# A Computational Approach to Find Deceptive Opinions by Using Psycholinguistic Clues

Mayank Saini[#1], Aditi Sharan [*2]
[#] Jawaharlal Nehru University, New Delhi, India
[1] mayanksaini1986@gmail.com
[*] Jawaharlal Nehru University, New Delhi, India
[2] aditisharan@mail.jnu.ac.in

*Abstract*— **The product reviews and the blogs play a vital role in giving the insight to end user for making a decision. Direct impact of reviews and ratings on the sale of the product raises a strong possibility of fake reviews. E-commerce sites are often indulged in writing fake reviews to promote/demote particular products and services. These fictitious opinions that are written to sound authentic are known as deceptive opinion/review spam. Review spam detection has received significant attention in both business and academia due to the potential impact fake reviews can have on consumer behaviour and purchasing decisions. To curb this issue many e-commerce companies have even started to certify the reviewers. But it covers an only small chunk of reviewers, so this technique couldn't be enough to deal with the problem of deceptive opinion spamming. Manually, it is difficult to detect these deceptive opinions. This work primarily focuses on enhancing the accuracy of existing deceptive opinion spam classifiers using psycholinguistic/sociolinguistic deceptive clues. We have formulated this problem in different ways and solve them with many machine learning techniques. This work carried out up on the publicly available gold standard corpus of deceptive opinion spam and achieved up to 92 percent cross-validation accuracy in restaurants and around 94 percent in hotels domain by the final classifier. A detail comparative results analysis has been done for all used machine learning algorithms.**

**Keyword -** Opinion Spamming, Opinion Mining, Web Mining, Psycholinguistic Features, and Machine Learning

## I. INTRODUCTION

Opinion spamming can be defined as writing fake reviews that try to mislead human readers deliberately by giving undeserving positive opinions or false negative opinions to promote or demote some target products, services or organizations. People with malicious intentions post fake opinions without disclosing their true identity, also known as opinion spammer. Opinion spam can be broadly classified as disruptive opinion and deceptive opinion spam. Most of the previous work has focused on disruptive opinion spam, which is in the form of advertisement and other irrelevant non-opinion text. But deceptive opinion where people intentionally try to mislead others by writing fake reviews, remained a less explored field. Disruptive opinion spam can easily be identified and ignored by the user as they have quite distinguishable features that correspond to the advertisement and other commercial interests. On the other hand deceptive opinions are neither identifiable by a human reader nor even easily ignored as they have a serious impact on revenue generation and reputation. A study conducted on the impact of consumer reviews in restaurant domain finds that one-star increase in Yelp rating leads to a 5-9 percent increase in revenue [1]. Several high-profile cases have been reported in the news. The main motive behind the spamming is the monetary benefits.

Opinion spam classifier can be seen as a two class text classification problem , however it is different from general text classification in terms of features. Traditional text classifiers mainly use syntactic, semantic, statistical etc. feature for classification purpose. Such features may be useful for classifying spam opinions also. But for detecting deceptive opinions, we need to keep in mind that these opinions are intentional, so a link needs to be established between use of regular words and deceptive behaviour to catch spammers.

The problem of linking the opinion and opinion holder (reviewer) behaviour is not an easy task. Moreover, the task becomes more difficult in absence of information regarding opinion holder in most of the cases. To build a spam review detection model, researchers may use reviews or reviewer's characteristics. But in most of the cases and domains, they have to rely on review text due of unavailability of reviewer details. Most opinions are found in form of reviews so opinion and review is used interchangeably.

Word play is deceptive and so is the human being. Review Language plays a major role in identifying the hidden intentions. Our main focus behind this work is to explore the use of psycholinguistic/sociolinguistic features in order to study the deceptive behaviour of the reviewer. A lot of study has been done by the linguists and psychologists to find verbal and non-verbal clues to deception [2] and establish an association between psycholinguistic features and deception. However in our opinion many of these associations have not been utilized for opinion spam detection. On this basis, we propose an intermediate layer where we identified various computational psycholinguistic measures/matrices and identified their association with the deceptive behaviour of the person based on the studies conducted earlier. To achieve our objective, we build with different computational matrices and on benchmark dataset (classified opinions as spam and non-spam), we observed that these measures are significantly different in spam and non-spam reviews. These measures were used as features for training and testing various machine learning models. This work mainly focuses on:

- Formulation of opinion spam detection problem in different ways: genre identification (Informative vs. imaginative writing), linguistic deceptive detection, and traditional text classification problem.
- Use of psycholinguistic/sociolinguistic features such as emotion, negativity, tension, anger, personal concern, tone, etc. to understand the intention of the reviewer.
- Use of readability and lexical diversity as features in the context of opinion spamming. We have observed that these measures can contribute significantly towards detecting deceptive reviews. However, in our knowledge no preliminary study has been reported on the application of these measures in opinion spamming domain.
- Use of SVM (support vector machine), SLDA (stabilized linear discriminant analysis) and ensemble learning techniques to detect opinion spam.

We have performed experiment on restaurant and hotel domain on Myle Ott's gold standard dataset [3]. A comparative study and analysis of each approach and corresponding result is given. The rest of the chapter is organized as follows. The second section describes various works related to opinion spamming considering different approaches. Section 3 explains feature identification, construction and justifies their use both logically and statistically. Section 4 includes problem formulation and classification methodology that we have used in this work for deceptive spam detection. Section 5 contains experimental details along with statistical analysis of the result. The last section comprises of the conclusion as well as the future work.

## II. RELATED WORK

The basic definition of spamming refers to web spam that includes email spam or search engine spam which indulges in the action of misleading search engines to rank some web pages higher than they deserve [4].Going beyond the basic definition, Spamming also includes opinion spamming which is comparatively a new field of research. Though a lot of research is going on into the field of opinion mining and sentiment analysis. But only a few of these studies have focused on opinion spam problem and more specifically on deceptive opinion spam detection. Preliminary research has been reported on Amazon reviews [5]. They re-framed the review spam identification problem as duplicated reviews identification problem. Previous attempts for spam/spammer detection used reviewer's behaviors, text similarity, linguistics features, review helpfulness and rating patterns.

One of the finest works in the field of deceptive opinion spam identification has been done by integrating psychology and computational linguistics [3]. The author claimed that best performance was achieved by using psychological features with support vector machine (SVM) to detect the deceptive spam with accuracy up to 89 percent on hotel domain. They have also contributed a large-scale publicly available gold standard data set for deceptive opinion spam research.

In another approach, author proposed a complementary model to existing approach for finding subtle spamming activities [6]. Thus, it can be combined with other textual feature-based models to improve their accuracy. In their work, authors proposed a novel concept of a heterogeneous review graph and claimed to capture the interrelationship among reviewers, reviews, and stores that the reviewers have reviewed. This model tries to identify suspicious reviewer by exploring nodes of the graph. It also tried to establish the relationship between trustiness of reviewers, the honesty of the review and the reliability of the store. This work has achieved the precision up to 49 percent. However, authors claimed to identify those suspicious spammers that couldn't detect by other existing techniques.

As earlier studies suggest, ratings have a high influence on revenue. Higher rating results in higher revenue. Many companies are indulging in insidious practices to get undue benefits. Unfair and biased rating pattern has been studied in several previous works [7], [8]. In one of the approach author identified several characteristics behavior of review spammer and model this behavior to detect the spammer [9] . They derived an aggregated behavior scoring methods to rank 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.

Another approach may involve capturing the general difference of language usages between deceptive and truthful reviews [10]. This model tried to include several domain independent features that allow formulating general rules for recognizing deceptive opinion spam. They used part of speech (POS), psychological and some other general linguistic cues of deception with SAGE [11] and SVM model. The dataset used in this work include following domains, namely hotel, restaurant, and doctor. SAGE achieved much better result than SVM and were around 0.65 accurate in the cross-domain task. Another model that integrates some deep linguistic features derived from syntactic dependency parsing tree was proposed to discriminate deceptive opinions from normal ones [12]. They worked on Ott's data set and a Chinese data set and claim to produce a state of art results on both of the topics.

Opinion spamming can be done individually or may involve a group [13]. Group spamming can be even more damaging as they can take total control of the sentiment on the target product due to its size. Their work was based on the assumption that a group of reviewers works together to demote or promote a product. The author has used frequent pattern mining to find a candidate spammer group and used several behavioral model derived from the collusion phenomenon among fake reviews and relation models.

## III. THEORETICAL FRAMEWORK FOR FEATURE IDENTIFICATION AND CONSTRUCTION

In our work, we have considered various well-defined readability, lexical diversity and psychological features along with n-grams measures. Each of these measures can be used to characterize the review. These characteristic measures have been used as features of the review. This work is based on the observation that these features help us to distinguish between deceptive and truthful reviews.

### *A. Readability*

The creator of the SMOG readability formula G. Harry McLaughlin defines readability as: "the degree to which a given class of people find certain reading matter compelling and comprehensible [14]." It was in 1937 when US government for the first time decided to grade civilians rather than considering them as either literate or illiterate. According to National Center for Educational Statistics (1993), average US citizen reads at the 7th-grade level and when it comes to writing it degrade even further. It has been observed that a review written by average US citizen contains simple, familiar words and usually, fewer jargons compare to one written by professionally hired spammer. This simplicity and ease of words lead to better readability. In particular, we will test the hypothesis that all else equal, higher readability will be associated with the fewer chances of spam.

Various readability metrics have been suggested to identify the readability of text. Among them, we have considered only a few well establish readability metrics [14], [15]. To be specific, we computed Automated Readability Index (ARI), Coleman Liau Index (CLI), Chall Grade(CG), SMOG, Flesch-Kincaid Grade Level (FKGL) and Linsear (LIN). As a whole, readability features have been referred as READ throughout this paper.

Table 1 below shows the statistical measures for the readability matrices or restaurant domain with respect to truthful and deceptive opinions. Statistics in Table 1 show a significant difference in ARI( two tailed t-test p=0.0045), CLI(two tailed t-test p=0.03), CG(two tailed t-test p=0.02),SMOG(two tailed t-test p=0.01),FKGL(two tailed t-test p=0.01) and LIN(two tailed t-test p=0.03) for truthful and deceptive reviews.

TABLE 1: Descriptive statistics of selected readability measures for Truthful and Deceptive reviews for restaurants

| | **Truthful** | | | | **Deceptive** | | | |
|---|---|---|---|---|---|---|---|---|
| **Readability measure** | **Mean** | **Std dev** | **Min** | **Max** | **Mean** | **Std Dev** | **Min** | **Max** |
| **ARI** | 6.9967362 | 4.5626325 | 1.7455556 | 14.64463 | 5.866139 | 3.2525745 | 1.141977 | 13.51321 |
| **CLI** | 7.0527304 | 4.1398862 | 3.666659 | 14.6672 | 6.2356855 | 3.39331 | 2.915186 | 11.6046 |
| **SMOG** | 10.027345 | 3.7796699 | 6.1291 | 17.87935 | 9.1565359 | 4.1415575 | 3.883918 | 15.90319 |
| **FKGL** | 7.2908371 | 4.6020752 | 3.912698 | 15.68627 | 6.1180497 | 4.8111631 | 3.131442 | 12.54458 |
| **LIN** | 8.9204804 | 4.0913597 | 2.9976 | 24.5 | 7.8725818 | 5.7460003 | 2.916667 | 21.16667 |

Table 2 below shows various readability measures for hotel domain. These statistics show a significant difference in ARI( two tailed t-test p=0.0194), CLI(two tailed t-test p=0.0202), CG(two tailed t-test p=0.0122),SMOG(two tailed t-test p=0.0264),FKGL(two tailed t-test p=0.0353) and LIN(two tailed t-test p=0.03) for truthful and deceptive reviews.

TABLE 2: Descriptive statistics of readability measures for Truthful and Deceptive reviews for hotels

| | Truthful | | | | Deceptive | | | |
|---|---|---|---|---|---|---|---|---|
| **Readability measure** | **Mean** | **Std Dev** | **Min** | **Max** | **Mean** | **Std Dev** | **Min** | **Max** |
| **ARI** | 7.845504 | 4.593358 | 1.326875 | 17.9453 | 6.56951 | 6.162905 | 0.903924 | 16.18293 |
| **CLI** | 8.365558 | 3.690968 | 3.336489 | 14.60216 | 7.382387 | 4.684986 | 2.998978 | 13.230022 |
| **SMOG** | 11.63516 | 5.700851 | 5.985473 | 17.50586 | 10.30323 | 4.839728 | 4.790887 | 16.32454 |
| **FKGL** | 8.956987 | 6.167367 | 3.001217 | 19.04759 | 7.44399 | 7.355717 | 2.598499 | 17.98789 |
| **LIN** | 9.88652 | 8.345347 | 2.166667 | 28.16667 | 8.331239 | 6.218107 | 1.989098 | 25.3245 |

### *B. Lexical Diversity*

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 [16], [17]. According to the different studies, lexical diversity is genre-sensitive [17].

Various search engine optimization(SEO) companies are hired to influence products rating to give undue benefits to hiring companies. They write fake reviews to manipulate customer's opinion about the particular products. When done individually or in a group, an employee writes more than one review to make a significant impact. So these reviews have higher similarity and less lexical diversity. Not only this 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. However, in the case of truth teller, they come with a fresh idea, honest opinion, and experience that leads to higher lexical diversity in comparison to liars.

Numerous metrics for measuring lexical diversity exist and each of them has its pros and cons. For example, the traditional lexical diversity measure is the ratio of different words (types) to the total number of words (tokens), the so-called type-token ratio, or TTR [18]. Text samples containing a large number and tokens give a lower value to TTR and vice versa because of its sensitivity to sample size. While D measure which was developed by Brian Richards and David Malvern [19] is independent of sample size but it is also being criticized for being insensitive to sample size [20].

TABLE 3: Details of selected lexical diversity measures

| **Lexical diversity metrics** | **Abbreviation** | **Author** |
|---|---|---|
| Type-Token Ratio | TTR | Brian Richards , 1987 |
| Moving-Average Type-Token Ratio | MATTR | M.A. Covington, J.D. McFall, 2010 |
| Guiraud's Root TTR | R | P. Guiraud , 1954 |
| Carroll's Corrected TTR | CTTR | J.Carroll, 1964 |
| Dugast's Uber Index | U | D. Dugast, 1978 |
| Summer's index | S | C. J Summers , 1971 |
| Mean Segmental Type-Token Ratio | MS-TTR | C. W. Hess, K. M.Sefton , and R. G.Landry , 1986 |
| Herdan's C | LogTTR | G.Herdan , 1960 |

Even as a traditional classifier feature, lexical diversity can play a significant role. Here we tried to find that how effective lexical diversity is to identify deceptive opinion spam. The combination of all of the lexical diversity metrics is referred as LEX in this paper. Further we will test the hypothesis that all else equal, higher lexical diversity will be associated with the fewer chances of spam.

Table 4 below shows various lexical diversity measures for restaurant domain. Statistics shows a significant difference in TTR( two tailed t-test p=0.0272), CTTR(two tailed t-test p=0.0325), MA-TTR(two tailed t-test p=0.0288),MS-TTR(two tailed t-test p=0.0316),log-TTR(two tailed t-test p=0.0173), R(two tailed t-test p=0.0334), S(two tailed t-test p=0.0247), and U(two tailed t-test p=0.005) for truthful and deceptive reviews.

TABLE 4: Descriptive statistics of lexical diversity for Truthful and Deceptive reviews of restaurants

| | Truthful | | | | Deceptive | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Mean | Std dev | Min | Max | Mean | Std Dev | Min | Max |
| TTR | 0.7001894 | 0.313123 | 0.5040323 | 1 | 0.6408534 | 0.2128566 | 0.4939341 | 0.9 |
| CTTR | 5.4674648 | 4.01663 | 3.019318 | 8.165712 | 4.6138687 | 3.938569 | 3.34664 | 8.389614 |
| MA-TTR | 0.913294 | 0.1188 | 0.823404 | 1 | 0.89181 | 0.0817 | 0.766667 | 0.977273 |
| MS-TTR | 0.9135 | 0.1305 | 0.81666 | 1 | 0.889 | 0.0936 | 0.8032 | 0.9833 |
| LogTTR | 0.9284 | 0.14133 | 0.87233 | 1 | 0.8915 | 0.1665 | 0.874 | 0.9714 |
| R | 8.6321 | 3.1377 | 4.2699 | 11.548 | 7.8735 | 3.92733 | 4.7328 | 11.8647 |
| S | 0.8935 | 0.225 | 0.8132 | 1 | 0.8375 | 0.269981 | 0.78051 | 0.94112 |
| U | 30.1119 | 10.72322 | 14.321 | 82.1991 | 27.407 | 8.2907 | 56.0912 | 6.29078 |

Table 5 below shows lexical diversity measures for hotel domain. These statistics show a significant difference in TTR( two tailed t-test p=0.0307), CTTR(two tailed t-test p=0.0035), MA-TTR(two tailed t-test p=0.008),MS-TTR(two tailed t-test p=0.0067),log-TTR(two tailed t-test p=0.0043), R(two tailed t-test p=0.0012), S(two tailed t-test p=0.0213), and U(two tailed t-test p=0.001) for truthful and deceptive reviews.

TABLE 5: Descriptive statistics of lexical diversity for both Truthful and Deceptive reviews

| | Truthful | | | | Deceptive | | | |
|---|---|---|---|---|---|---|---|---|
| Variable | Mean | Std dev | Min | Max | Mean | Std Dev | Min | Max |
| TTR | 0.8006567 | 0.31765 | 0.5176543 | 1 | 0.7608224 | 0.4128566 | 0.2887961 | 0.89 |
| CTTR | 6.2174638 | 2.919873 | 3.019318 | 8.002322 | 5.38348787 | 4.258569 | 1.14324 | 7.28812 |
| MA-TTR | 0.8867654 | 0.341123 | 0.532104 | 1 | 0.8500281 | 0.19232 | 0.513667 | 0.91233 |
| MS-TTR | 0.892235 | 0.32053 | 0.59166 | 1 | 0.8543329 | 0.23116 | 0.54437 | 0.9611 |
| LogTTR | 0.905234 | 0.12133 | 0.85133 | 1 | 0.8823215 | 0.19125 | 0.813454 | 0.95814 |
| R | 8.232231 | 3.1231 | 5.26569 | 11.148 | 7.6122735 | 4.42433 | 4.72228 | 10.8347 |
| S | 0.881235 | 0.2232 | 0.70312 | 1 | 0.8578765 | 0.18001 | 0.62219 | 0.94092 |
| U | 28.13129 | 11.65434 | 18.13421 | 42.1991 | 26.407232 | 9.32451 | 11.0912 | 39.21028 |

### *C. Psychological and linguistic features*

It's a well-known fact that lying is undesirable, decent people rarely r lie. And this lack of practice makes them a poor liar. While falsehood communicated by mistake are not lies. People lie less often about their actions, experience and plans. And if they do so, they do lie in pursuit of material gain or to escape the punishment. Deception can be defined as a task to mislead others. People behave in quite different ways when they are lying compared to when they are telling the truth. Practitioners and laypersons have been interested in these differences for centuries[21].

In 1981, Zuckerman, DePaulo, and Rosenthal published the first comprehensive meta-analysis of cues to deception [6]. They reported a huge difference of verbal and nonverbal cues occurred in deceptive communications compared with truthful ones.This study shows that liars make a more negative impression and are more tense. Michal Woodworth revealed that liar produced more sense-based words [22]. In other words, deceptive reviewers are more subjective than the truthful ones. Deceptive liars also use fewer self-oriented words (I, me, mine, we, etc.) but more other-oriented words(You, they, etc.). According to study on deception, liars offer fewer details than the truth teller, not only because they have less familiarity with the domain but also to allow for fewer opportunities to be disproved [23].

To fetch psychological features from text reviews, we have used Linguistic Inquiry and Word Count (LIWC) [24] .It is a transparent text analysis program that counts words in psychologically meaningful categories. Empirical results using LIWC (version 2015) demonstrate its ability to detect meaning in a wide variety of experimental settings, including to show attentional focus, emotionality, social relationships, thinking styles, and individual differences. It is among most popular text analysis tool in social sciences. It has categorized its entire output variable into linguistic processes, psychological processes, personal concerns and spoken categories. We have used its linguistic process ($LIWC_{ling}$) and psychological process ($LIWC_{psy}$) feature sets. Psychological and linguistic features of LIWC jointly has been referred as $LIWC_{all}$ in this paper. Table 6 shows a list of few LIWC features.

TABLE 6: List of few examples of LIWC features

| FEATURES | EXAMPLE |
|---|---|
| FUNCT | Total Function words |
| PRONOUN | I, Them, Itself |
| I | I, Me, Mine |
| WE | We, Us, Our |
| YOU | You, Your, Thou |
| Negate | NO, Not, Never |
| SWEAR | Damn, Piss, Fuck |
| AFFECT | Happy, Cried, Abandon |
| POSEMO | Love, Nice, Sweet |
| NEGEMO | Hurt, Ugly, Nasty |
| CERTAIN | Always, Never |
| EXCL | But, Without, Exclude |
| SEXUAL | Horny, Love, Incest |
| RELATIV | Area, Bend, Exit, Stop |
| SPACE | Down, In, Thin |
| DEATH | Bury, Co n, Kill |

### *D. N-Gram*

To get the context of the review we have used unigrams (UG) and bigrams (BG). Some generic preprocessing like removing stop words, extra white spaces are done before generating DTM (Document-Term Matrix). Top UG and BG were filtered based on their term frequency and inverse document frequency score. Jointly we have referred UG and BG as N-gram(NG) in this paper.

## IV. PROPOSED WORK

As discussed earlier this paper primarily focusses to improve opinion spam classifiers accuracy by identifying domain independent lingustic and psycholinguistic features. This section of proposed work is divided in three sub sections,. The first subsection focuses on feature identification and construction with the explanation of their significance for opinion spam detection. The second subsection deals with possible ways for problem formulation and explains different strategies and their corresponding feature set to solve them. The third subsection talks about various classification methods used in this work.

### *A. Problem Formulation*

There are various ways to formulate the problem of detecting opinion spam. Opinion spam can be identified by either using duplicate detection or using classification techniques. Many of the existing literature over opinion spamming have framed opinion spam identification as duplicated opinion identification problem. However, this assumption is not appropriate [33]. Based on the type of spam, this paper reports the study on deceptive opinion spamming. We have tackled the problem to identify opinion spam detection in following three ways.

#### *1) Genre identification between informative vs. creative/imaginative writing*

The problem of finding deceptive opinion spam can be constituted as genre identification task that whether it's imaginative or informative writing. Imaginative writing is quite different from informative writing. Imaginative writing relies heavily on imagination and motive behind it. It includes representation of ideas, feelings and mental images in words.People behave differently, when have to write something that they have not experienced. For example, when you imagine something rather than experiencing it, you tend to be more negative and tense. However, informative writing comprises much of truth, facts, and experience. It primarily provides information through explanation, description, argument and analysis. An imaginative writing might use metaphor to translate ideas and feelings into a form that can be communicated effectively. We can easily relate imaginative writer to the deceptive reviewer who leaves clues such as more sense based words, lesser facts, etc.

Psychological features can play a vital part to distinguish between deceptive and truthful review. People lie most frequently about their feelings and their preference, but less often about their experience, actions and plans. And their lie is clearly visible in their writing when they write a false review about their experience of a product or service. On the other hand,Studies suggest lexical diversity is genre-sensitive [17]. As discussed earlier, vocabulary richness would be higher in informative reviews because of originality in their content. On the other hand when someone tries to write something she/he has not experienced then he might borrow the

words and use them repetitively that leads to low lexical diversity. That's why we have used $LIWC_{psy}$ and LEX feature sets to train our classifiers for genre identification problem.

*2) Linguistic deception detection*

This whole problem can also be treated as linguistic deception detection. It focuses upon how effectively linguistic features alone can detect deception. The study suggests that to the extent that liars deliberately try to control their feelings, expressive behavior, and thoughts. Higher are the chances that their performance would be compromised [2]. They would seem less forthcoming, less convincing, less pleasant and more tense. Deceptive spammer leaves various linguistic cues when lies about something. To obtain linguistic deceptive cues we have used $LIWC_{ling}$ feature set. $LIWC_{ling}$ features subsume most of the linguistic features used in the previous research works. Apart from these, we have also used READ feature set. With both of these feature sets, we have developed our linguistic classifiers for this approach.

*3) Traditional text classification problem*

In a most traditional way, this problem can be constructed as text classification problem using various feature sets. We trained various classifiers with all possible combination of our feature sets. Rather than reporting all classifiers we have enlisted only top performing ones.

*B. Classifiers*

This section describes various machine learning approach used in this work. For the given set of features, we have trained SVM, stabilized linear discriminant analysis (SLDA), random forest (RF), decision tree(DT), neural network(NN), maximum entropy(ME), bagging and boosting for all three approaches mentioned earlier. Out of all these classifiers SVM, SLDA, RF, bagging and boosting performed better than the rest.

SVM [25] is one of the the most powerful technique for non-linear classification. SVM has performed in the related work [5]. It tries to find optimal separating hyperplane between the classes. It uses kernel methods to map the data into higher dimensions using some non-linear mapping. We have used 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.

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

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

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

SLDA is 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].

Apart from SVM and SLDA we also have focused on ensemble methods bagging, boosting and random forest. Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution of all trees in the forest. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them [26] . While bagging combines multiple classification models or same model for different learning sets. In bagging the final classification is the most often predicted class, voting by these classifiers. Boosting also combines the result from multiple classifiers, but it uses to derive weights to combine the predictions from those models into a single prediction or predicted classification. In both bagging and boosting, we have used decision tree as an individual classifier.

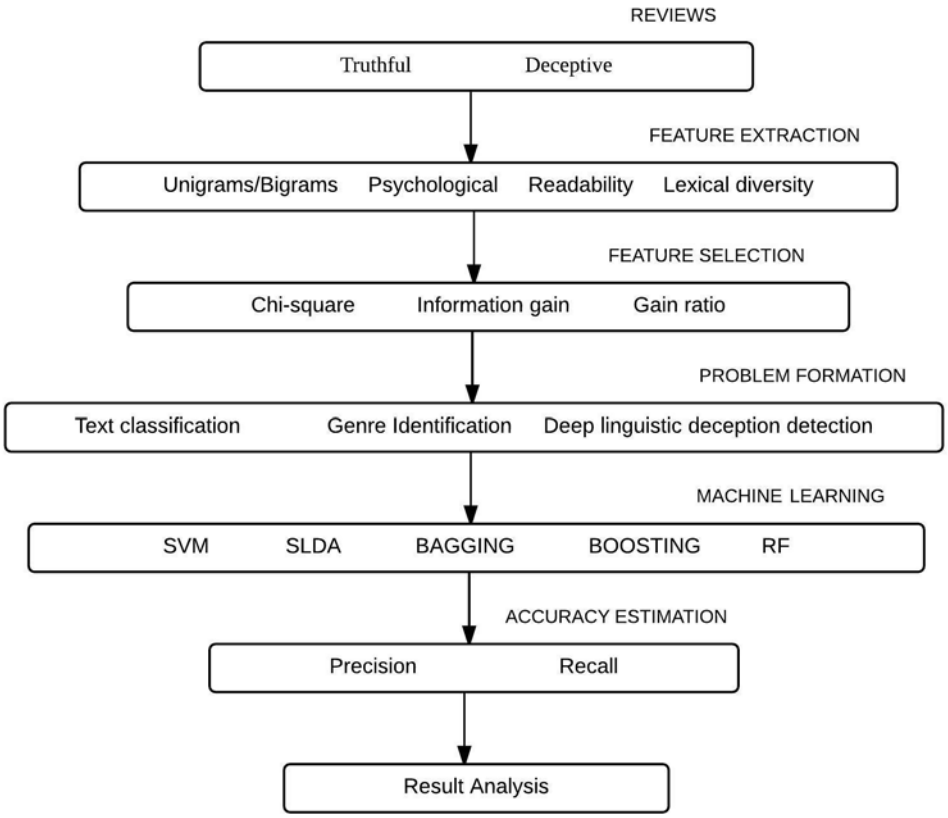


Fig. 1: General framework for detecting deceptive reviews

### *C. Dataset*

As mentioned earlier, we have used the publicly available gold standard deceptive opinion spam corpus for our experiments [3]. This data set is generated through crowdsourcing and domain expert. To construct the dataset, Author mined the truthful reviews of 20 hotel near Chicago from TripAdvisor following the work of Yoo and Gretzel [27]. While, to solicit deceptive reviews, they used anonyms online workers (knowns 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. The earlier version of the dataset has reviews of hotels domain only (400 truthful, 400 deceptive). The current version of the dataset have reviews from restaurant domain (200 truthful, 200 deceptive reviews) and hotel domain(800 truthful, 800 deceptive reviews). We have performed our experiments on a current dataset for both domains. The baseline results are shown on an earlier version of the dataset on hotels domain.

## V. EXPERIMENTS AND RESULTS

This work is an extension of myle ott's work on finding deceptive opinion spam. In that work, NG and psycholinguistic features is used to achieve the best accuracy on SVM. This accuracy we have used as baseline result for our experiments as shown in Table 9. They used psycholinguistic features extracted from earlier version of LIWC (version 2007) which we referred as $LIWC_{old}$ in this paper. Author has performed their experiments on an earlier version of the dataset on hotel domains only.

To build the classifiers we have extracted 92 text dimensions as a text features from LIWC, twelve metrics of lexical diversity, eight metrics for readability along with unigrams and bigrams from R packages. 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 [] as a feature selection tool. We tried every attribute evaluation method available in Weka to select best features. Chi-square, information gain, and gain ratio outperformed others. R software is used for the simulation.

To check the effectiveness of feature sets, we trained classification models for each of them individually. The table 7 and table 8 shows how different feature set performed individually with different learning methodologies for hotel domain and restaurant respectively. In terms feature sets, we find psychological process most effective to differentiate between truthful and deceptive opinions. A Newer version of LIWC features set ( $LIWC_{all}$) is giving a better result than the older version ($LIWC_{old}$). The reason of improved performance is the inclusion of new text dimension such as tone, authentic, informal etc.

TABLE 7: 10-fold cross-validation accuracy averaged over ten runs for psycho-linguistic, readability and lexical diversity. Boldface indicates the highest accuracy on particular feature set for hotels domain

| Feature Set | SVM | SLDA | BOOSTING | BAGGING | RF |
|---|---|---|---|---|---|
| **LIWCold** | 76.80 | | | | |
| **LIWCall** | 77.90 | 74.02 | **81.96** | 80.06 | 80.87 |
| **READ** | 68.03 | 65.45 | **71.07** | 70.79 | 69.20 |
| **LEX** | 69.77 | 67.78 | **73.59** | 71.26 | 71.04 |

TABLE 8: 10-fold cross-validation accuracy averaged over ten runs for psycho-linguistic, readability and lexical diversity. Boldface indicates the highest accuracy on particular feature set for restaurant domain

| Feature Set | SVM | SLDA | BOOSTING | BAGGING | RF |
|---|---|---|---|---|---|
| **$LIWC_{all}$** | 74.90 | 69.20 | **79.60** | 78.60 | 78.16 |
| **READ** | 66.03 | 63.13 | **69.07** | 68.13 | 67.28 |
| **LEX** | 68.87 | 65.87 | **70.13** | 70.78 | 71.14 |

Apart from that, $LIWC_{all}$ feature set is also performing better than LEX and READ also. The difference between these psychological processes supports Zukerman's claim that psychological processes likely to occur more or less often when people are lying compared with when they are telling the truth [28]. To understand this result better, we have to go in what $LIWC_{all}$ subsumes. It determines the degree any text uses positive or negative emotions, self-references, casual words, and 80 other language dimensions. We have also observed a difference in word count, sentence, etc. which also included in this feature set empowers Vrij claim that liars offer fewer details to allow for fewer opportunities to be disapproved [29]. Even though $LIWC_{all}$ is showing good classification accuracy compare to LEX and READ. But $LIWC_{all}$ has more features comparative to both READ and LEX and moreover all these feature sets can work as complementary to each other.

Statistics in Table 1,2 and 4,5 shows a clear difference in readability and lexical diversity between deceptive and truthful reviews. Two tail t-test easily rejected the null hypothesis and showing a significant difference. Classification accuracies on these feature sets also give strength to both our hypothesizes which we have assumed earlier about readability and lexical diversity.

Table 9 and Table 10 shows the result of all three strategies along with their feature set for all learning models for hotel and restaurant domains respectively. All learning models trained only on n-grams have performed comparatively better than those trained on LEX, READ and $LIWC_{all}$ 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. On the other hand among all the classifiers, ensemble methods (mainly bagging and boosting) have outperformed others on most of the occasions.

TABLE 9: Automated classifiers 10-fold cross-validation accuracy averaged over ten runs for all three strategies for the hotel. Boldface indicates the highest value in respective rows

| Strategy | Feature set | SVM | SLDA | BOOSTING | BAGGING | RF |
|---|---|---|---|---|---|---|
| **Baseline** | NG, $LIWC_{old}$ | 89.80 | | | | |
| **Text** | UG | 87.13 | 85.60 | 90.65 | 89.39 | 89.12 |
| **Classification** | UG, $LIWC_{all}$ | 90.47 | 87.41 | 92.09 | 91.34 | 91.21 |
| | UG, READ | 88.18 | 86.11 | 91.23 | 90.53 | 89.69 |
| | UG, LEX | 88.80 | 86.92 | 91.81 | 90.14 | 89.43 |
| | UG, $LIWC_{all}$, READ | 91.23 | 88.96 | 93.22 | 92.54 | 91.66 |
| | UG, $LIWC_{all}$, READ,LEX | 91.77 | 89.13 | 93.86 | 92.84 | 92.12 |
| | NG | 87.90 | 86.82 | 91.11 | 89.81 | 89.82 |
| | NG,$LIWC_{all}$ | 90.70 | 88.58 | 92.49 | 91.90 | 91.67 |
| | NG,READ | 89.09 | 87.26 | 91.53 | 90.14 | 90.08 |
| | NG,LEX | 89.64 | 87.61 | 92.06 | 90.43 | 90.88 |
| | NG, $LIWC_{all}$, READ | 91.10 | 88.72 | 93.81 | 92.55 | 92.09 |
| | NG, $LIWC_{all}$, READ, LEX | **92.24** | **89.99** | **94.55** | **93.05** | **92.99** |
| **Genre Identification Informative vs. Creative Writing** | $LIWC_{psy}$, LEX | 80.39 | 75.18 | 83.29 | 82.35 | 82.13 |
| **Linguistic deception detection** | $LIWC_{ling}$, READ | 77.14 | 73.81 | 80.13 | 79.02 | 79.77 |

TABLE 10: Automated classifiers 10-fold cross-validation accuracy averaged over ten runs for all three strategies for restaurant domain. Boldface indicates the highest value in respective rows.

| Strategy | Feature set | SVM | SLDA | BOOSTING | BAGGING | RF |
|---|---|---|---|---|---|---|
| **Text** | UG | 87.41 | 79.71 | 88.82 | 87.80 | 87.61 |
| **Classification** | UG, $LIWC_{all}$ | 88.87 | 80.58 | 89.19 | 89.90 | 88.87 |
| | UG, READ | 88.17 | 80.16 | 89.13 | 88.34 | 88.18 |
| | UG, LEX | 88.33 | 79.86 | 89.16 | 88.44 | 87.78 |
| | UG, $LIWC_{all}$, READ | 89.11 | 81.37 | 90.21 | 90.07 | 89.09 |
| | UG, $LIWC_{all}$, READ,LEX | 89.74 | 82.69 | 91.76 | 90.17 | 90.81 |
| | NG | 88.04 | 80.08 | 89.15 | 87.92 | 88.32 |
| | NG,$LIWC_{all}$ | 89.69 | 81.04 | 90.39 | 90.14 | 89.72 |
| | NG,READ | 89.18 | 81.18 | 90.28 | 89.33 | 89.69 |
| | NG,LEX | 88.24 | 81.43 | 90.42 | 89.54 | 89.04 |
| | NG, $LIWC_{all}$, READ | 90.12 | 82.26 | 90.82 | 89.84 | 90.06 |
| | NG, $LIWC_{all}$,READ,LEX | **90.71** | **83.80** | **92.12** | **90.75** | **91.23** |
| **Genre Identification Informative vs. Creative Writing** | $LIWC_{psy}$, LEX | 78.44 | 74.66 | 81.51 | 80.22 | 79.10 |
| **Linguistic deception detection** | $LIWC_{ling}$, READ | 76.32 | 73.88 | 79.11 | 78.65 | 77.33 |

Baseline results show 89.90 % accuracy using n-gram and psycholinguistic features on SVM. By adding READ, LEX and advance LIWC feature we got accuracy up to 92.24 % in hotels and 90.71% in restaurant domain using the same SVM classifier. On the other hand, boosting has given the significantly better result as 94.55 % in hotels and 92.12% in restaurants domain. Ensemble methods in general have worked better than other classifiers. SLDA has shown up to 83.80 % accuracy for restaurant domain and 89.90% accuracy for hotels. Boxplot in figure 2 and 3 shows the best performance for each classifier on 10-fold cross-validation using NG, $LIWC_{all}$, READ and LEX feature sets for hotels and restaurants domain respectively.

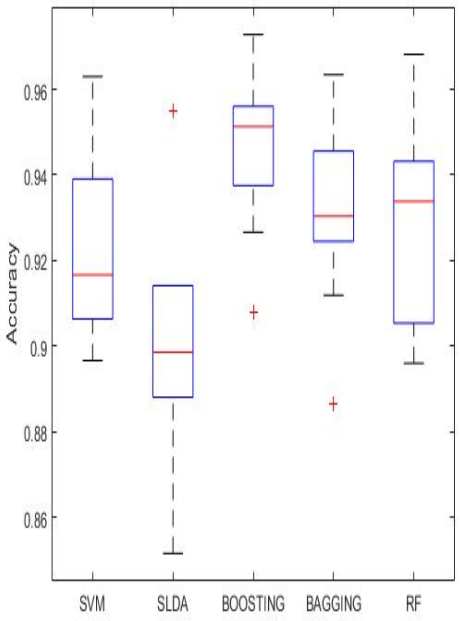


Figure 2: 10-fold cross-validation accuracy for each classifier for restaurant domain

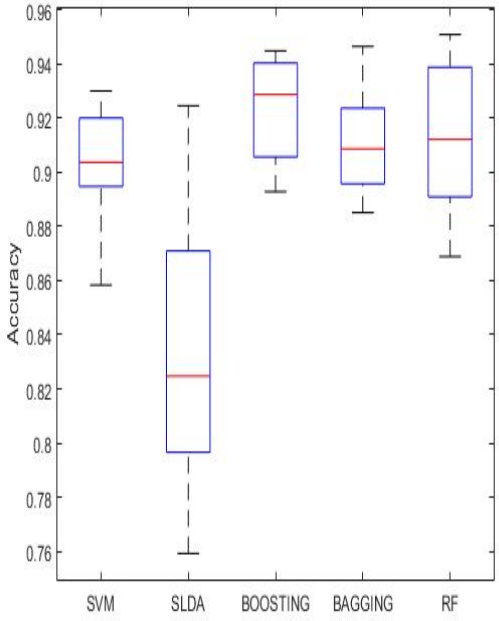


Figure 3: 10-fold cross-validation accuracy for each classifier on for hotel domain

By treating deceptive spam detection as a genre identification task we used only genre-sensitive feature sets $LIWC_{psy}$ and LEX. We achieved accuracy up to 83 % for hotels and 81% for the restaurants . In the previous work[3] they used part of speech (POS) as genre identification feature and achieved up to 73% accuracy for hotels. On the other hand treating the problem as linguistic deception detection and using only linguistic features we achieve up to 80% for hotels domain and 79% for restaurant domain. Table 11 and 12 shows micro precision, recall, and f-score for a best-performing method for all strategies on the respective feature set.

TABLE 11: Micro-averaged precision, recall and f-score for top performing classifier for each strategy and corresponding feature set

| Strategy | Feature set | Method | Accuracy | Precision | Recall |
|---|---|---|---|---|---|
| **Text** | UG | BOOSTING | 88.82 | 86.75 | 91.2 |
| **Classification** | UG, $LIWC_{all}$ | BOOSTING | 89.19 | 89.17 | 91.15 |
| | UG, READ | BOOSTING | 89.13 | 88.23 | 90.7 |
| | UG, LEX | BOOSTING | 89.16 | 91.56 | 88.83 |
| | UG, $LIWC_{all}$, READ | BOOSTING | 90.21 | 89.8 | 91.05 |
| | UG, $LIWC_{all}$, READ, LEX | BOOSTING | 91.76 | 92.65 | 90.82 |
| | NG | BOOSTING | 89.15 | 90.1 | 88.13 |
| | NG, $LIWC_{all}$ | BAGGING | 90.39 | 91.18 | 89.9 |
| | NG,READ | BOOSTING | 90.28 | 88.37 | 91.15 |
| | NG,LEX | BOOSTING | 90.42 | 89.56 | 90.11 |
| | NG, $LIWC_{all}$ ,READ | BAGGING | 90.82 | 92.67 | 88.83 |
| | NG, $LIWC_{all}$ ,READ,LEX | BOOSTING | 92.12 | 93.5 | 91.9 |
| **Genre Identification Informative vs. Creative Writing** | $LIWC_{psy}$, LEX | BOOSTING | 81.1 | 77.6 | 85.73 |
| **Deep psychological Deceptive Detection** | $LIWC_{ling}$, READ | BOOSTING | 85.55 | 82.5 | 87.15 |

TABLE 12: Micro-averaged precision, recall and f-score for top performing classifier for each strategy and corresponding feature set

| Strategy | Feature set | Method | Accuracy | Precision | Recall |
|---|---|---|---|---|---|
| **Text** | UG | BOOSTING | 90.65 | 88.74 | 92.13 |
| **Classification** | UG, $LIWC_{all}$ | BOOSTING | 92.09 | 89.55 | 93.05 |
| | UG, READ | BOOSTING | 91.81 | 88.23 | 90.7 |
| | UG, LEX | BOOSTING | 91.23 | 92.16 | 90.13 |
| | UG, $LIWC_{all}$, READ | BOOSTING | 93.22 | 93.81 | 93.05 |
| | UG, $LIWC_{all}$, READ, LEX | BOOSTING | 93.86 | 93.15 | 92.82 |
| | NG | BOOSTING | 91.11 | 92.12 | 90.33 |
| | NG, $LIWC_{all}$ | BAGGING | 92.49 | 94.08 | 90.91 |
| | NG,READ | BOOSTING | 91.53 | 92.37 | 91.15 |
| | NG,LEX | BOOSTING | 92.56 | 92.16 | 93.01 |
| | NG, $LIWC_{all}$ ,READ | BOOSTING | 93.81 | 94.17 | 93.43 |
| | NG, $LIWC_{all}$ ,READ,LEX | BOOSTING | 94.55 | 93.5 | 95.91 |
| **Genre Identification Informative vs. Creative Writing** | $LIWC_{psy}$, LEX | BOOSTING | 83.29 | 81.16 | 85.73 |
| **Deep psychological Deceptive Detection** | $LIWC_{ling}$, READ | BOOSTING | 86.13 | 85.15 | 87.05 |

In our experiments, we have noticed that in most of the cases no significant difference in accuracies between RF and SVM. And an advantage of using random forest over bagging and boosting is that it is faster and relatively robust to outliers and noise. Apart from that, it gives an internal estimation of correlation and importance of the feature which has been shown in Table 13.

TABLE 13: Shows the top weighted features from each category given by RF.

| Categories | | | | |
|---|---|---|---|---|
| **UNIGRAMS** | **BIGRAMS** | **LIWC** | **LEXICAL DIVERSITY** | **READABILITY** |
| Definitely | will definitely | Pronoun | C | SMOG |
| Recommend | signature room | Space | K | CLI |
| Restaurant | highly recommend | Funct | CTTR | LIN |
| Atmosphere | grill restaurant | Posemo | | ARI |
| Back | first time | Negemo | | |
| Greeted | definitely recommend | Sexual | | |
| Bar | purple pig | I | | |
| Good | recommend restaurant | Swear | | |
| Love | will back | Affect | | |
| Anyone | joes seafood | We | | |
| Prime | gibsons bar | Certain | | |
| Great | can't wait | | | |
| Tasty | food delicious | | | |
| Experience | just right | | | |
| Food | food delicious | | | |

In this study, we also contrasted with some of the findings in previous research. .For example, across the studies, it has been found that deceptive statements are moderately descriptive and distanced from self compared to truthful ones [30]. In the case of deceptive reviews, we found less total word count and sentence but more self-referencing. Deceptive reviews are less descriptive, and the reason behind it might be the fear of being caught. It also has been observed that to make an impact, spammers either go extremely positive or extremely negative. A clear difference has been seen in negative and positive feature values in both types of reviews.

By using different linguistic measures, researchers found that non-naive individuals assigned to be deceptive compared with naive individuals who were truthful showed less diversity and complexity [31]. Our study also supports both the claims as we also found fewer exclusion words that are also a marker of complexity in deceptive reviews and less diversity because in the lack of real experience the spammer borrow experience from other reviewers.

## VI. CONCLUDING REMARKS

It's a widely accepted fact that deceptive spam detection is difficult to detect manually. In this work, we have trained a automated classifier with high accuracy using domain-independent features. We have discovered the relationship between deceptive opinions and lingustic features like readability, lexical diversity. This work has shown different ways to form the problem of deceptive spam detection and effective strategies to solve them. A detail experiment and analysis has been shown for various machine learning algorithms. This paper made many theoretical contributions and contrasted some deceptive assumptions and also strengthen many.

Spammers are getting smart every day that's why for future, both domain specific and independent deceptive clues needed to be discovered. One of the possible future direction to evaluate these deception clues to other domains.

## REFERENCES

[1] M. Luca, "Reviews, Reputation, and Revenue: The Case of Yelp.com," Business, pp. 1–40, 2011.
[2] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," Psychol. Bull., vol. 129, no. 1, pp. 74–118, 2003.
[3] M. Ott, Y. Choi, C. Cardie, and J. T. Hancock, "Finding Deceptive Opinion Spam by Any Stretch of the Imagination," p. 11, 2011.
[4] H. Garcia-molina, "Web Spam Taxonomy," 2008.
[5] N. Jindal and B. Liu, "Opinion spam and analysis," Proc. Int. Conf. Web search web data Min. WSDM 08, p. 219, 2008.
[6] G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection," Proc. - IEEE Int. Conf. Data Mining, ICDM, pp. 1242–1247, 2011.
[7] C. Dellarocas, "Mechanisms for coping with unfair ratings and discriminatory behavior in online reputation reporting systems," Proc. twenty first Int. Conf. Inf. Syst., pp. 520–525, 2000.
[8] G. Wu, D. Greene, B. Smyth, and P. Cunningham, "Distortion as a Validation Criterion in the Identification of Suspicious Reviews," Proc. 1st Work. Soc. Media Anal., pp. 10–13, 2010.
[9] E.-P. Lim, V.-A. Nguyen, N. Jindal, B. Liu, and H. W. Lauw, "Detecting Product Review Spammers using Rating Behaviors," Proc. 19th ACM Int. Conf. Inf. Knowl. Manag., pp. 939–948, 2010.
[10] J. Li, M. Ott, C. Cardie, and E. Hovy, "Towards a General Rule for Identifying Deceptive Opinion Spam," Acl-2014, pp. 1566–1576, 2014.
[11] J. Eisenstein, A. Ahmed, and E. P. E. Xing, "Sparse additive generative models of text," Proc. 28th Int. Conf. Mach. Learn., pp. 1041–1048, 2011.
[12] Q. Xu and H. Zhao, "Using Deep Linguistic Features for Finding Deceptive Opinion Spam.," COLING (Posters), no. 61170114, pp. 1341–1350, 2012.
[13] A. Mukherjee, B. Liu, and N. Glance, "Spotting fake reviewer groups in consumer reviews," Proceeding WWW '12 Proc. 21st Int. Conf. World Wide Web, pp. 191–200, 2012.
[14] W. DuBay, "The principles of readability. 2004," Costa Mesa Impact Inf., p. 77, 2008.
[15] N. Wright, "Towards a better readability measure–The Bog index," Retrieved June, pp. 1–16, 2009.
[16] M. A. K. (Michael A. K. Halliday, Spoken and written language. Oxford University Press, 1989.
[17] V. Johansson, "Lexical diversity and lexical density in speech and writing : a develop- mental perspective," Work. Pap., vol. 53, pp. 61–79, 2008.
[18] B. Richards, "Type/Token Ratios: what do they really tell us?," J. Child Lang., vol. 14, no. 2, p. 201, Jun. 1987.
[19] D. Malvern, Lexical diversity and language development : quantification and assessment. Palgrave Macmillan, 2004.
[20] P. M. McCarthy and S. Jarvis, "vocd: A theoretical and empirical evaluation," Lang. Test., vol. 24, no. 4, pp. 459–488, Oct. 2007.
[21] P. V. Trovillo, "A History of Lie Detection," J. Crim. Law Criminol., vol. 29, no. 6, p. 848, 1939.
[22] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On Lying and Being Lied To: A Linguistic Analysis of Deception in Computer-Mediated Communication," Discourse Process., vol. 45, no. 1, pp. 1–23, 2007.
[23] A. Vrij, K. Edward, K. P. Roberts, and R. Bull, "Detecting Deceit via Analysis of Verbal and Nonverbal Behavior," J. Nonverbal Behav., vol. 24, no. 4, pp. 239–263, 2000.
[24] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The Development and Psychometric Properties of LIWC2015," 2015.
[25] X. Zhang, "Support vector machines," Encycl. Mach. Learn., vol. 1, pp. 941–946, 2010.
[26] L. Breiman, "Random Forests," J. Mach. Learn., vol. 45, no. 1, pp. 5–32, 2001.
[27] U. Gretzel and K. H. Yoo, "Use and Impact of Online Travel Reviews," in Information and Communication Technologies in Tourism 2008, Vienna: Springer Vienna, 2008, pp. 35–46.
[28] M. Zuckerman, R. Driver, and A. Siegman, "Telling lies: Verbal and nonverbal correlates of deception," nonverbal Behav., 1985.
[29] A. Vrij and P. A. Granhag, "Eliciting cues to deception and truth: What matters are the questions asked," J. Appl. Res. Mem. Cogn., vol. 1, no. 2, pp. 110–117, 2012.
[30] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: predicting deception from linguistic cues," Personal. Soc. Psychol. Bull., vol. 29, no. 5, pp. 665–675, 2003.
[31] L. Zhou, J. K. Burgoon, J. F. Nunamaker, and D. Twitchell, "Automating Linguistics-Based Cues for Detecting Deception in Text-Based Asynchronous Computer-Mediated Communications," Gr. Decis. Negot., vol. 13, no. 1, pp. 81–106, Jan. 2004.

## BIOGRAPHY

Mayank Saini is currently a Ph.D. research scholar at School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India. His research interests include Text Mining, Opinion Spamming and data mining etc. He received his M. Tech degree in Computer Science and Technology from School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India in 2012. Mr. Saini has published papers in International Journals and Conference including Springer and IEEE.

Dr. Aditi Sharan has been working as an assistant professor for the past 12 years at the School of Computer and Systems Sciences, Jawaharlal Nehru University, India. She has a doctoral degree in computer science. She is involved in teaching undergraduate and graduate courses like database management, information retrieval, data mining, natural language processing and semantic web. She has published several research papers in international conferences and journals of repute