FINDING AND CLASSIFYING THE DECEPTIVE SPAM REVIEWS USING LIWC DICTIONARY VARIABLES AND DECISION TREE CLASSIFIER

Pandi. Chiranjeevi[#1], D. Teja Santosh[*2], B. Vishnu Vardhan[#3]
[#]Research Scholar, CSE, JNTUCEH, JNTU, Hyderabad, India
[*]Assistant Professor, CSE, GITAM, Hyderabad, India
[#]Professor of CSE, Vice-Principal of JNTUHCEM , Peddapally, India
[1]chiruanurag@gmail.com, [2]tejasantoshd@gmail.com, [3]mailvishnuvardhan@gmail.com

Abstract: **Now-a-days, online reviews in the e-commerce website are increasingly written by the consumers of the product. These reviews have become an important source of information for the new customers to research about these products online. The curious customer research often leads to decision making towards purchasing the product. However, these e-commerce websites still contain deliberately misleading reviews (also called as deceptive spam reviews). Finding these reviews from the huge reviews collection is not an easy task. In order to find whether a review is spam or not, the reviews are pre-processed first. Then, typed dependency parsing is carried out on the pre-processed reviews. Further, the context and sense of these words are understood. The SentiWordNet scores are assigned to these senses and the sense of the review is determined. The deviation of the review itself cannot distinguish the spam nature of the reviews. Additionally, Linguistic Inquiry and Word Count (LIWC) dictionary categories namely social words and cognitive process words scores are analyzed on the review words. In addition to these categories, the summary variables namely clout and authenticity scores from LIWC are also analyzed. Finally, each review's deviation, social words score, cognitive words score, clout score and authenticity score are considered as deep linguistic features and are provided as parameters to the decision tree classification algorithm to learn the spam review classifier.**

Keywords: spam reviews, typed dependency parsing, SentiWordNet score, LIWC, dictionary categories, summary variables, deep linguistic features, decision tree classification

1. INTRODUCTION

With the rapid growth of e-commerce over the past two decades, more and more products are sold on the web. Many people are purchasing the products online. In order to enhance the customer shopping experience, the online merchants have enabled their customers to write reviews on the products that they have purchased and experienced. These reviews contain crucial pieces of information namely the product features and their opinion words. On the other side, the increase in the customer reviews has paved way for some of the group of writers to provide inappropriate or fraudulent opinionated reviews. Such kind of reviews is called as deceptive spam reviews [1]. These reviews give the wrong impression to readers who are intended to purchase the products online.

Identifying these spam reviews from the huge reviews collection is not an easy task. In order to find whether a review is spam or not, the reviews are pre-processed first. Then, typed dependency parsing is carried out on the pre-processed reviews. Further, the context and sense of these words are understood. The SentiWordNet [2] scores are assigned to these senses and the sense of the review is determined. The deviation of each review from the sense of the review is calculated to identify the spam review.

The deviation of the review itself cannot distinguish the spam nature of the reviews. However, when the Linguistic Inquiry and Word Count (LIWC) [3] dictionary categories scores and the scores of newly available summary variables from LIWC 2015 version are analyzed, the first research question is accentuated as follows.

Are the social words, cognitive process words, clout representative words and authenticity based word occurrences among the review sentences help the machine to identify the deliberately written reviews?

This research question implies the understanding of the need of newly available summary variables from LIWC 2015 dictionary for the task of spam reviews analysis.

Finally, each review deviation, social words score, cognitive words score, clout score and authenticity score are considered as deep linguistic features and are provided as parameters to the decision tree classification algorithm to learn the spam review classifier.

The organisation of the paper is as follows: The contributions in this direction are critically reviewed in Section 2, the description of the dataset used in this work is written in Section 3, the proposed method is explained in Sections 4, the experimental results and discussion is explained in Section 5, and finally, conclusions and future work are specified in Section 6.

2. RELATED WORK

Analysis of online opinionated reviews is a popular research topic over the last two decades. Also, a good amount of research has been carried out on spam reviews identification and classification. Jindal and Liu specified [1] that opinion spam is widely spread on the Internet. This is the first work on spam reviews analysis. These researchers worked on spam reviews by using duplicate spam reviews as positive training examples and other reviews as negative examples. They built logistic regression classifier. They reported the model accuracy as 98.7% with the included text features and without feedbacks. Wu et al. worked [4] on opinion spam reviews by making the comparisons among their popularity rankings. Ott et al. collected [5] a gold-standard dataset of 400 truthful and 400 deceptive opinionated reviews and implemented Naive Bayes and SVM based automatic deception classifiers using these reviews. Chen et al. introduced [6] spam detection in Chinese forums. The researchers have analyzed the writings of deceptive writers and identified them using machine learning classifiers. Spam writer identification is also investigated by Lim et al.[7] and Mukherjee et al. [8].

The researchers in their work [5] used LIWC 2007 version dictionary to detect the deceptive reviews. The remaining research works [4,6,7,8] concentrated on detecting the deceptive reviews using various statistical features from the reviews. As opposed to LIWC 2007 dictionary which is available in several languages, the LIWC 2015 version is exclusively an original English dictionary version. The summary variables in 2015 version are the non-transparent dimensions that provide the psychometric scores based on the word co-occurrences in the English sentences.

Also, all the above works never concentrated on analyzing what types of features are the most beneficial (feature subset selection method) for efficient classification of deceptive reviews by the machine. This was also specified as a research requirement in the survey carried out on review spam detection using machine learning techniques by Michael et al. [9]. This never intervened research requirement motivated us to state the second research question as follows.

Does the usage of the machine learning algorithm upon the considered four linguistic features helps to identify the most beneficial features towards efficient classification of deceptive reviews by the machine?

Different from the above works, the current work focuses on understanding the four important types of words used in the reviews which stand as deep linguistic features towards classifying the deceptive reviews. The learned decision tree machine learning model upon these features in this work is also viewed as the basic heuristic method for attribute subset selection.

The research carried out by Ott et al. in their work [5] used 15 types of words summarized under four categories as psycholinguistic features to classify spam reviews. The researchers of the current work argue that only four types of words are enough to classify the spam nature of the reviews. These four features also improve the accuracy of the trained machine learning classifier. This argument raised the final research question as follows.

Are the considered four types of words as psycholinguistic features enough to classify the spam nature of reviews and thereby improve the accuracy of the trained machine learning classifier?

3. DESCRIPTION ON THE COLLECTED DATASET

3.1 GOLD STANDARD ENGLISH DATASET

Ott et al. created [5] gold-standard English reviews dataset in which both truthful and deceptive opinions are provided. The dataset was prepared by them by collecting the genuine reviews from TripAdvisor website. These reviews have the following characteristics. These are namely reviews rated with 5 stars, only English reviews, more than 150 characters and are not written by the first time authors.

The truthful deceptive reviews also known as positively polarized spam reviews were collected from Amazon Mechanical Turk (AMT). The AMT is an online marketplace where workers select their tasks from the available load of tasks and work according to their convenient time. These workers get paid for the work done as per the need of the requesters. The turkers were provided with the name and website of a hotel and were asked to craft their deliberate reviews.

After filtering out the insufficient quality reviews (reviews whose length is less than 150 characters and written by first time authors), 400 golden deceptive reviews were created. These opinions were used as deceptive reviews set.

4. DECEPTIVE SPAM REVIEWS CLASSIFICATION USING LIWC AND DECISION TREE CLASSIFIER

4.1 REVIEWS PRE-PROCESSING

During the pre-processing of gold standard reviews dataset, initially the dataset is divided into individual reviews. Then each review is broken down to a list of words. Subsequently, the stop words that are used across all the reviews are removed as they cannot actually infer any meaning. The stop words are compiled from the reviews itself. This compilation is carried out by sorting the terms in the decreasing order of collection frequency and thereby hand-filtering those terms for their semantic content relative to the reviews domain.

The detailed analysis of the gold standard reviews dataset revealed that the capital words and exclamation marks are very useful indicators for learning about deceptive reviews [9]. So, in pre-processing stage, these are not removed.

4.2 TYPED DEPENDENCY PARSING OF PRE-PROCESSED REVIEWS

The process of identifying spam reviews is carried out initially by understanding the context and sense of review words. Once the sense of the word is finalized, the SentiWordNet scores are assigned to these words and the overall sense of the review is statistically determined. For the considered deceptive review sentence “The rooms are BEAUTIFUL and the staff very attentive and wonderful!!” the sense of each and every review word is determined from WordNet. To do this, the context of each word is identified from the sentence using sentence dependency parsing. The typed dependency parsing for the given sentence is shown in below figure.

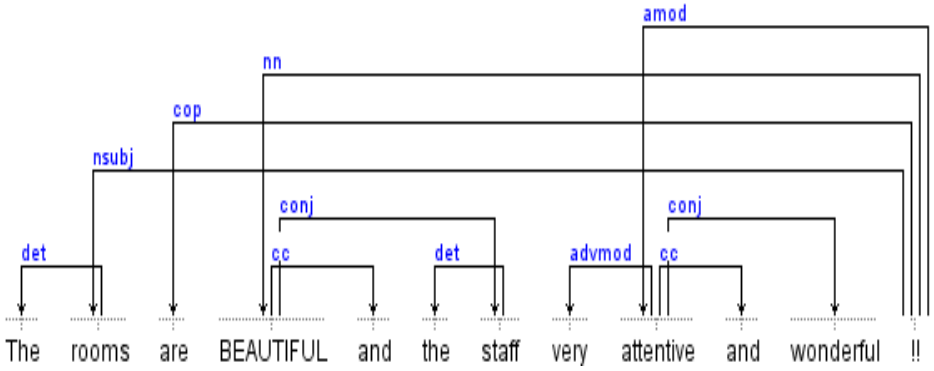

Figure 1. Type dependencies in the deceptive review sentence

From the generated word dependencies in the sentence the grammatical relations nsubj(), nn(), conj(), advmod(), amod() are the contextual clues for the words rooms, BEAUTIFUL, staff, very, attentive, wonderful. The window size of n + 1 words (where n=0,1,2,...) till that word is selected as the contextual clue. With these clues the sense value for each word is finalised by using WordNet sense similarity software package [10]. The SentiWordNet objective score for the word room is 0. The SentiWordNet positive score for the word BEAUTIFUL is 0.75. The SentiWordNet objective score for the word staff is 0. The SentiWordNet positive score for the word very is 0.25. The SentiWordNet positive score for the word attentive is 0.5. The SentiWordNet positive score for the word wonderful is 0.75. The overall sense of the review is calculated as;

$$Sense(\text{Re}\,view) = \frac{\sum_{i=1}^{n} wordsense\_swnscore(w_i)}{Numberofwords\,\text{in}\,thereview} \quad \ldots (1)$$

The overall sense of the review sentence is 0.375.

For the considered truthful review sentence “The rooms are modern and very comfortable”, the typed dependency parsing for the given sentence is shown in below figure.

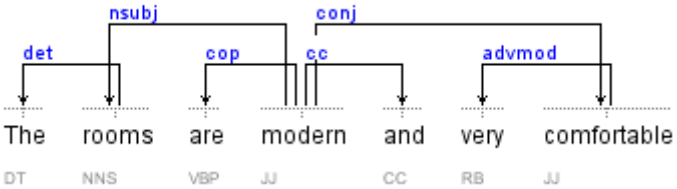

Figure 2. Type dependencies in the truthful review sentence

From the generated word dependencies in the sentence the grammatical relations nsubj(), conj(), advmod() are the contextual clues for the words rooms, modern, very, comfortable. The window size of n + 1 words (where n=0,1,2,...) till that word is selected as the contextual clue. With these clues the sense value for each word is finalised by using WordNet sense similarity software package. The SentiWordNet objective score for the word room is 0. The SentiWordNet positive score for the word modern is 0. The SentiWordNet positive score for the word very is 0.25. The SentiWordNet positive score for the word comfortable is 0.75. The overall sense of the review sentence is 0.25.

The average value of the two review senses is calculated. The value is 0.3125. As there are only two reviews that are considered for this analysis, the standard deviation of each review is found by taking the difference between the overall sense of the review sentence and the average value of the two review senses and squaring it. The result of the line plot of two reviews deviations is shown in below figure.

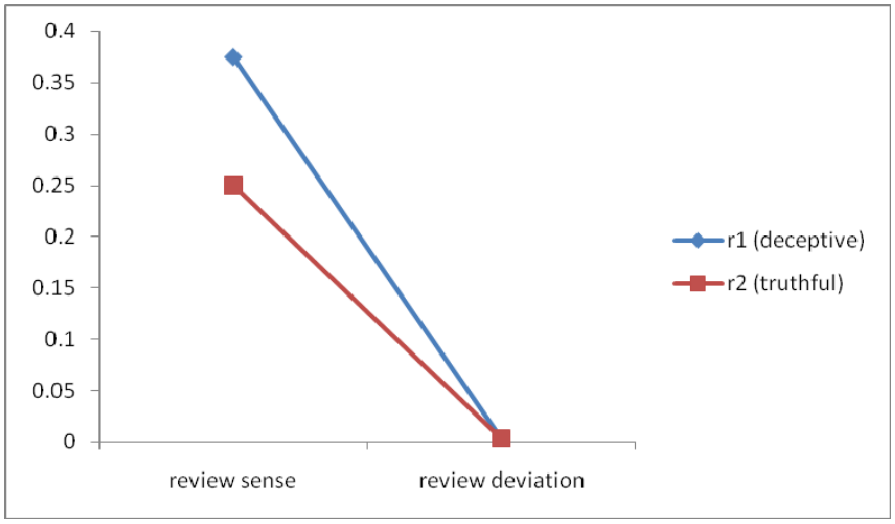

Figure 3. Reviews sense and deviation comparison

It is observed from the above figure that the deviation values of these two reviews overlap with each other. Hence, it is concluded that mere values of deviations of the reviews cannot distinguish the spam review with truthful one.

4.3 SUPPORT FROM LINGUISTIC INQUIRY AND WORD COUNT (LIWC) DICTIONARY CATEGORIES AND SUMMARY VARIABLES

The deviation of the review itself is not enough to say whether the review is a truthful review or a spam review. In order to determine clearly the class of the review to which it belongs, the Linguistic Inquiry and Word Count (LIWC) dictionary is considered which provides additional information on the words like beliefs, thinking patterns, social relationships and personalities that are written in the sentence.

The LIWC 2015 version dictionary outputs 90 summary variables for the given natural language text. For the considered gold standard reviews, it is observed that two summary language variables scores namely clout and authenticity and two psychological constructs scores namely social words and cognitive words are reducing the uncertainty in the classification of the reviews.

4.4 EMPIRICAL EVALUATION OF THE FOUR LIWC VARIABLES

4.4.1 Social Words:

These are the large group of words in LIWC that suggests human interaction. The analysis on the truthful reviews and deceptive reviews under positive polarity reviews with respect to social words specified that 60% of the truthful reviews have less social words written as a part of the review. The analysis on the truthful reviews and deceptive reviews under negative polarity reviews with respect to social words specified that 70% of the deceptive reviews have more social words written as a part of the review. This clearly specifies that the truthful reviews contain more information on the personal experience obtained on a particular thing.

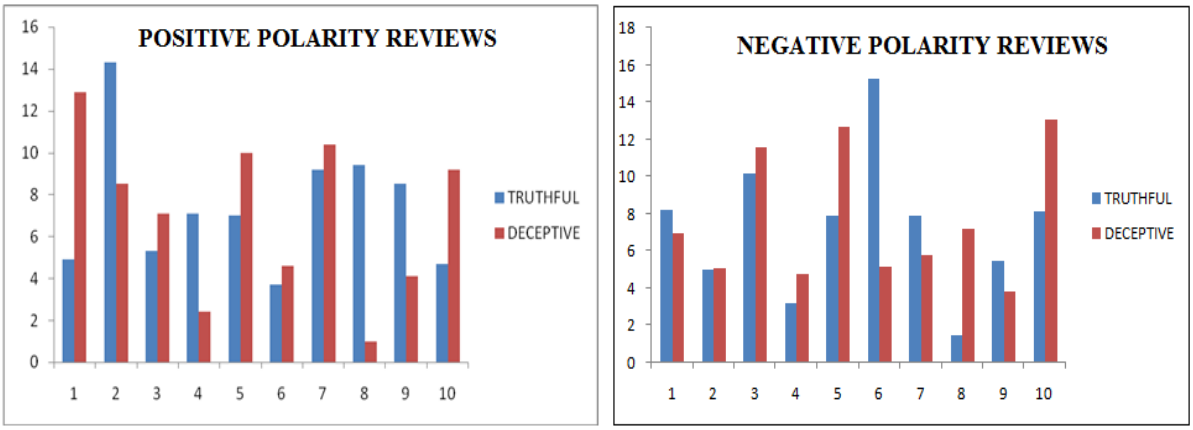

Figure 4. Social words scores comparsion between positive polarity reviews and negative polarity reviews

4.4.2 Cognitive process:

The words used by the reviewers in writing the reviews based on their knowledge levels. The analysis on the truthful reviews and deceptive reviews under positive polarity reviews with respect to cognitive processes specified that 50% of the truthful reviews are written with fewer cognizance based words. The analysis on the truthful reviews and deceptive reviews under negative polarity reviews with respect to cognitive processes specified that 70% of the deceptive reviews are written with higher cognizance based words. This clearly specifies that the truthful reviews are written without any deliberation.

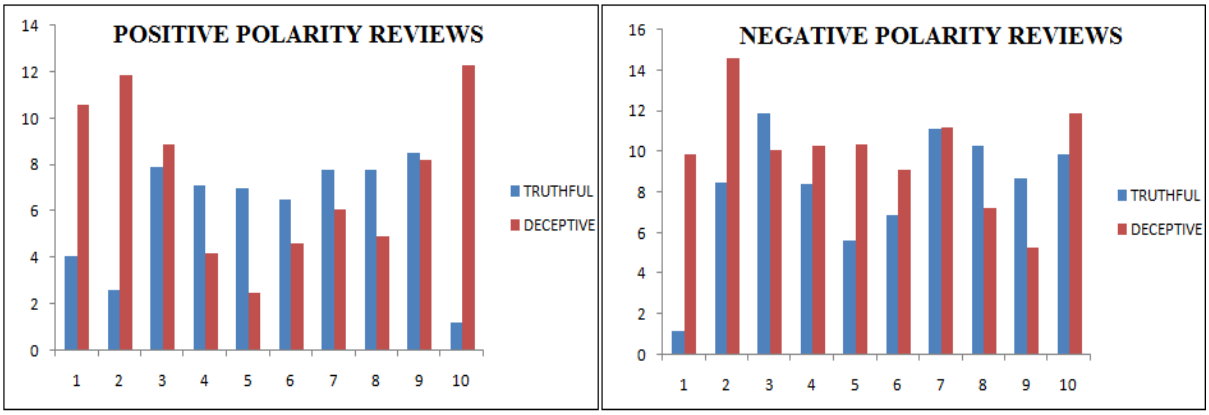

Figure 5. Cognitive process scores comparsion between positive polarity reviews and negative polarity reviews

4.4.3 Clout:

Clout taps writing that is authoritative, confident, and exhibits leadership. The analysis on the truthful reviews and deceptive reviews under positive polarity reviews with respect to authoritative words specified that 60% of the truthful reviews are written with average number of leadership words. The analysis on the truthful reviews and deceptive reviews under negative polarity reviews with respect to authoritative words specified that 60% of the truthful reviews are written with average number of leadership words. This clearly specifies that the truthful reviews are written concentrating on the experience rather on the authoritative words.

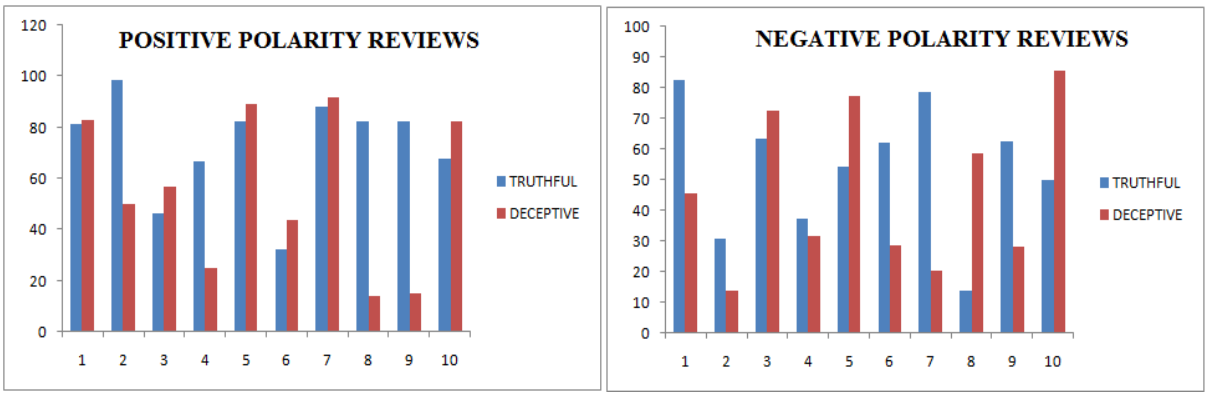

Figure 6. Clout scores comparsion between positive polarity reviews and negative polarity reviews

4.4.4 Authenticity:

Authenticity refers to writing that is personal and honest. The analysis on the truthful reviews and deceptive reviews under positive polarity reviews with respect to honest words specified that 70% of the truthful reviews are written with acceptable number of genuine words. The analysis on the truthful reviews and deceptive reviews under negative polarity reviews with respect to honest words specified that 80% of the deceptive reviews are written with high number of genuine words in order to defame the competitor. This clearly specifies that the truthful reviews are written honestly.

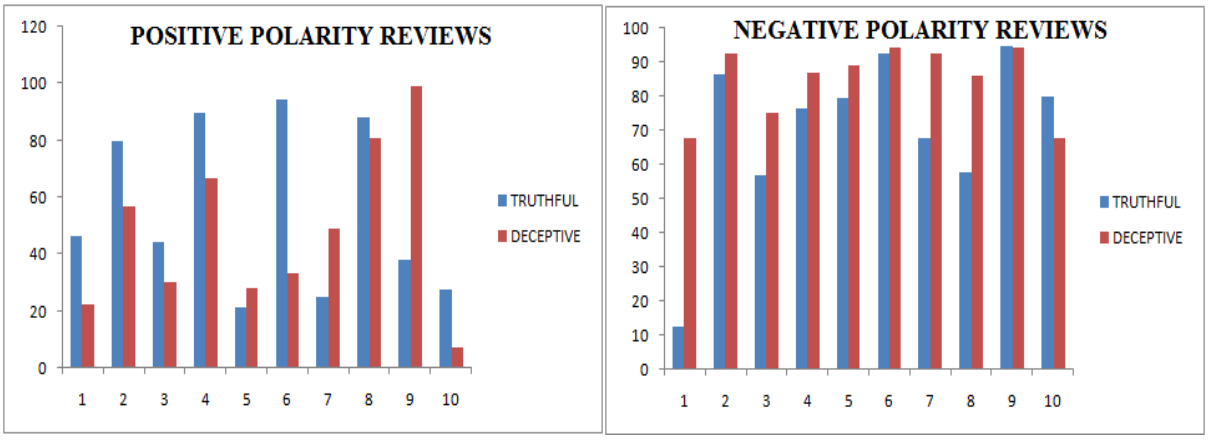

Figure 7. Authenticity scores comparsion between positive polarity reviews and negative polarity reviews

4.5 DECISION TREE CLASSIFIER FOR IDENTIFYING TRUTHFUL AND DECEPTIVE REVIEWS BASED ON LIWC VALUES

The four LIWC variables with the review deviation are treated as deep linguistic features on the considered reviews and are used to train decision tree classifier. This is a well known machine learning model. The trained decision tree classifier contains class labels as the leaf nodes of the tree. The probability of correct classification to a class c_p is;

$$P_c(c_p) = \prod_{i_p=1}^{n_p-1} P_c(c_p \mid N_{ip}) \ldots (2)$$

where $(N_1,\ldots,N_{np})$ is the tree branch ended with the leaf node which is assigned with c_p label. $P_c(c_p \mid N_i)$ is the probability of correct decision at node N_i. Subsequently, the probability of correct classification with the entire tree is;

$$P_c(T) = \sum_{c_P \in C} P(c_p) \prod_{i_p=1}^{n_p-1} P_c(c_p \mid N_{ip}) \ \ldots (3)$$

The class labels considered for this work are truthful and deceptive respectively. The learned model in the form of decision tree from the constructed dataset with deviation and four LIWC variables is tabulated in Table 1.

TABLE 1. Decision Tree model learning

Machine Learning inputs	Values generated
Classifier learned	Limited Search induction tree algorithm
Splitting Attribute	Clout
True positives and False positives percentage	90%

It is observed from the analysis that the review deviation feature did not appear in the learned decision tree. The decision tree algorithm treated the review deviation feature as irrelevant in the model learning process.

5. EXPERIMENTAL RESULTS AND DISCUSSION

As discussed earlier, the data corpus used for spam reviews classification is the gold standard opinion spam reviews collection. The analysis is carried out on 400 truthful positive reviews as collected from TripAdvisor and 400 deceptive positive reviews collected from Amazon Mechanical Turk. A total of 800 reviews were collected. The data corpus also contains 400 truthful negative reviews collected from various websites namely Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor and Yelp and 400 deceptive negative reviews from Amazon Mechanical Turk. The average score of four LIWC variables on both positive polarity reviews and negative polarity reviews is given below in the table.

TABLE 2. Average scores of four LIWC variables

LIWC Variable	Positive Polarity reviews		Negative polarity reviews	
Name	Truthful reviews	Deceptive reviews	Truthful reviews	Deceptive reviews
Social words	7.41	7.02	7.28	7.63
Cognitive words	6.05	7.42	8.25	10
Clout	72.89	55.17	53.55	46.32
Authenticity	55.52	47.42	70.56	84.79

From the above table scores it is observed that the deliberate writings of the turkers are better understood with deceptive reviews when compared with the truthful reviews for both positive and negative polarity reviews.

The gold standard reviews data corpus is pre-processed by removing stop words and non English words. Then, the sense of the review is determined. The deviation of each review is calculated and a comparison is performed between truthful reviews deviation and deceptive reviews deviation. It is found that there is no significant difference between the two deviations. So, the help of LIWC dictionary is taken in order to detect and classify the spam reviews. Four LIWC variables namely the social words score of the review, the cognitive process score of the review, the clout score of the review and the authenticity score of the review. These variables with their corresponding scores and the corresponding review deviation are used in learning the decision tree classifier.

The accuracy of the learned decision tree is best interpreted by Area under Receiver Operating Characteristic (AUC) curve. The accuracy of model AUC is 95.5%. The AUC plot is presented below.

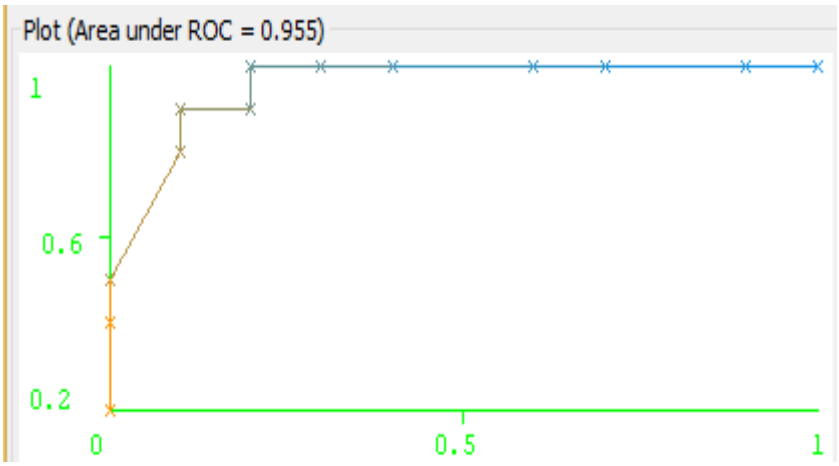

Figure 8. Decision Tree AUC Curve

The learned decision tree is evaluated on the 20% of the test set. It is observed that 90% of the truthful reviews from the training data are correctly classified as truthful reviews by the decision tree model. Also it is observed that 90% of the deceptive reviews from the training data are correctly classified as deceptive reviews by the decision tree model.

The comparison of accuracy of the learned decision tree model using LIWC variables as deep linguistic features in terms of AUC is carried out against the Ott et al. [5] learned SVM model using LIWC based 15 types of words as psycholinguistic features. They obtained the accuracy of 76.8%. The proposed approach outperformed the Ott et al. approach. This is because Ott et al. used 15 LIWC features in which the emotional words were also used as feature. The emotional words tend to detect more positive sentiment related words rather than psycholinguistic related words written in the reviews. Also, Ott et al. used LIWC 2007 version

dictionary in which clout and authenticity variables were not present. The current work uses LIWC 2015 variables which includes clout and authenticity. The classifier performance parameters with LIWC versions comparison is provided in the below table.

TABLE 3. Classifier performance parameters and LIWC versions comparison

Classifier used	ROC%	LIWC version used	Summary variables from LIWC
SVM [5]	76.8%	2007	No
Decision Tree (our work)	95.5%	2015	Yes

6. CONCLUSION AND FUTURE WORK

The classification of spam reviews based on four LIWC variables as deep linguistic features in decision tree model was carried out successfully. In the process of this work it is found that one of the LIWC summary variables named Clout was the deciding feature in the growth of decision tree. The very high ROC accuracy with the four variables is an encouraging factor towards using them in the advancing psychometric research works.

In future, the truthful reviews as classified by the decision tree are used to identify various aspects and corresponding opinions from the reviews. Also, the considered four deep linguistic features with extracted aspects and opinions help in estimating the helpfulness of the review in a better manner.

REFERENCES

[1] Jindal, N., & Liu, B. (2008, February). Opinion spam and analysis. In Proceedings of the 2008 International Conference on Web Search and Data Mining (pp. 219-230). ACM.
[2] Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In LREC (Vol. 10, pp. 2200-2204).
[3] Pennebaker, J. W., Boyd, R. L., Jordan, K., & Blackburn, K. (2015). The development and psychometric properties of LIWC2015.
[4] Wu, G., Greene, D., Smyth, B., & Cunningham, P. (2010, July). Distortion as a validation criterion in the identification of suspicious reviews. In Proceedings of the First Workshop on Social Media Analytics (pp. 10-13). ACM.
[5] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011, June). Finding deceptive opinion spam by any stretch of the imagination. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies- Volume 1 (pp. 309-319). Association for Computational Linguistics.
[6] Chen, C., Wu, K., Srinivasan, V., & Zhang, X. (2013, August). Battling the internet water army: Detection of hidden paid posters. In Advances in Social Networks Analysis and Mining (ASONAM), 2013 IEEE/ACM International Conference on (pp. 116-120). IEEE.
[7] Lim, E. P., Nguyen, V. A., Jindal, N., Liu, B., & Lauw, H. W. (2010, October). Detecting product review spammers using rating behaviors. In Proceedings of the 19th ACM international conference on Information and knowledge management (pp. 939-948). ACM.
[8] Mukherjee, A., Liu, B., Wang, J., Glance, N., & Jindal, N. (2011, March). Detecting group review spam. In Proceedings of the 20th international conference companion on World wide web (pp. 93-94). ACM.
[9] Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. Journal of Big Data, 2(1), 23.
[10] Pedersen, T., Patwardhan, S., & Michelizzi, J. (2004, May). WordNet:: Similarity: measuring the relatedness of concepts. In Demonstration papers at HLT-NAACL 2004 (pp. 38-41). Association for Computational Linguistics.

AUTHOR PROFILE

Pandi. Chiranjeevi is presently working as Assistant Professor in CSE at ACE Engineering College, Hyderabad. He is a research scholar in the faculty of Computer Science and Engineering, JNTUH, Hyderabad. His areas of interest are data mining, machine learning, natural language processing. He is one among the research scholars of B. Vishnu Vardhan.

D. Teja Santosh is presently working as Assistant Professor in CSE at GITAM (Deemed to be University) Hyderabad. He holds a doctoral degree in the faculty of Computer Science and Engineering from JNTUK, Kakinada. His areas of interest are data mining, machine learning, natural language processing. He was the Additional Reviewer for 2015 International Conference on Advances in Computing, Communications and Informatics (ICACCI) in Kerala, India. He is the member of professional bodies of international repute namely IEEE, ACM, IAENG.

B. Vishnu Vardhan is presently serving in the capacity of Vice-Princpal at JNTUH College of Engineering Manthani, Peddapally. He is also the Professor in the faculty of Computer Science and Engineering. He was an active member in Free Software Movement in India. His areas of interest are linguistic processing, data mining, natural language processing, information security and other elite fields of engineering. He has completed Government of Andhra Pradesh funded project worth 5 lakhs from Ministry of IT on localisation activity. As a co-investigator he completed another UGC funded project worth 9 lakhs. He is a technical member for International Journal of Information and Electronics Engineering.