

# **Fake Product Review Monitoring and Analysis using Opinion Mining**

B.Saikoti reddy	B.Mahith	CH.Venkata siva sai
Student of CSE	Student of CSE	Student of CSE
Dr.M.G.R Educational And	Dr.M.G.R Educational And	Dr.M.G.R Educational And
Research Institute	Research Institute	Research Institute
Chennai, India	Chennai, India	Chennai, India
bhavanamsaikotireddy@gmail.co	mahith.bommireddy2002@gmail.com	chirumamillavenkatasivasai@gmail.co
<u>m</u>		<u>m</u>

Mrs.Maheswari.A Professor, Dept of CSE Dr.M.G.R Educational And Research Institute Chennai, India <u>Maheswari.it@drmgrdu.ac.in</u> Dr.P.Dhivya Professor, Dept of CSE Dr.M.G.REducational And Research Institute Chennai, India <u>dhivya.cse@drmgrdu.ac.in</u>

#### Abstract

Product reviews play a vital role in determining the sales of products on online shopping websites such as Flipkart, Amazon, and Snapdeal. This paper proposes a framework to detect spam or fake reviews using Opinion Mining, or Sentiment Analysis. Here, we examine customer views in text form. First, we determine if a review is for a particular product using a Decision Tree method. In order to identify spam, we use a Spam dictionary to identify words related to spam in the reviews. There are different algorithms used in Text Mining, and thus

we are driven to certain conclusions based on these calculations.

**Keywords**: text mining, decision trees, opinion mining, and spam review identification.

# I. INTRODUCTION

Product reviews in today's marketplace are critical in influencing consumers' purchasing decisions online. Most people go through these reviews prior to purchasing a product. Reviews may be negative or positive, where positive reviews tend to attract more customers than negative reviews. This means that product reviews have a great effect on a business, as they could help the business gain profit or incur a loss. Customers often comment on products they buy, but with real reviews come many fake reviews that can mislead buyers.

There are a lot of reviews on online shopping sites and apps such as Amazon and Flipkart, which are largely motivated by money. Although there are a lot of reviews by actual customers describing their experiences, there are many false reviews also available for the same purpose. For example, if a person leaves a bad review of an iPhone 8 on an Apple review website because of bad service, it can give a bad impression of the product to potential customers, which can ultimately damage sales. To offset the effect of such genuine reviews, Apple Inc. may go to the extent of employing people to leave good reviews to improve its reputation. This practice may deceive customers into buying inferior products, while quality products or truthful services are denied by



false negative reviews. Individuals who write such fake reviews with the intention to deceive consumers or opinion analysis programs are commonly called spammers, and their reviews are termed as spam reviews. To show how difficult it is to detect spam, let us take an example of a review written on oyo.com (Indian version). This review, as with all of the reviews done by the one selling the hotel, makes it challenging to separate what is true and what isn't: "If you don't mind small room size (usual in India), then you will fall in love with this hotel.". It's a ten- minute walk from the railway station, cleanliness is a top priority for the hotel, and they offer a great OYO breakfast with tea/coffee, along with outstanding service at a very cheap price. There is absolutely nothing to be unhappy about.

Other people's thoughts and opinions have always played a significant role in the decisionmaking process. Even prior to the widespread use of the internet, individuals would ask other people for their opinions about products or use surveys to assist

them in making a decision.

With the advent of the World Wide Web, consumer behavior has changed, and now more individuals are buying products online. With the increase in internet usage, there has also been an increase in the habit of people sharing their reviews, ratings, and comments. Sharing feedback has been found to be highly beneficial for other consumers who make their purchases based on online sites.

Consequently, the number of reviews has increased, and they play an important role in influencing purchasing decisions. Best- selling items receive hundreds of reviews that may enhance or decrease their sales, hence the importance of these views. Our application is designed to sift through the reviews, separating the real reviews from the spam, to enhance the precision of the existing system and drive the sales of genuine products based on users' actual experiences.

# **II. LITERATURE SURVEY**

To fulfill the research goals, the following methodology will be used (see Figure 2.1):



Fig. 2.1 Methodology to accomplish objectives of the research.

**Research and Survey**: It consists of a complete analysis of the latest research very much connected with the theoretical problem statement and research questions.

After the study of current literature in spam detection within non-Arabic opinion reviews, we will determine the inadequacies of current techniques and create ways to fill these inadequacies within our research. **Data Acquisition**: Here, we will generate an inhouse data set of spam reviewers and reviews, taken from online e-commerce websites such as Amazon and Flipkart.

The data will be acquired by humans and will differ in attributes and sizes, with records chosen randomly from the available data on these websites.

**Data Integration**: Here, we will combine data from various review sources into a single format.

**Spam Identification** Labelling: In this phase, we will inspect the merged dataset for various kinds of spam and label each record as spam or non-spam manually.

**Pre-processing**: This process includes applying pre-processing techniques to handle missing, noisy, and inconsistent data. Techniques may be case folding, removal of characters, tokenization, handling slang words, stop word removal, stemming, and number normalization.

- **Processing Stage**: Here, we will execute the following steps:

- 1. Data mining grouping.
- 2. Classification of text mining.
- 3. Classification of Data-Text mining.

We will utilize each of them through various classification techniques.

**Evaluate the Approach**: We will then check the results and determine the effectiveness of our method by comparing it with existing ones.

The Public Opinion Survey contained various questions to break down the role played by product reviews for various consumers while buying a product. All the volunteers used to purchase stuff online.

The survey was planned to estimate how people's perceptions were shaped while relying on product reviews when making a purchase. The results fell in line as expected and bolstered arguments established previously.

# **III PROBLEM DISCRIPTION**

The main objective is to increase customer satisfaction and enhance the experience of shopping online. For this purpose, numerous online stores now facilitate customers to post reviews about the products they bought. With more people getting used to the internet, a large number of people are rising to write reviews and

post them on websites, which helps fellow shoppers. Reviews also affect the profitability of the business of e- commerce.

Consequently, the number of reviews for products is growing exponentially. Best- selling items can have thousands of reviews on leading retail platforms. With anyone being able to leave an opinion, this also provides opportunities for individuals and entities to leave deceptive or spam reviews to support or discredit particular products. The system in place today does not successfully filter out spam and useless reviews and comments.

Hence, an urgent need has arisen to develop an intelligent system that can automatically process opinions and classify them as spam or non-spam.

### IV. EXISTING APROACH

By taking several public opinion polls, it is possible to analyze that individuals do read and act upon online ratings and product

reviews. A survey by a leading website found that:

- More than 80% of the online buyers go through the existing reviews.

- 50% base their purchasing decisions on the ratings of the products.

- 30% of the consumers compare ratings of competing products prior to making their decision.

Evidently, consumers appreciate the opinions of fellow



users, and businesses that offer these products do too. Websites, blogs, discussion forums, and other media are repositories of customer recommendations, which are valuable sources of text data. Therefore, both younger and older generations significantly depend on online reviews. This means that individuals tend to make a choice about buying a product by considering and thinking over available opinions on it. If a prospective customer forms an honest overall perception of a product by taking into account existing sentiments around it, there is every chance that they will proceed to make the

purchase. Generally, if the percentage of positive and impactful opinions is large, the general impression will be highly positive. On the other hand, if the general impression is bad, chances are that they will not purchase the product.

Nowadays, customers are free to write any opinion, which can encourage people and organizations to write false reviews to promote or discredit some products.

#### **Naive Bayes Theorem**

Naive Bayes Classification is a supervised machine learning algorithm that is simple yet extremely efficient for classification problems. It works under certain assumptions, and even slight breaks in these assumptions usually do not affect the performance of the algorithm considerably. The Naive Bayes theorem relies on the following main assumptions:

1. All features are conditionally independent.

2. All random variables in the assumptions are independent of one another and are drawn from similar distributions.

3. The Naive Bayes theorem is derived from the mathematical Bayes Theorem in probability.

The Bayes theorem states:

#### **Posterior Probability**

The posterior probability can be

understood as: "What is the probability that a specific object belongs to class i, given its observed feature values?"

For instance,

$$P(A_j|E_i) = \frac{P(E_i|A_j)}{P(E_i)} \cdot P(A_j)$$

In this equation,

 $\ (A_{i} \ ) \ is the \ (i \ )th class of the classes (1, 2, 3, m)$ 

 $(E_{i} ))$  is the feature vector of the (i ) th sample (1, 2, 3, n)

In other words, the posterior probability shows the probability of sample i being in class  $(A_{i} ))$ , given the feature vector  $(E_{i} ))$ .

#### V. PROPOSED APPROACH

Sentiment Analysis

Review analysis includes extracting and aggregating positive and negative opinions from product reviews



using opinion mining. This paper aims to generate feature-based summaries from user reviews of items on e-commerce sites.

From a dataset of a particular item, e.g., a phone, the problem can be divided into three subproblems: (1) Determining the product features described by the customers in the review field. (2) For each product feature, identifying the review sentences that reflect an opinion (positive or negative). (3) Preparation of a summary report based on the different observations collected. Also, the analysis involves the detection of spam reviews by identifying fictitious or misleading content. Spammy language-based reviews can be screened out so that there can be a proper assessment of the product.

#### **Decision Tree**

A decision tree is a useful decision-making support tool. It is tree-shaped, and it is a model of a decision and its potential outcomes, such as utility, cost of resources, and outcomes of chance. This is one of the methods of visualizing conditional control statements.

Among the several uses of decision trees, they are used most widely in operations research, specifically for analyzing decisions to develop a plan for reaching intended results. They are also an integral part of machine learning.

A decision tree is a graph-like representation where every node corresponds to a "potential outcome" given an attribute (for instance, the outcome of rolling a die, which may be 1, 2, 3, 4, 5, or 6). Every branch signifies the outcome of an action, and every leaf node signifies a class label (the decision made after checking all test cases). The route from the root node to a leaf shows the classification rules.

In decision analysis, a decision tree and its corresponding diagram are utilized as a graphical aid and decision-making tool, where the expected utilities of rival alternatives are determined. decision tree, and they are:

1. Chance nodes – with circles used to symbolize them.

2. Decision nodes with squares used to represent them.

3. End nodes with triangles used to indicate them.

Decision trees find extensive applications in operations research and management. Wherever decisions must be taken on-line without an option to correct them and

where the information is incomplete, one should apply a probability model with the decision tree as a best choice model or an online selection model algorithm.

Business, health economics, and public

health undergraduate students learn about decision trees, influence diagrams, utility functions, and other tools and methods of decision analysis, which fall under operations research or management science.

To train a model, we use a decision tree, which predicts values or classes of the target variables by learning decision rules from training data.

The Decision Trees algorithm is simpler to comprehend than other classification

algorithms. It uses a tree structure to solve problems. In the structure, every internal

node represents an attribute, and every leaf node represents a class label.

#### Decision Tree Algorithm Pseudo code

1. The root of the tree must have the best attribute of the dataset.

There are three kinds of nodes utilized in a

2. The training set is split into subsets, with each subset having the same values for a given attribute.

Steps 1 and 2 are iterated for each subset until leaf nodes are formed in all branches of the decision tree.

# VL PERFORMANCE ANALYSIS

Sentiment is a feeling, thought, or opinion, and is usually linked to opinion mining,

which analyzes how individuals feel about different entities. The internet is full of

sources that give sentiment information. From the user's point of view, people can post their opinions on many social

networking sites, such as forums and microblogs.

Figure 6.1 illustrates our suggested model for classification, which is also used in this paper as the framework. Our primary

contributions are contained within Phases 2 and 3. In Phase 2:

1) We suggest and develop an algorithm to detect negative phrases;

2) We add a mathematical solution for calculating sentiment scores;

3) We introduce a feature termed as the vector generation method for classifying sentiment polarity.

In Phase 3



Fig 6.1: Sentiment Polarity Categorization Process



# VIL CONCLUSION

In the paper, we have proposed a framework for finding spam or forged product reviews based on Opinion Mining, or commonly known as Sentiment

Analysis. This is performed by analyzing opinions of customers reflected in text

form. First, we determine whether a review relates to a given product using a Decision Tree. For spam detection, we make use of a Spam dictionary and identify spam- words in the reviews. Moreover, we utilize different algorithms in Text Mining through which we can obtain certain

outcomes from our analyses.

#### REFERENCES

1. Bing Liu, "Sentiment Analysis and Opinion Mining", Synthesis Lectures on Human Language Technologies, 2012.

2. Bing Liu, "Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data", Springer, 2011.

3. Xing Fang and Justin Zhan, "Sentiment analysis using product review data", Fang and Zhan Journal of Big Data (2015).

4. Shirani-Mehr, Houshmand. "SMS spam detection using machine learning approach." (2013)

5. http://www.cs.uic.edu.

6. Xie, Sihong, Guan Wang, Shuyang Lin, and Philip S. Yu. "Review spam detection via temporal pattern discovery."

Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD 12, 2012.

7. Fake product review monitoring survey, https://goo.gl/formx/alqmW0q0tObl.mIU2 3

8. NJindal and B.Liu, "Review spam detection", In Proceedings of the 16th

International Conference on World Wide Web, pp1189-1190, 2007.

9. N. Jindal and B. Liu, "Opinion spam and analysis", In Proceedings of the

International Conference on Web Search and Web Data Mining (WSDM), pp219- 230, 2008.

10. G. Wang, S. Xie, B. Liu, and P. S. Yu, "Review graph based online store review spammer detection", IEEE 11th

International Conference on Data Mining, pp 1242-1247,2011.

11. Liu B (2012) Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers

12. Tan LK-W, Na J-C, Cheng Y-L, Chang K (2011) Sentence-level sentiment polarity classification using a linguistic approach.

In: Digital Libraries: For Cultural

Heritage, Knowledge Dissemination, and Future Creation. Springer, Heidelberg, Germany.

13. Hu M, Liu B (2004) Mining and summarizing customer reviews. In:

Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, New York, NY, USA.

14. Gann W-JK, Day J, Zhou S (2014)

Twitter analytics for insider trading fraud detection system. In: Proceedings of the second ASE international conference on Big Data. ASE

15. Roth D, Zehenko D (1998) Part of

speech tagging using a network of linear separators. In:Coling-Acl, The 17th

International Conference on Computational Linguistics.

16. Zhang Y, Xiang X, Yin C, Shang L (2013) Parallel sentiment polarity

classification method with substring feature reduction. In: Trends and

Applications in Knowledge Discovery and Data Mining, volume 7867 of Lecture

Notes in Computer Science. Springer Berlin Heidelberg, Heidelberg, Germany.

17. Ranks.nl, "Stopwords". [Online]. Available: http://www.ranks.nl/stopwords.

Han, Jiawei, Micheline Kamber, and Jian
Pei. Data mining: concepts and
techniques 3rd Edition. Morgan Kaufmann
Publishers, 2013.

19. Han, Jiawei, Jian Pei, and Yiwen Yin."Mining frequent patterns without candidate generation." ACM Sigmod Record.Vol. 29. No. 2. ACM, 2000.

20. Manleen Kaur Kohli, Shaheen Jamil Khan, Tanvi Mirashi, Suraj Gupta, "Fake Product Review Monitoring and Removal for Genuine Online Product Reviews Using Opinion Mining".

21. Fake Product Review Removal for Actual Product Reviews Through Opinion Mining, take off edu group, A young mind technology solutions P. Ltd.