Fake Review Detection

Sam Clastine Jesumuthu
Department of Computer Science,
City, University of London
London, United Kingdom
sam.jesumuthu@city.ac.uk

***Abstract*—Reviews now contributes significantly in buying and selling a product. At present people buy product based on how good the review is. So, promoting the product and to increase the business revenue, this individual getting their focus towards optimizing the reviews as good as possible. Unfortunately, manipulating consumer purchasing Decision.**

In this paper, we highly focused in review-based features which helps in detecting pattern of deceiver. In particular we have analyzed each feature how they vary from truthful. As this proposed research is based on supervised Learning so the algorithm use for the study is logistics Regression and Random Forest classifier. Furthermore, by considering several factors to extract features. And scaling features that will improve the prediction to classify the deceivers.

***Index Terms*— Reviews, product, deceptive reviews, consumer, detecting pattern, truthful, supervised Learning, algorithm, logistics Regression, Random Forest classifier, deceivers.**

I. Introduction

The problem of fake reviews poses a serious and persistent challenge for websites that host customer reviews, as well as for customers who trust these reviews. According to study of squareup.com shows that around 3.8 million dollars in global spend on ecommerce in 2021 was driven by online reviews [10]. To be specific 52 percent of reviews posted on Walmart and 30 percent in Amazon are unreliable [2]. Sometime the fake negative review will have greater impact in selling the product and business reputation on the away to sink. Additionally, there is no oversight of the reliability of the reviews produced on e-commerce platforms, which encourages the development of numerous low-quality evaluations. To increase the purchase of their online goods and services, several businesses pay people to publish false reviews [15]. This fraudulent review causes both sellers and consumer. Extensive study has been done on how to classify deceptive reviews. The most common method to classify deceptive reviews review based and reviewer based. This paper is completely focused on review based it detects the misleading review based on the similarities and linguistics features of the reviews. It pulls out relevant structural features, POS tags and Semantic features. The main goal is to analyse the features which recognizes the deceptive review and then use machine learning algorithm to classify the reviews.

II. Analytical Question and Data

In this paper, we address deceptive review detection based on review-centric approach through the following research question:

1. *How does POS tag properties associates to deceptive and Truthful review?* Detecting patterns in deceptive speech is challenging but filtering linguistic features from review may show how comparatively it differ from truthful. Hancock et.al. suggest that overall, when senders were lying to their partners, they produced more words, used more "other" pronouns (e.g., "he," "she," "they"), and used more terms that described the senses (e.g., "see," "hear," "feel") than when they were telling the truth [5].
2. *How does the deceptive review's sentiment polarity for product categories varies in comparison to Truthful reviews?* The polarity for each product category of reviews will be negative and positive, Deceiver may write most positive or negative reviews based on the product.
3. *Will reviewer use offensive words during deceptive interaction while degrading the product?* Offensive words or swear word can be used to degrade a brand or product. These words may be increasing the potential to detect deceiver.
4. *Is there deceiver in verified purchase in compared to truthful?* Highly motivated deceivers may most likely try to make is review read and believe by everybody and people in online accepts the reviews which is by verified purchase.
5. *How useful are the characteristics obtained from various POS tags, sentiment polarity and readability score of review? Does it improve machine learning algorithm performance to detect deceptive reviews?* POS tags, sentiment polarity and readability score may have greater impact on the finding fake reviews and to distinguish that how each of them varies from other based on accuracy score of the algorithm.

III. Data(material)

To research and study about how deceptive review varies from truthful one it is important to have a labeled dataset. The Amazon review dataset which is labeled by Alsubari et.al. [1] with an outstanding accuracy of 95%, for labeling this dataset is the Amazon filtering algorithm that is employed by the Amazon website [1].

The Dataset contains 21,000 deceptive and truthful reviews of 30 product category with equally sized. Each of the purchased product has its product id, product title, verified purchase, review title, review text, label and rating. The reviews' average rating value was 4.13, and 55.7% of the data were identified as verified purchases [1].

IV. Analysis

A. Data Preprocessing

Raw dataset is always difficult to understand for our analysis it is important to transform data, so that we can visualize and analyze it easily. To get useful insights from the dataset, a traditional way preprocessing technique need to be applied in the dataframe. The preprocessing process will be implemented are:

- Renaming the target variable (label1= Deceptive, label2 = Truthful)
- Replacing short form words to a proper phrase.
- Removing all URLs (https//:) and punctuations from reviews
- Translating each text to lower case
- Stopwords Removal: Stopwords are available in abundance in any human language. By removing these words, we remove the low-level information from our text in order to give more focus to the important information. Removal of stop words definitely reduces the dataset size and thus reduces the training time due to the fewer number of tokens involved in the training [12].
- Lemmatization: Lemmatization usually involves using a vocabulary and morphological analysis of words, removing inflectional endings, and returning the dictionary form of a word (the lemma) [13]. It has the greater efficiency than stemming which chops of the words.
- Word Tokenization: It split raw text into chunks of words. This word helps the model to understand the context of the sentence.
- Vectorization: After processing and filtering all the data, now comes the main part that's converting our textual data to numeric values also known as word embeddings. In this study we are using TF-IDF a single float value per word that solves a very particular problem that may come in handy in text classification a lot i.e., word importance [14].

B. Feature Engineering

Feature extraction from review text is vital section of the study. As our study is based on review centric, we need to extract some linguistic features, Semantic features and other structural features.

- linguistic features: Some of the properties are POS tags which are both in percentages and counts. The linguistic features which are used in this study are Negation, offensive words and part of speech tags like Noun, Pronoun, Adjectives, Adverb, Conjunction and Preposition.
- Semantic features: To extract sentiment polarity, we had used VANDER (Valence Aware Dictionary for Sentiment Reasoning) is a model used to apply directly on unlabeled text data available in NLTK package.
- Structural features: The features used are Word count, Sentence Length, Capital Letters Count and Numeric count
- Readability Score: It quantitatively says how ease the text is readable. Some of the readability test used for study are: Flesch Reading Ease, Gunning Fog Index and dale-chall readability score

The above feature extraction will represent review-centric analysis on deceptive reviews and it has about 16 extracted features.

C. Encoding

There are many methods for encoding the categorical variable. The methods taken for consideration are One Hot Encoding and Label Encoding. To Identify proper method that's suitable for model for the data is Label Encoding, the reason for not using is it produces multiple variables causing the size of the data which may affect the model accuracy.

D. Feature Analysis

To determine the patterns and the distribution for the paired variable shows the relationship. While the we have pair grid(figure 1) with two plots one is scatter and other one is Kernel Density Estimates which is diagonally plotted.
Kernel Density estimated offers greater flexibility to understand curve of the distribution. It is calculated by weighing the distance of all data points in each specific location along the specific location along the distribution. This multimodal distribution helps us to understand the variable in a better way compare to other distribution function. Scatter Plot shows the relationship between the two variables i.e., Positive or Negative.
In this pair grid the we had used linguistics and readability features to visualize distribution and correlation. In Scatter plot, most of the POS tag variables show positive relationship between them but pronouns person 1 and person 2 variable has less relationship with other POS tag features. While other features like structural and readability are not at all corelated with post tag features.

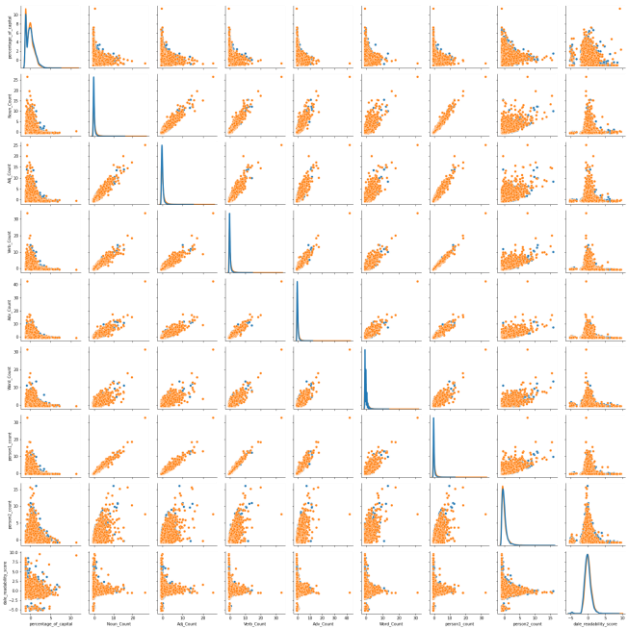

Figure 1

In KDE plots, all the features are right skewed leaving dale readability score, so as this says that median will always higher that mean. And we had plots with multimodal distribution that is percentage of capital variable has two peaks this says that a set of data which has more than one mode. This pair grid plot visualized each and every in better and efficient way and helped to find that pos tags are positive correlation between them, it also demonstrated the multimodal distribution of the variable using KDE and also most of variables has outliers.

E. Feature selection

It is important to find the most influencing features to target variable (deceptive or truthful). In this study, the algorithm which used to select the best features is Extra Trees Classifier an ensemble learning technique, and it make decision based on Gini Index. The feature which is given as input are all linguistic, semantic, structural and readability variables.
In Figure 2 it has illustrated the top 10 features which will be used in modelling.

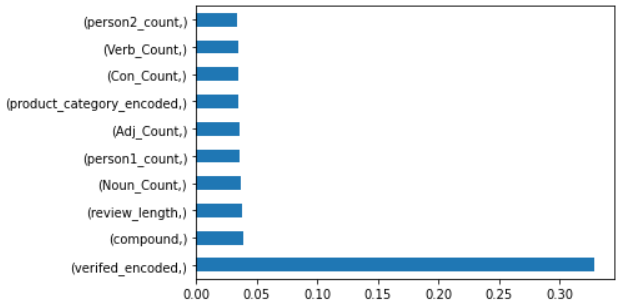

Figure 2

F. Modeling and Algorithm

In this paper, the algorithm used is Logistic Regression and Random Forest Classifier. Logistic Regression is easier to train at less time and the most efficient algorithm another main reason is most of the features (POS tags) has linear relationship. Random Forest Classifier is ensemble learning method one of the main advantages of this algorithm is outperform in accuracy of predicting the outcome.
To find the efficiency of how each feature extracted from the algorithm helps to detecting the deceptive reviews. So we created data are only reviews, reviews and readability Score, Reviews and POS tags, reviews and structural features and reviews and variable from feature selection (for feature selection we had used all features semantic, structural and linguistic)
Thus comparing the model performance for each data would give overview of how efficient is our data performance in predicting the label.

V. FINDINGS

1. *How does POS tag properties associates to deceptive and Truthful review? Does POS tag identify deceptive reviews?*

The data shows in Figure 2 there is quite difference between the deceptive and truthful reviews. Taking truthful review in consideration, deceptive reviews have less counts when compared to truth expect preposition count for both are almost the same. To conclude that, deceiver may had tried to create a review with vague details, and as there is less variation in deceptive text, we can't derive that POS tag identifies the deceiver.

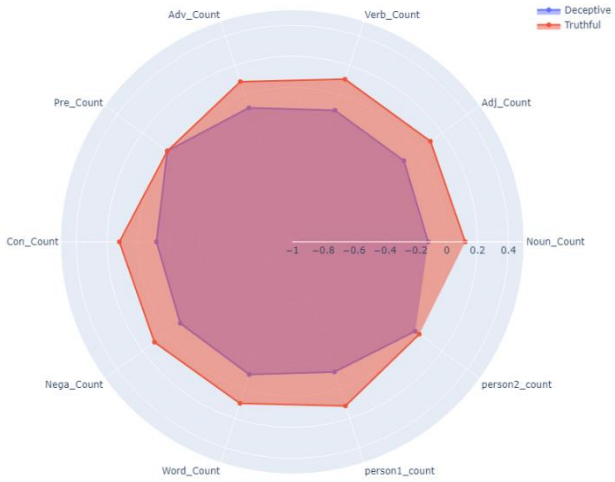

Figure 3

2. *How does the deceptive review's sentiment polarity for product categories varies in comparison to Truthful reviews?*
Categorizing positive and negative sentiment to the deceptive and respectively for the truthful as well. Sampling 10 products out of 30. The plot (Figure 4) illustrates that there is high number of positive reviews to deceptive, on the other side fake review has slight low negative sentiment. In products, furniture and Books have most negative reviews when compare to other. And most of them are almost evenly distributed. As a result, deceiver creates large number of positive sentiment review than negative.

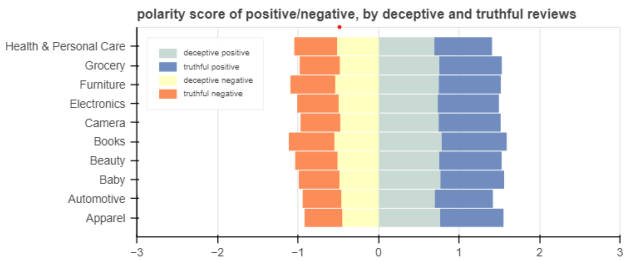

Figure 4

3. *Will reviewer use offensive words during deceptive interaction while degrading the product?*

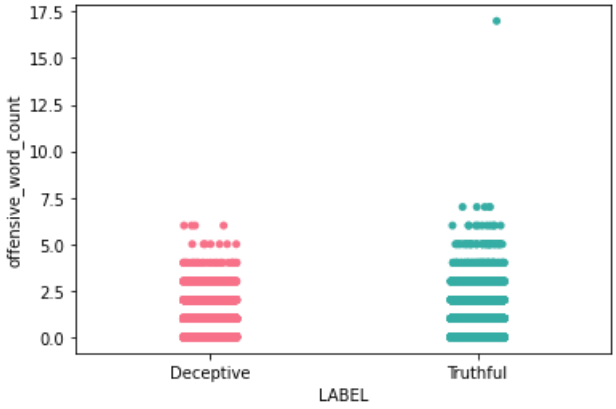

Figure 5

Offensive word is one of the important features extracted from review text. The graph (Figure 5) shoes most of the truthful reviewers used bad words. The highest count is around 7 in deceptive review and 17 in truthful.

4. *Is there deceiver in verified purchases in compared to truthful*

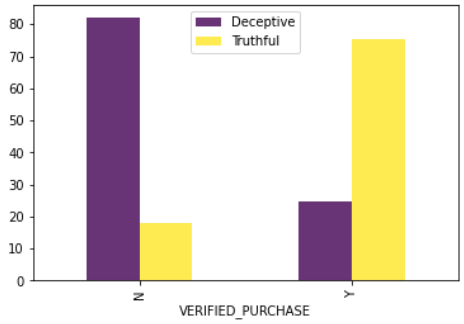

Figure 6

The figure 6 demonstrates, that in verified purchases there are very few less deceivers and most of them are not verified which is around 82% (Table 1).

VERIFIED PURCHASE	Deceptive	Truthful
N	81.950118	18.04988
Y	24.593948	75.406052

Table 1

And 75% in Truthful are verified, so there is large variation between deceptive and truthful in verified purchases.

5. *How useful are the characteristics obtained from various POS tags, sentiment polarity and readability score of review?*

Thus, the Table 2 shows how each data with different show the contrast performance between two algorithms i.e., Random Forest and Logistics Regression but both this algorithm performed well in variable selected from feature selection

Algorithm	Data	Train		Test	
		Accuracy	MSE	Accuracy	MSE
Logistic Regression	Reviews	0.67	0.326	0.56	0.439
	Reviews + Readability Score	0.68	0.318	0.57	0.421
	Reviews + POS tags	0.68	0.312	0.58	0.413
	Reviews + Structural Features	0.68	0.318	0.57	0.423
	Reviews + Variable from Feature Selection	0.81	0.186	0.79	0.203
Random Forest	Reviews	0.89	0.105	0.56	0.439
	Reviews + Readability Score	0.99	0.004	0.56	0.435
	Reviews + POS tags	1.0	0	0.60	0.392
	Reviews + Structural Features	0.98	0.012	0.58	0.415
	Reviews + Variable from Feature Selection	1.0	0	0.80	0.198

Table 2

VI. Reflections, Further Work

In this investigation of detecting fake reviews utilizing linguistics, structural, readability and semantic features. And demonstrates the solution to questions, some provide most satisfied indication and others vague. But after this research it has potential to solve this problem.

Most of real time data review data are unlabelled, research was made labelling the data with fake review corpus using active learning method and vectorizing text with ELMO a pretrained model would have extensive focus on context of each sentence and Transformer based model increases the focus on detecting the patterns.

Further comparing both models would help to get a ground truth result on detection.

VII. WORDCOUNT

Section	Expected	Actual
Abstract	150	141
Introduction	300	220
Analytical Question and Data	300	275
Data	300	107
Analysis	1000	989
Finding, reflections and further work	600	569
Total	2650	2598

REFERENCES

[1] Alsubari, S.N., Deshmukh, S.N., Al-Adhaileh, M.H., Alsaade, F.W. and Aldhyani, T.H.H. (2021). Development of Integrated Neural Network Model for Identification of Fake Reviews in E-Commerce Using Multidomain Datasets. *Applied Bionics and Biomechanics*, 2021, pp.1–11. doi:10.1155/2021/5522574.

[2] https://www.cbsnews.com/news/buyer-beware-a-scourge-of-fake-online-reviews-is-hitting-amazon-walmart-and-other-major-retailers/

[3] Burgoon, Judee, et al. Detecting Deception through Linguistic Analysis Active Shooter Events View Project Cognitive Bias Mitigation View Project Detecting Deception through Linguistic Analysis. 2003. Accessed 12 Dec. 2022.

[4] Feng, Song, et al. *Syntactic Stylometry for Deception Detection*. Association for Computational Linguistics, 2012.

[5] Hancock, J.T., et al. "Automated Linguistic Analysis of Deceptive and Truthful Synchronous Computer-Mediated Communication." *IEEE Xplore*, 1 Jan. 2005. Accessed 12 Dec. 2022.

[6] Jindal, Nitin, and Bing Liu. *Review Spam Detection*.

[7] Ott, Myle, et al. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. 2011.

[8] Yoo, Kyung-Hyan, and Ulrike Gretzel. "Comparison of Deceptive and Truthful Travel Reviews." *Information and Communication Technologies in Tourism 2009*, 2009, pp. 37–47. Accessed 12 Dec. 2022.

[9] Zhou, Lina, et al. An Exploratory Study into Deception Detection in Text-Based Computer- Mediated Communication Sea-Based Battle Lab View Project Interpersonal Deception View Project an Exploratory Study into Deception Detection in Text-Based Computer-Mediated Communication *. 2014.

[10] https://squareup.com/gb/en/townsquare/tackling-fake-business-reviews

[11] Fontanarava, Julien, et al. "Feature Analysis for Fake Review Detection through Supervised Classification." *IEEE Xplore*, 1 Oct. 2017, ieeexplore.ieee.org/document/8259828/.

[12] https://towardsdatascience.com/text-pre-processing-stop-words-removal-using-different-libraries-f20bac19929a

[13] https://www.engati.com/glossary/lemmatization

[14] https://medium.com/data-science-in-your-pocket/text-vectorization-algorithms-in-nlp-109d728b2b63

[15] Wang, Zehui, et al. Fake Review Detection on Yelp.