UNFAIR REVIEWS ON DIFFERENT PRODUCTS USING SENTIMENT ANALYSIS

Gundam Nikita
ECE Department
Amrita School of Engineering
Kollam, India
nikitagundam@gmail.com

Kalla Vishwas Uday Kiran
ECE Department
Amrita School of Engineering
Kollam, India
kallavishwasudaykiran@gmail.com

Kuruba Aparna
ECE Department
Amrita School of Engineering
Kollam, India
aparnakurba07@gmail.com

Singamsetty ChanduPriya
ECE Department
Amrita School of Engineering
Kollam, India
chandupriya.2802@gmail.com

***Abstract*— Sentiment analysis is the examination of feelings and viewpoints in any kind of literature. The expression of the opinions of the masses, the group, or any person is facilitated greatly by sentiment analysis of the data. To evaluate and gain insights from text data from numerous sources, like Facebook, Twitter, and Amazon, sentiment analysis is one of the most prominent fields. It is essential for enabling firms to actively work on enhancing their company strategies and gaining a thorough understanding of the customer feedback on their products. It comprises computational analysis of a person's purchasing interests and subsequent opinion mining about a company's corporate entity. To accomplish this, we must first pre-process the dataset, then extract the adjectives from the dataset that have some meaning, which is known as a feature vector, then select the feature vector list, and finally apply machine learning based classification algorithms such as Naive Bayes, Maximum entropy, and SVM, as well as the Semantic Orientation based WordNet, which extracts synonyms and similarity for the content feature.**

I. INTRODUCTION

Reviews may contain user reviews on products, destined to help other users in their buying decision making. Huge numbers of reviews exist, which makes it difficult for a consumer to read them all and make a decision. With increasing volume of reviews, customers sometimes make product buying decisions based on unfair or fake feedback reviews. The NB, DT-J48, LR, and SVM, KNN algorithms are used in this study's sentiment classification to separate the datasets of reviews into unfair and fair reviews. The author noted that utilizing the stop word removal method is more efficient for detecting unfair evaluations than without using stop words. This growth of E-commerce gave importance to customer needs and opinions which in turn gave rise to an important aspect of online shopping known as 'User Reviews'. User reviews are customer suggestions and opinions about the product which helps other customers make decisions about that product. Such review systems form the backbone of E-commerce. In this research we are looking forward to understanding and analyzing the dataset with the help of different visualization techniques.

II. LITERATURE REVIEW

A. Paper 1 - A Survey on Fake Review Detection using Machine Learning Techniques[1]

This paper discusses different fake reviews detection techniques that are based on unsupervised, supervised as well as semi supervised methodologies. It has different features in detail like linguistic features, behavioral and relational features. Also compared different techniques to identify fake reviews. Major challenges are discussed on fake review detection. When fake reviews are intentionally fabricated like genuine review, it would be hard to decide genuine review. When only one review is available for a particular item, it is difficult to identify rating behaviors. Due to rapid development of the internet, the size of the reviews of the items / products increases. These huge amounts of information are generated on the Internet. There is no analysis of quality of reviews that are written by consumers. Anyone can write anything which conclusively leads to fake reviews or some companies are hiring people to post reviews. Some of the fake reviews that have been intentionally fabricated to seem genuine, capability to identify that fake online reviews are crucial.

B. Paper 2 - Unfair reviews detection on Amazon reviews using sentiment analysis with supervised learning techniques [2]

In this paper, the author analyzes the sentiment of three different Amazon reviews datasets using four supervised machine learning techniques. These techniques are also used to identify unfair positive and negative reviews. Our study's primary objective is to apply Sentiment Analysis algorithms

and supervised learning approaches to datasets classify Amazon reviews into fair and unfair reviews. A supervised learning method with three classes negative, neutral, and positive can be used to define detection procedures for sentiment classification based on machine learning. Supervised learning algorithms include two stages: extraction and, in particular, reviews' selection. Introduces a feature selection technique for the SA classification task and evaluates it against stop word removal to determine the most accurate supervised learning algorithm. Four sentiment classification algorithms have been used in the research results to demonstrate their performance and accuracy in identifying unfair reviews.

C. *Paper 3 - A Study on Identification of Important Features for Efficient Detection of Fake Reviews [3]*

This paper utilizes the process of readability features and topic features to prove them as more effective than sentiment analysis for fake review detection. They used Readability features such as Gunning-Fog Index, Flesch-Kincaid Reading Ease Index, Automated Readability Index and Coleman-Liau Index. And they used topic features such as Average Topic Probabilistic, Big Topic Probabilistic and Review Topic Distribution. Finally, they classified the fake reviews individually by different combinations of the above features by using Support Vector Machine (SVM), k-Nearest Neighbor (k-NN) and Logistic Regression (LR). According to the classification results of SVM, readability features (FOG + FK + ARI + CLI) are the most effective feature set for fake review detection. The readability feature FOG with topic features (ATP + BTP + RTD) produces the second-best performance for fake review detection with SVM. The performance of other classifiers is found to be maximum when all the features are used.

D. *Paper 4 - Opinion spam detection framework using hybrid clasification scheme [4]*

This paper proposes development of a unified framework for Opinion Spam detection utilizing a hybrid categorization approach is the proposed work's primary contribution. The identification of spam has generally been studied in a variety of areas, including video spam, search engine spam, and email spam. Opinion spam may be seen in product review sites that aim to promote or harm the reputation of competitor brand products by giving unjustified favorable or negative ratings to some products in order to deceive users. According to reports, 15% of reviews are simply copies of prior evaluations, which might be considered opinion spam. For the purpose of identifying Opinion Spam, the author distinguished between fake reviews, brand reviews, and non-reviews. Even expert readers may effortlessly identify brand reviews and non-reviews when reading using a supervised learning logistic regression model. The fake reviews are challenging to spot. Thus, they are detected by labelling duplicate spam review.

E. *Paper 5 - Sentiment Analysis on Consumer Reviews of Amazon Products [5]*

This paper understands and analyzes the Amazon user review dataset with the help of different visualization techniques like Common words used in reviews i.e., word cloud and Number of helpful reviews by user. Also, from the dataset it is observed that most of the products have either 4-star or 5-star ratings. The statistical analysis determines the relationship between the price of the product and the number of helpful reviews. Also, products with high prices have less rating than the products at low prices i.e., most of the consumers have more review on low price products. So, we can say that product price is inversely proportional to number of helpful reviews indicating a negative correlation between them. Also done a chi-square test to determine the p-value for items in various categories which discovered that the observed values and our expected values disagree, and the p-value is likewise quite small. As a result, the analysis is largely biased towards higher rating and null hypothesis in almost all sub-categories it miserably failed, and rather chose alternative hypothesis. The number of reviews on products gradually increased from the year 2014 to 2017.

III. Problem Statement

One of the major challenges confronting Sentiment Analysis today is how to detect unfair negative reviews, unfair neutral reviews and unfair positive reviews from opinion reviews.

Dishonest reviews or ratings have already become a serious problem in practice. Thus, in this research, our primary goal is detecting unfair reviews through Sentiment Analysis using supervised learning techniques.

IV. Block Diagram

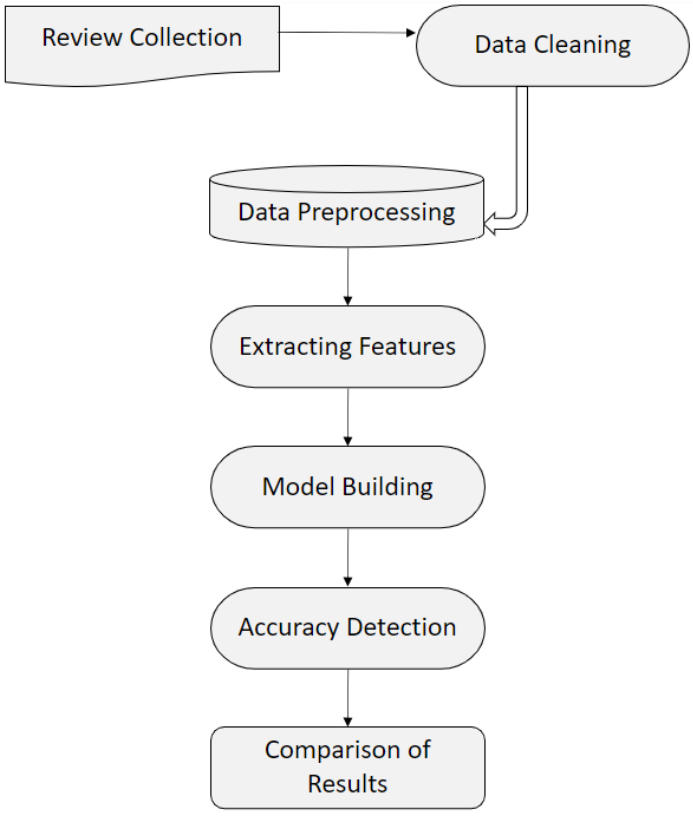

Fig. 1. Block Diagram of our Project

V. Initital implementation and report

A. Reviews Collection

Our experiment is based on a machine learning algorithm analysis of the sentiment value of the standard dataset. To evaluate our review categorization algorithms, we used the dataset of Amazon reviews and analyzed the datasets.

B. Data Cleaning

The dataset used in our experiment was broken down into five rating scales: 1-star, 2-stars, 3-stars, 4-stars, and 5-stars. It was derived from Amazon product data. The original dataset is typically not very clean and is difficult to model. Some blank rows that were confusing the analytical process have been eliminated.

C. Data Preprocessing

Data preprocessing is a significant step and plays crucial role to many supervised learning approaches.

StringToWordVector

Depending on whether a word appears in the document or not, the StringToWordVector filter assigns each single word in the transformed datasets an attribute value that is either Positive, Negative, or Neutral. Stopwords Removal and Tokenization are two of the subprocesses that use this filtration procedure.

Stopword Removal and Tokenization

Stop words are frequent words that need to be filtered out before the classifier is trained. Some of those words are frequent, but they don't add anything to our labelling scheme or improve the meaning of a sentence; instead, they cause difficulty for our classifier.

Attribute Selection

The process of choosing a subset of relevant features to be used in building a model is known as attribute selection in machine learning, sometimes known as feature selection. Selection of the right attributes can make it better and improve classification accuracy.

VI. Results

The results of the experiments will demonstrate their accuracy and performance in detecting unfair reviews using sentiment classification algorithms.

Acknowledgment

We would like to convey our deep gratitude to our mentor Remya M S for the valuable and constructive suggestions, enthusiastic encouragement and useful tips of our literature review and Mini Project. Besides our mentor, we would also like to thank Rahul Lal for guiding us and helping us to know the instructions and also constantly giving us tips on how to present the literature review.

References

[1] *N. A. Patel and R. Patel, "A Survey on Fake Review Detection using Machine Learning Techniques," 2018 4th International Conference on Computing Communication and Automation (ICCCA), 2018.*

[2] *Elshrif Elmurngi and Abdelouahed Gherbi, "Unfair reviews detection on Amazon reviews using sentiment analysis with supervised learning techniques" Article in Journal of Computer Science, 2018.*

[3] *Ting-You Lin, Basabi Chakraborty, Chun-Cheng Peng, "A Study on Identification of Important Features for Efficient Detection of Fake Reviews" Article in International Conference on Data Analytics for Business and Industry (ICDABI), 2021.*

[4] *Asghar, M.Z., Ullah, A., Ahmad, S. et al. "Opinion spam detection framework using hybrid classification scheme", Soft Comput 24, 3475–3498 (2020)*

[5] *Aamir Rashid and Ching-yu Huang. "Sentiment Analysis on Consumer Reviews of Amazon Products". Article in International Journal of Computer Theory and Engineering, Vol. 13, No. 2, May 2021*