have used and compared both heterogeneous and homogeneous ensemble learning methodologies.

3.2.1 Ensemble learning with heterogeneous learners

Krogh and Vedelsby[3] have shown that the accuracy of good ensemble approaches depends on upon the diversity and accuracy of base learners. We have chosen diverse machine learning algorithms SVM (Support Vector Machine), SLDA (Stabilized Linear Discriminant Analysis), NB (Naïve Bayes) and Decision Tree (DT) as base learners. Here, we have given a brief introduction to each method that we have used to perform our experiments.

SVM [3] is one of the the most powerful techniques for non-linear classification. It tries to find optimal separating hyper plane between the classes. It uses kernel methods to map the data into higher dimensions using some non-linear mapping. We have used the C++ implementation by Chih-Chung Chang and Chih-Jen Lin with C-classification and RBF kernel. Data are scaled internally to zero mean and unit variance for better class prediction.

$$k(x,x') = (exp(-||x - x_i||^2/ 2\sigma^2) \quad (1)$$

$$F(x) = \sum_{i=1}^{N} \alpha_i y_i (exp(-||x - x_i||^2/ 2\sigma^2) + b \quad (2)$$

$$subjectto\ 0 \leq \alpha_i \leq c;\ i = 1,2,...l$$

SLDA is a linear discriminant analysis based on left-spherically distributed linear scores. We have used the implementation of LDA for q-dimensional linear scores of the original *p* predictors derived from the PCq rule [18].

$$W = (X - \bar{X})^T (X - \bar{X}) \quad (3)$$

NB classifier is a probabilistic classifier based on Bayes rule. It relates the conditional probability to the inverse conditional probability. NB is based on the strong assumption of conditional independence in features given the review class.

The C4.5 algorithm is used to implement **DT**. At each node of the tree, C4.5 chooses one attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. It uses information gain to split the data at each node. Information gain is calculated for remaining attributes and then attribute with highest normalized information gain is used for node splitting.

These learners were generated in concurrent/independent manner. To combine the results, average weighting for regression and majority voting for classification are the most popular methods/schemes. These classifiers were trained in parallel style and their outputs were combined using weighted majority vote. This new classifier is referred as **COM** in this paper.

3.2.2 Ensemble learning with homogenous learners

Here, we have used the most representative ensemble methods like Bagging, Boosting, and Random Forest. Decision tree is used as a base learning algorithm for all three bagging, boosting and random forest. Decision tree produces different generalization behavior with a small change in training set; hence works great for ensemble methods. Here, we have given a brief introduction to each method that we have used to perform our experiments.

Boosting: Freund and Schapiro's AdaBoost is used for implementing boosting algorithm. Initially, all training instances were given equal weights to produce the initial base learner. Subsequently, weights are modified depending on the error rate by previously learned model. Wrongly classified instances were given higher weight in next iteration.

Suppose *D* is the review dataset such as $D = \{(x_1, y_1),$

(x_2,y_2), (x_3,y_3).....(x_t,y_t)}, where x_t is the reviews and y_t is corresponding reviews class and $i \in 1, 2,3....n$. D_i is the weight for i^{th} iteration. $h(x)$ is the learned model and L is the base learning algorithm. $\in_m$ is the error of h_m at m^{th} iteration.

$$D_1(i) = \frac{i}{t} \text{ and } h(x) = L(D, D_m) \quad (4)$$

$$D_{m+1} = \frac{D_m(i)}{Z_m} e^{-\alpha_m y_m h_m(x_t)}$$

Where Z_m = normalization constant (5)

$$\alpha_m = \frac{1}{2} \ln \frac{1-\epsilon_m}{\epsilon_m}$$

where $\epsilon_m = P_{Dm}(h_m(x) \neq y)$ (6)

For a new testing point (x', y')

$$H(x') = sign\ [\sum_{t=1}^{T} \alpha_m h_m(x')] \quad (7)$$

Bagging: The word Bagging is derived from bootstrap aggregating. In bagging, every base learner is trained on different data sample. These bootstrap samples are collected using sub-sampling as the main dataset with replacement.

$$D_m = Bootstrap\ (D); \quad (8)$$

$$h_m = L\ (D_m) \quad (9)$$

$$H(x) = argmax_{y \in Y} \sum_{m=1}^{M} 1\ (y = h_m(x)) \quad (10)$$

Due to sub-sampling, some instances may repeat in D_m and some might not appear even once. In bagging, each instance is chosen with equal probability, while in boosting, instances are chosen with probability proportional to their weight. To classify an unknown instance composite bagged classifier $H(x)$ return the most predicted class based on majority vote.

Random forest: RF is a slight variation of bagging. At each tree split, a random sample of m features is drawn, and only those k features are considered for splitting. Typically

$$k = \sqrt{p}\ or\ log_2 p,$$

Where p is the total number of features (11)

4. Experiments and results

4.1 Experimental setup

As mentioned earlier, we have used the publicly available gold standard deceptive opinion spam corpus for our experiments[7]. This data set is generated through crowd sourcing and the domain expert system. To construct the dataset, the author crawled the truthful reviews of 20 hotels near Chicago from Trip Advisor following the work of Yoo and Gretzel [31]. While to solicit deceptive reviews, they used anonymous online workers also known as turkers. These turkers were told to assume themselves as an employee in the marketing department of the company. These turkers were paid one dollar to write a fake review for the hotel/restaurant. This dataset contains 400 reviews in restaurant domain (200 truthful, 200 deceptive reviews) and 1600 reviews hotel domain(800 truthful, 800 deceptive reviews). They achieved the highest accuracy using NG and psycholinguistic features by SVM. We have used this accuracy as the baseline result for our experiments.

Clue	Features	Source
More words	WC	LIWC
Intellectual	Analytic	LIWC
Personal pronoun	Ppron	LIWC
Impersonal pronouns	Ipron	LIWC
Parts of speech	POS	R packages
Negative words	Negate	LIWC
Swearing Words	SWEAR	LIWC
Formality	Formality score	R packages
Contextual	Informal score	R packages
Positive emotion words	POSEMO	LIWC
Negative emotion words	NEGEMO	LIWC
Subjectivity	Subjectivity Score	Alchemy API
Objectivity	Objectivity score	Alchemy API
Informal language	Informal	LIWC
Vocabulary richness	LEX	R packages
Readability	READ	R packages

Table 2. List of few examples of features for corresponding clues and sources

To build the classifiers, we have extracted 92 text dimensions as text features from LIWC, eighteen parts of speech, twelve metrics of lexical diversity, eight metrics for readability along with unigrams and bigrams from R packages. The Table 2 focus some features and their corresponding sources. We have used some standard feature selection techniques to avoid overfitting, improve accuracy, and reduce training time. Not only that but also some time to include redundant features can be misleading to modeling algorithm. We have used Weka [30] as a feature selection tool. After building the feature sets, Weka's Cfs-Subset selector[31] with Subset forward selection was applied to each one to include only those features that contribute most to accurate classification.

4.2 Results and analysis

Deception is a complicated psychological behavior which is related to cognitive processes and mental activity. We have used personality traits as features to detect deceptive reviews and reviewers. Deceptive reviewers are not very expressive as they might have a fear of being identified. Deceptive reviewers are closed in nature and use more non-committal verbs (guess, suppose, assume etc.) and vague quantifiers (more or less, you might say etc.). Some personality traits such as the extraversion are more reflective in the language. We found strong relations between the extraversion and conscientiousness traits and truthful reviews, and between neuroticism and disagreeableness and deceptive reviews.

Figures 2 and 3 show the performance of each feature sets for hotel and restaurant domain respectively. Personality traits performed significantly better than other two (two tail t-test for p <=0.05) but less than n-grams. The better performance of N-grams shows the importance of language structures to detect deception.

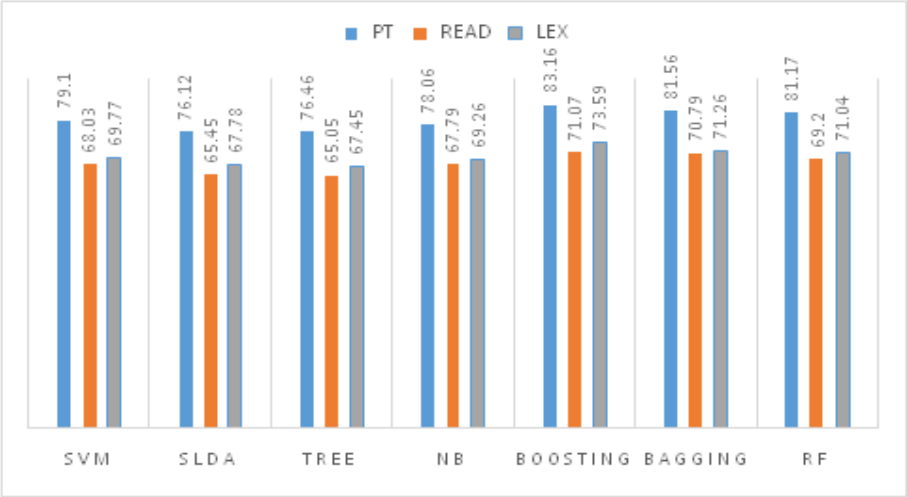

Figure 2. 5-fold cross-validation accuracy averaged over ten runs for Personality traits, Readability and Lexical diversity features set. Boldface indicates the highest accuracy on particular feature set for hotel domain

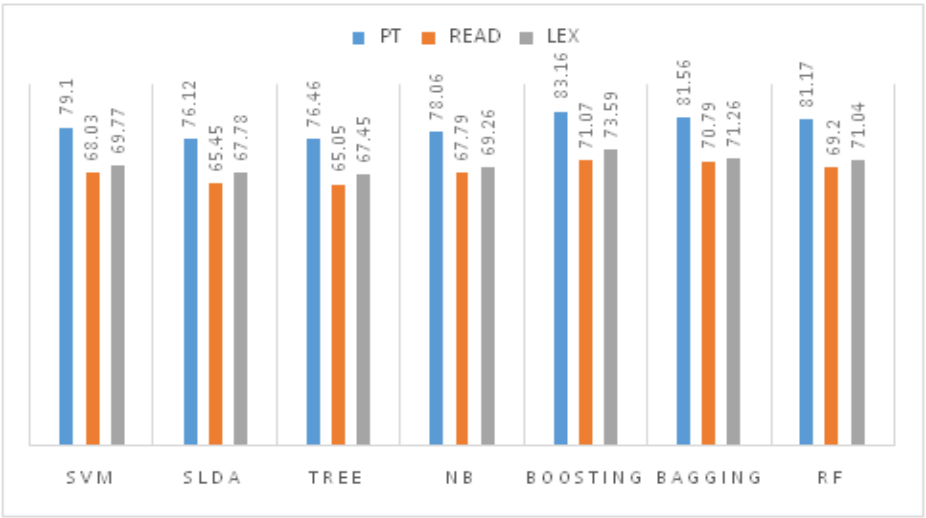

Figure 3. 5-fold cross-validation accuracy averaged over ten runs for Personality traits, Readability and Lexical diversity feature set. Boldface indicates the highest accuracy on particular feature set for restaurant domain

Extrovert and truthful reviews share strong correlation. Moreover, deceptive reviewer behaves more like an introvert. A reviewer who uses more words, less formality score, more positive emotion words and lower anxiety tend to be an extrovert and has a good chance of being truthful. On the other hand, an introvert reviewer is less forthcoming, more formal, higher anxiety and more negative emotion words have greater chances of being deceptive. We also contrasted with some of the findings in previous research. For example, across the studies, it has been found that even though the low lexical diversity is found in extroverts but also in deceptive reviews.

Reviewers with neurotics' characteristics use more 1st person singular pronoun, more frequent and concrete words, and more positive emotion words. While reviewer with agreeable personality traits also uses less negative emotion words and fewer articles. Both neurotics and agreeable personality reviewers are strongly correlated to truthful reviews.

As conscientious reviewers avoid negations and words reflecting discrepancies (e.g., *would* and *should*) tend to be more truthful. Finally, openness to experience is characterized by a preference for longer words and words expressing tentativity (e.g., *perhaps* and *maybe*), as well as the avoidance of 1st person singular pronouns and present tense forms. With all these traits openness to experience have a higher chance of being truthful.

Tables 3 and 4 show the results for hotel and restaurant domains respectively. The tables show the classification accuracy of different individual classifiers along with the ensemble classifier (COM) which has combined the output of these classifiers based on the weighted majority. In most of the occasions, we didn't find any significant difference between SVM and NB performances. But concurrent heterogeneous ensemble learner COM has clearly performed better than all individual/base classifiers. Personality traits based features have clearly made significant improvement in classification accuracy across all classifiers.

Strategy	**Feature set**	**SVM**	**SLDA**	**TREE**	**NB**	**COM**
Baseline	NG, LIWC	89.80				
Text	UG	87.13	85.60	85.15	86.39	89.21
Classification	UG, PT	89.47	86.41	88.90	89.34	91.41
	UG, PT,READ	90.18	87.11	89.23	89.53	92.19
	UG, PT, LEX	90.80	87.92	89.81	89.14	91.13
	UG, PT, READ,LEX	91.43	88.13	90.86	90.84	93.12
	NG	87.90	85.82	86.11	87.81	89.52
	NG,PT	90.70	87.58	88.49	89.90	91.67
	NG,PT,READ	91.09	88.26	89.53	90.34	93.08
	NG,PT, LEX	91.64	89.61	89.06	90.33	92.88
	NG, PT, READ, LEX	**92.14**	**89.89**	**90.55**	**92.05**	**94.42**

Table 3. Automated classifiers 10-fold cross-validation accuracy averaged over ten runs for hotel domain. Boldface indicates the highest value in respective column

Strategy	**Feature set**	**SVM**	**SLDA**	**TREE**	**NB**	**COM**
Text	UG	87.41	79.71	77.25	87.39	89.76
Classification	UG, PT	88.17	82.18	79.23	88.14	90.66
	UG,PT, READ	88.37	83.16	80.22	88.13	91.11
	UG, PT,LEX	89.33	82.86	81.89	88.04	90.73
	UG, PT, READ,LEX	89.64	83.69	82.01	89.54	92.52
	NG	88.04	80.08	79.25	87.81	90.22
	NG,PT	89.19	81.04	80.43	89.40	91.17
	NG,PT,READ	89.88	83.18	82.33	89.11	91.78
	NG,PT, LEX	90.24	84.43	82.15	88.03	92.88
	NG, PT, READ, LEX	**91.55**	**85.18**	**83.51**	**90.55**	**93.10**

Table 4. Automated classifiers 10-fold cross-validation accuracy averaged over ten runs for hotel domain. Boldface indicates the highest value in respective column

Tables 5 and 6 show the results for other ensemble methods for both hotel and restaurant domains respectively. Boosting has performed better than other methods. In our experiments, we have noticed that in most of the cases no significant difference in accuracy between RF and BAGGING. But an advantage of using random forest over bagging and boosting is that it is faster and relatively robust to outliers and noise. All learning models trained only on n-grams have performed comparatively better than those trained on PT, READ, LEX feature set. It shows that context of the documents needs to be considered, and all other feature sets worked as complementary to improve the accuracy further. All classifiers have performed best across domain when trained with NG, PT,READ and LEX combine.

When compared between homogeneous classifiers and heterogeneous ensemble learning classifiers, former clearly has an advantage in classification accuracy. The main advantage of using the heterogeneous learner is the more diversity in learning. Tables 7 and 8 show micro precision, recall, and accuracy for the best-performing method for hotel and restaurant domain respectively.

We achieved the best accuracy of 93.10% in the restaurant and 94.42% in hotel domain using n-grams, personality traits, readability and lexical diversity feature sets. In this study, we also contrasted with some previous findings. For example, a general hypothesis [32] about deceptive persons is that they use fewer self-reference words compared to truthful ones. But in our results, we find the the use of more self-referencing compared with deceptive reviewers. We also found that frequency-based model for weighting n-grams features had shown higher accuracy

Strategy	Feature set	BOOSTING	BAGGING	RF
Text	UG	90.65	89.39	89.12
Classification	UG, PT	90.09	89.34	89.21
	UG, PT,READ	90.23	89.53	89.69
	UG, PT,LEX	91.81	90.14	89.43
	UG, PT, READ,LEX	92.86	91.84	91.12
	NG	91.11	90.81	89.82
	NG,PT	91.49	90.90	90.67
	NG,PT, READ	92.53	92.14	91.08
	NG,PT, LEX	92.06	91.43	91.08
	NG, PT, READ, LEX	**93.35**	**93.05**	**92.50**

Table 5. Automated classifiers 10-fold cross-validation accuracy averaged over ten runs for hotel domain. Boldface indicates the highest value in respective column

Strategy	Feature set	BOOSTING	BAGGING	RF
Text	UG	88.82	87.80	87.61
Classification	UG, PT	89.19	89.90	88.87
	UG, READ	89.13	88.34	88.18
	UG, LEX	89.16	88.44	87.78
	UG, PT, READ,LEX	91.76	90.17	90.81
	NG	89.15	87.92	88.32
	NG,PT	90.39	90.14	89.72
	NG,READ	90.28	89.33	89.69
	NG,LEX	90.42	89.54	89.04
	NG, PT, READ, LEX	**92.12**	**90.75**	**91.23**

Table 6. Automated classifiers 10-fold cross-validation accuracy averaged over ten runs for hotel domain. Boldface indicates the highest value in respective column

Strategy	Feature set	Method	Accuracy	Precision	Recall
Text	UG	COM	89.76	86.75	91.2
Classification	UG, PT	COM	90.66	90.17	91.15
	UG, PT, READ	COM	91.11	92.23	90.71
	UG, PT, LEX	COM	90.73	90.56	90.83
	UG, PT, READ,LEX	COM	90.21	91.52	89.05
	NG	COM	91.76	91.22	92.12
	NG, PT	COM	91.17	89.1	93.13
	NG, PT, READ	COM	91.78	91.18	92.59
	NG, PT, LEX	COM	92.88	93.37	92.15
	NG, PT, READ, LEX	COM	**93.10**	94.56	92.10

Table 7. Micro-averaged accuracy, precision, and recall for top performing classifier with the corresponding feature set for hotel domain

Strategy	Feature set	Method	Accuracy	Precision	Recall
Text	UG	COM	89.21	87.75	91.2
Classification	UG, PT	COM	91.41	90.17	91.15
	UG, PT, READ	COM	92.19	93.23	91.7
	UG, PT, LEX	COM	91.13	91.56	90.83
	UG, PT, READ,LEX	COM	90.21	91.12	89.05
	NG	COM	91.76	90.52	91.92
	NG, PT	COM	91.67	91.88	91.43
	NG, PT, READ	COM	93.08	93.18	93.79
	NG, PT, LEX	COM	92.88	93.17	92.15
	NG, PT, READ, LEX	COM	**94.42**	95.16	94.11

Table 8. Micro-averaged accuracy, precision, and recall for top performing classifier with the corresponding feature set for hotel domain

comparing to Boolean models which give either 0 or 1 weight based on absence and presence of the features.

5. Conclusion and Future Work

This paper mainly aimed at automatic deceptive review detection. Most of the existing opinion spam detection techniques have focused on detecting deceptive reviews. While in this approach, we also tried to identify deceptive reviewers using personality traits. Personality traits were not used in this domain before. We have used standard big five model to study individual behavior. We found strong relations between the extraversion and conscientiousness traits and truthful reviews, and between neuroticism and disagreeableness and deceptive reviews. Some novel linguistic features such as readability, lexical diversity, and formality helped to improve classification accuracy. Both homogeneous and heterogeneous ensemble learning methodologies ensured effective use of features through diversity between classifiers. All ensemble classifiers clearly are shown much higher accuracy in comparison to baseline classifier.

Some personality traits such as extraversion can easily be recognized in the text, while others are not expected. But in the lack of big deceptive review dataset, many personalities could not be analyzed in detail. However, the results shown is certainly suggestive, and expect to broaden the analysis on bigger datasets in future.

References

[1] Pang, B., Lee, L. (2006). Opinion Mining and Sentiment Analysis, *Found. Trends InformatioPang, B., Lee, L. (2006).* Opin. Min. Sentim. Anal. Found. *Trends Inf. Retrieval*, 1(2) 91–231.

[2] Liu, B., Cardie, C. (2014). Sentiment Analysis and Opinion Mining, *CI2014*, No. May.

[3] Luca, M. (2011). Reviews, Reputation, and Revenue:

The Case of Yelp.com, *Business*, p. 1–40.

[4] Bond,C. F., DePaulo, B. M. (2006). Accuracy of deception judgments., *Pers. Soc. Psychol. Rev.*, vol. 10 (3) 214–234.

[5] Phares, E. (1998). Introduction to personality. 2nd ed.

[6] Vrij, A., Granhag, P. A. (2012). Eliciting cues to deception and truth: What matters are the questions asked, *J. Appl. Res. Mem. Cogn.*, 1 (2) 110–117.

[7] Ott, M., Choi, Y., Cardie, C. J. T., Hancock, C. (2011). Finding Deceptive Opinion Spam by Any Stretch of the Imagination, p. 11.

[8] Li, J., Ott, M., Cardie, C. Hovy, E. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam, *Acl-2014*, p. 1566–1576.

[9] Banerjee , S., Chua, A. Y. K (2014). Applauses in hotel reviews: Genuine or deceptive?, *In*: Proc. 2014 Sci. Inf. Conf. SAI 2014, p. 938–94.

[10] Lai, C. L., Xu, K. Q., Lau, R. Y. K., Li, Y., Jing, L. (2010). Toward a language modeling approach for consumer review spam detection,*In*: Proc. - IEEE Int. Conf. E-bus. Eng. ICEBE 2010, p. 1–8, 2010.

[11] Jindal, N., Liu, B. (2008). Opinion spam and analysis, *Proc. Int. Conf. Web search web data Min.* WSDM 08, p. 219.

[12] Fei,G., Mukherjee, A., Liu, B., Hsu, M., Castellanos,M., Ghosh, R. (2013). Exploiting Burstiness in Reviews for Review Spammer Detection, *In*: Proc. Seventh Int. AAAI Conf. Weblogs Soc. Media, p. 175–184.

[13] Lim, E.-P., Nguyen, V.-A., Jindal, N., Liu, B., Lauw, H.W. (2010). Detecting Product Review Spammers using Rating Behaviors,*In*: Proc. 19th ACM Int. Conf. Inf. Knowl. Manag., p. 939–948.

[14] Günnemann,S., Gunnemann, N., Faloutsos, C. (2014). Detecting Anomalies in Dynamic Rating Data/ : A Robust Probabilistic Model for Rating Evolution,*In*:ACM SIGKDD Conf., p. 841–850.

[15] Dellarocas, C (2000). Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems, *In*: Proc. twenty first Int. Conf. Inf. Syst., p. 520–525.

[16] Algur, S. P., Patil, A.P., Hiremath, P. S., Shivashankar, S. (2010). Conceptual level similarity measure based review spam detection,*In*: Signal Image Process. (ICSIP), 2010 Int. Conf., p. 416–423.

[17] Hernández Fusilier, D., Montes-y-Gómez, M., Rosso, P., Guzmán Cabrera, R. (2015). Detecting positive and negative deceptive opinions using PU-learning, *Inf. Process. Manag.*, 51 (4) 433–443.

[18] Hancock, J. T. (2013). Negative Deceptive Opinion Spam, *Naacl*, No. June, p. 497–501.

[19] Costa, H., Merschmann, L.H.C. Barth, F. Benevenuto, F. (2014). Pollution, bad-mouthing, and local marketing: The underground of location-based social networks, *Inf. Sci. (Ny).*,V. 279, p. 123–137.

[20] Li, F., Huang, M., Yang, Y. Zhu, X. (2011). Learning to identify review spam, *IJCAI Int. Jt. Conf. Artif. Intell.*, p. 2488–2493.

[21] Wang, G., Xie, S., Liu, B., Yu, P. S. (2011). Review graph based online store review spammer detection, *Proc. - IEEE Int. Conf. Data Mining, ICDM*, p. 1242–1247.

[22] Ye, J., Kumar, S., Akoglu, L. (2016). Temporal Opinion Spam Detection by Multivariate Indicative Signals, *$Icwsm16*, p. 743–746.

[23] Malmstrom, M. (2015). Personality Traits and Deception Detection Ability Among College Students with Primary Psychopathic Traits.

[24] Oberlander, J., Nowson, S. (2006). Whose thumb is it anyway/ ? Classifying author personality from weblog text, No. July, p. 627–634.

[25] John O. P., John, O. P (1999). The Big-Five Trait Taxonomy: History, Measurement, and Theoretical Perspectives, No. 510.

[26] McLaughlin, G. H. (1969). SMOG grading: A new readability formula, *J. Read.*, 12 (8) 639–646.

[27] DuBay, W. (2008). The principles of readability. 2004, *Costa Mesa Impact Inf.*, p. 77.

[28] Johansson, V. (2008). Lexical diversity and lexical density in speech and writing/ : a develop- mental perspective, *Work. Pap.*, V. 53, p. 61–79.

[29] Heylighen,F., Dewaele, J.-M. (2002). Variation in the Contextuality of Language: An Empirical Measure., *Found. Sci.*, 7 (3) 293–340.

[30] Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I. H. (2009). The WEKA data mining software, *ACM SIGKDD Explor.*, 11 (1) 10–18.

[31] Hall, M. a., Smith, L. a. (1998). Practical feature subset selection for machine learning, *Comput. Sci.*, V. 98, p. 181–191.

[32] DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck,L., Charlton, K., Cooper, H. (2003). Cues to deception, *Psychol. Bull.*,129 (1) 74–118.