ISSN:1533-9211
SEYBOLD REPORT
www.seybold-report.com

Twitter Data Pre-Processing and Detection of Fake Reviews

Shrinivas Biradar
Reseach Scholar Dept. of Computer Science and Engineering, SJC Institute of Technology, Chickaballapur

Dr. G T Raju
Professor Dept. of Information Science and Engineering, SJC Institute of Technology, Chickaballapur

Divakar K M
Reseach Scholar Dept. of Computer Science and Engineering, SJC Institute of Technology, Chickaballapur

ISSN:1533-9211
SEYBOLD REPORT
www.seybold-report.com

ISSN: 1533 - 9211

CORRESPONDING AUTHOR:

Shrinivas Biradar
shrinivas.biradar067@gmail.com

KEYWORDS:

Sentiment Analysis, Machine Learning, Logistic Regression, Naïve Bayes, SVM, Twitter, Pre-Processing

Received: 26 December 2023
Accepted: 06 January 2024
Published: 13 January 2024

TO CITE THIS ARTICLE:

Biradar, S., Raju, G. T., & Divakar, K. M. (2024). Twitter Data Pre-Processing and Detection of Fake Reviews. *Seybold Report Journal*, *19*(1), 72-86. DOI: 10.5110/77. 1097

Abstract

Over last few decades, Twitter has grown enormously. It is utilized now by a million people who disseminate knowledge about their lives and feelings. However, to process and analyze this information, evaluation of sentiment and theme modeling methods are used by the applications. The preliminary processing procedures necessary to derive characteristics using Twitter data are detailed in this study. The technique enables an improvement in sentiment estimation efficiency. In the past couple of decades, sentiment analysis of Tweets has advanced significantly by conducting many experiments in this area of expertise. Several machine-learning algorithms have been employed over the years and anticipate being employed in the future for sentiment analysis in the upcoming days to accomplish further accurate results. Nowadays people possess knowledge about platforms a place where people could convey opinions freely, hence this provides a chance for sentimental analysis could be carried through and the findings utilized for various goals since it may provide insight into a person's intense condition. There are three main techniques employed in Sentiment Analysis: 1. Machine learning-based 2. Sentiment lexicon-based 3. Hybrid approach. In this study, we propose an approach that involves extracting attributes once an initial preprocessing of tweets. Furthermore, we utilize Logistic Regression, Naïve Bayes, and Support Vector Machine (SVM) Classifiers to detect fraudulent reviews. The rise of online platforms and the growing influence of user-generated content have prompted a rise in the predominance of fake reviews. Fake reviews can be harmful to businesses and consumers alike, as they mislead potential customers and mislead market dynamics. Detecting fake reviews has become a critical task in maintaining the integrity of online review systems. Nevertheless, fake review detection poses numerous challenges. Confrontational attacks by individuals or groups trying to avoid detection classifiers require continuous refinement and adaptation of detection techniques. In this paper, the main objective is to highlight the significance of data pre-processing techniques and show how it could benefit to enhance the accuracy. In this paper, we additionally elucidate the factors contributing to the enhanced accuracy through a meticulous examination of each approach.

Introduction

Sentiment analysis is a developing method in data analytics that can be employed to categorize subjective viewpoints concerning specific subjects within documents. Machine Learning techniques have been broadly applied in customer reviews, social networks, news, politics, and blogs. It's important to note people purchase and respond to evaluations of the things they purchase have evolved dramatically as a result of the web's quick and vast expansion. [7]. However, in the swift development of information technology, the Internet has emerged as an exceptionally crucial platform for capturing customer opinions, attitudes, feelings, and emotions. Given the continuous creation and dissemination of new documents in real-time on the Internet, there is a pressing need for an expert system capable of aggregating people's opinions from these documents and promptly discerning their sentiment [8]. During the previous period, internet traffic has nearly doubled [1]. In this digital age, the internet's evolution has elevated various online platforms like Facebook, Twitter, LinkedIn, and others to prominence. Within this dynamic digital landscape, trends swiftly emerge and gain popularity over Online Social Networks (OSNs). Various methods of sharing and communication rely not only on the content itself but also on its repetition [1]. Twitter, a renowned micro-blogging and public networking podium, lets operators to share, convey, and interpret 140-character messages known as tweets [1]. With 320 million periodic active users, Twitter is accessible over both its website and mobile interfaces. It has been asserted 80% of users remain engaged through mobile phones. In micro-blogging services, operators tend to make spelling errors and employ emoticons to convey their opinions and emotions. Natural Language Processing plays a pivotal character and could be utilized to analyze the sentiments expressed. Methods for sentiment analysis rely on supervised machine learning. The following are the three primary machine learning categories used for sentiment analysis:

1. Naive Bayes
2. Logistic Regression Model
3. Support Vector Machine Algorithm.

The feature extraction and the examined datasets determine how accurate the classifiers are. As an illustration, SVM performs more effectively while simply using unigrams, and adding bigram parameters would decrease its precision. [4]. Currently, in order to get greater outcomes Machine learning approaches, employ a training data set to absorb information from previous analyses and produce superior findings, which is why they are frequently employed. Machine learning algorithms and classifiers accurately detect the emotional content associated with every sentence. [5]. In this paper, we present a methodology that involves gathering datasets, preprocessing Twitter data, and employing feature extraction techniques. Furthermore, in this work, we also have dealt with the detection of fake reviews and to achieve this we have used above mentioned classifiers. Reviews are classified as fictitious because attempts to influence them are seen as dishonest since reviews are seen as real feedback-sharing platforms. As a result, identifying fraudulent evaluations has become and is currently a need for scientific evaluations. [10].

Related Work

In sentiment analysis, data pre-processing is a very important phase to bring more quality and accuracy to the outcome. There are multiple techniques accessible for data pre-processing. Researchers [1, 3, 4] worked extensively on tweeter data pre-processing and have contributed to improving the data quality moreover, the improvement in the outcomes. A great challenge for researchers has been the outcome of fake reviews/opinions and enormous work has been done, even though still it is crucial to deal with this issue. Researchers [11, 13, 15] show good results for the detection of fake reviews using the desired machine learning approaches.

Challenges of Tweets Pre-processing and Fake review detection

Twitter users employ informal language, creating their identifiable words, implying shortcuts, punctuation, jargon, novel terms, URLs, and genre-specific terminologies and shortenings. Therefore, such script requires careful handling. Consequently, for dataset analysis, it is necessary to eliminate slang words, emoticons, stop words, punctuation, URLs, and so on. We utilized Stanford Parser Tools for POS tagging and the Natural Language Toolkit for stop-word removal plus lemmatization. It is imperative to scrutinize operators who are ranking the products, amenities, and services delivered by numerous websites. This entails analyzing user behavior, perspectives, and attitudes, and subsequently normalizing them. Identifying fake reviews poses a significant challenge in the realm of knowledge discovery and classification tasks. For the past decade, researchers have concentrated on various perspectives for detecting deception in review data [11].

Data Collection

In this work, data was gathered from the Twitter application. By utilizing the SNScrape Python library, real-time collection of tweets about cell phones from users was conducted, focusing on specific keywords. From 1 January 2020 to 1 January 2023. Python, along with the Natural Language Toolkit (NLTK), is employed for processing the data. The dataset considered in this work is presented in Fig. 1 with all possible attributes.

```
       Unnamed: 0      phone name        date          username  \
0               0  realme c11 2021  31-10-2021  flipkartsupport
1               1  realme c11 2021  23-10-2021      realmecarein
2               2  realme c11 2021  13-10-2021  flipkartsupport
3               3  realme c11 2021  10-10-2021  flipkartsupport
4               4  realme c11 2021  18-09-2021  wahinya_charlie
...           ...              ...         ...               ...
62558       62558         realme 8  15-05-2018         91mobiles
62559       62559         realme 8  15-05-2018      telecomtalk
62560       62560         realme 8  15-05-2018  indiatodaytech
62561       62561         realme 8  15-05-2018      realmeindia
62562       62562         realme 8  04-05-2013        crackshot

                                                   tweet language  \
0       Sorry for the experience. We'll surely help y...       en
1       Hi Ravinder, we would like to inform you that...       en
2       However, the product realme C11 2021 (Cool Bl...       en
3       We understand your concern about the delivery...       en
4      Need a phone and you're on a budget  Get the  ...       en
...                                                  ...      ...
62558  Realme 1 with 6-inch FHD+ display and Helio P6...       en
62559  Oppo Realme 1 Smartphone With Dual 4G Support ...       en
62560  Realme 1 launched with Helio P60, Android Oreo...       en
62561  Realme 1 features a 6.0" FHD + Full screen, th...       en
62562  I have 29 longboxes of comics in my mom's base...       en

       replies_count  retweets_count  likes_count  \
0                  1               0            0
```


```
       Unnamed: 0      phone name        date          username  \
0               0  realme c11 2021  31-10-2021  flipkartsupport
1               1  realme c11 2021  23-10-2021      realmecarein
2               2  realme c11 2021  13-10-2021  flipkartsupport
3               3  realme c11 2021  10-10-2021  flipkartsupport
4               4  realme c11 2021  18-09-2021  wahinya_charlie
...           ...              ...         ...               ...
62558       62558         realme 8  15-05-2018         91mobiles
62559       62559         realme 8  15-05-2018      telecomtalk
62560       62560         realme 8  15-05-2018  indiatodaytech
62561       62561         realme 8  15-05-2018      realmeindia
62562       62562         realme 8  04-05-2013        crackshot

                                                   tweet language  \
0       Sorry for the experience. We'll surely help y...       en
1       Hi Ravinder, we would like to inform you that...       en
2       However, the product realme C11 2021 (Cool Bl...       en
3       We understand your concern about the delivery...       en
4      Need a phone and you're on a budget  Get the  ...       en
...                                                  ...      ...
62558  Realme 1 with 6-inch FHD+ display and Helio P6...       en
62559  Oppo Realme 1 Smartphone With Dual 4G Support ...       en
62560  Realme 1 launched with Helio P60, Android Oreo...       en
62561  Realme 1 features a 6.0" FHD + Full screen, th...       en
62562  I have 29 longboxes of comics in my mom's base...       en

       replies_count  retweets_count  likes_count  \
0                  1               0            0
```


Figure 1 Data Collection

Methodology

Logistic Regression: This regression method is well-suited for modeling a dependent binary variable. It establishes the linking among a single binary dependent variable and other independent variables at nominal, ordinal, interval, or ratio levels. As an alternative to continuous or numerical values, it outputs a probabilistic value. Probabilistic values, sometimes referred to as probability values, are consistently in the 0 to 1 band.

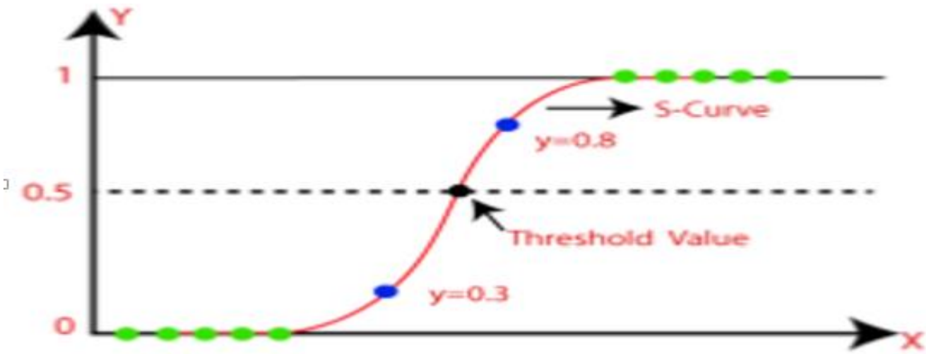

FIGURE 1 LOGISTIC REGRESSION CURVE

Fig.2 demonstrates the logistic regression classifier's data flow schematic. In the previously discussed image, the outcome will be assigned to class 1 if the probability outcome is greater than 0.5, it will fall into class 1; otherwise, it will be classified as class 0 [5]. Three distinct forms of logistic regression exist. 1. Binary Logistic Regression: This is used when the response variable is dichotomous, meaning it falls into one of two categories. 2. Multinomial Logistic Regression: Is applied when the response variable can be classified into three or more categories, with no inherent ordering among them. 3. Ordinal Logistic Regression: This type of regression assumes that the categories are ordered and that the response variable falls into one of three or more ordinal categories. Naive Bayes, an efficient probabilistic classification model, is also utilized in this study. It is grounded on Bayes' theorem with the assumption of feature independence. It's generally used in several machine learning projects, including text classification, spam filtering, and sentiment analysis. The Support Vector Machine (SVM) is a prominent classifier used to discover the hyperplane that separates a dataset into multiple categories in a multi-dimensional space. This hyper land requires a comparable range between it, and two hyper lands that include the closest data points belonging to two distinct categories, consequently, are considered to be on either side of the hyper land. SVM employs kernel features to transform data elements from a non-linearly separable space into a linearly separable one, especially for datasets that are not linearly distinguishable [7].

Data Processing

Pre-processing is an essential data preparation step for sentiment classification [6]. In this study, data was extracted from Twitter application. Pre-processing encompasses tasks such as data cleansing, case folding, tokenization, stop-word removal, and lemmatization, detect the fake review with a Naïve Bayes and Logistic Regression classifiers. The complete process of text normalization is depicted in the fig.3 [1].

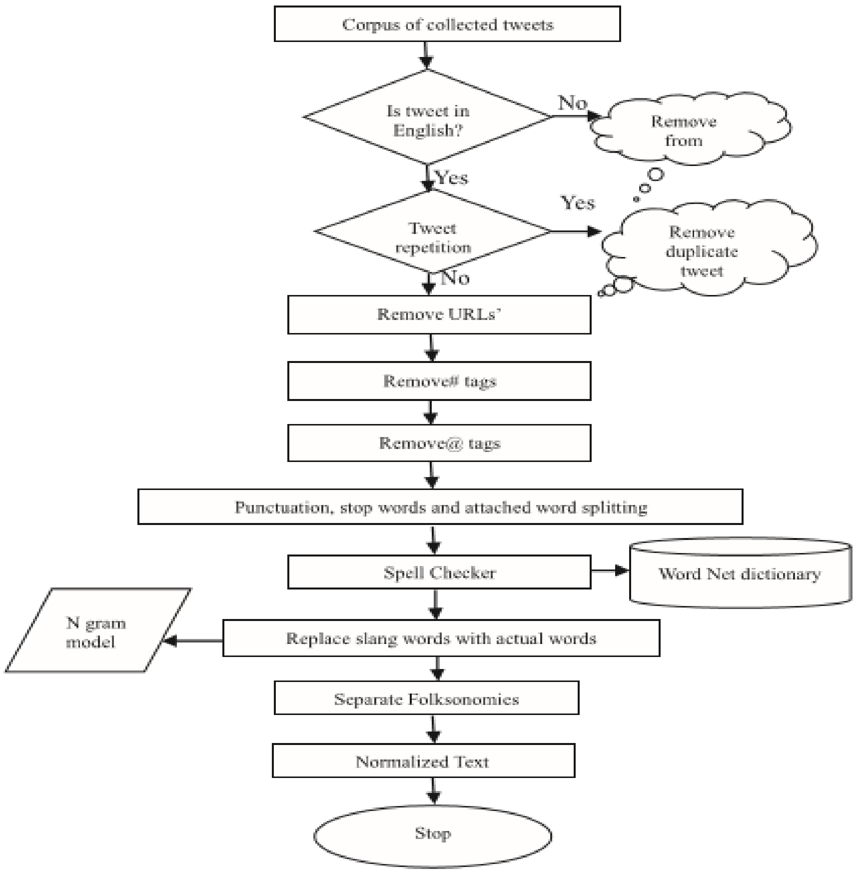

FIGURE 2 TEXT NORMALIZATION PROCESS

• Data scrubbing: The process of removing extraneous or irrelevant information from tweet data to ensure its relevance.

• Case Normalization: The procedure of changing words to a consistent case, such as all uppercase or all lowercase letters.
• Tokenization is the process of breaking down text into individual units, which can be words, phrases, or other important elements.

• The elimination of mutual and regularly used words, without significantly altering the sentence structure, is known as stop word elimination.

• Lemmatization: The method of reducing a word to its base or root form by removing affixes and suffixes [2]. Text Mining: The practice of uncovering and extracting information from vast, unstructured textual resources. It has the ability to make use of unstructured data sources. Text mining encompasses a series of steps [2]: Data retrieval

- Preliminary, cursory, and comprehensive language examination

- Recognize relevant entities and gather information about them
- Data mining
- Merge and establish connections between extracted information.

- The performance metrics used in this work include precision, recall, and the F-measure. Precision and recall are particularly beneficial for evaluating unbalanced datasets. These metrics can be calculated using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) measures, where TP is the rate of correctly classified instances as positive, TN is the rate of correctly classified instances as negative, and FP refers the rate of incorrectly classified instances as positive. The fraction of erroneously identified negative instances is denoted by FN. Accuracy: It is a statistical measure used to assess the performance of a prediction model. The application of such measures is performed using the following formulas [12]:

$$\boxed{\text{Accuracy} = \text{TP+TN / TP+TN}}$$

Precision: It assesses the accuracy of positive predictions by measuring the proportion of true positives.

$$\boxed{\text{Precision=TP+TN / TP+FP}}$$

Remember: This metric, known as sensitivity, evaluates a model's capacity to predict positive labels.

$$\boxed{\text{Remember =TP+TN / TP+FN}}$$

F-Measure: This measurement takes Remember and Precision into account and can be viewed as a weighted average of Precision and Remember values ranging from 0 (worst) to 1 (best). F-measure is calculated with the equation given below.

$$\boxed{\text{F−Measure=2×Precision×Recall /}}$$

Lower uppercase letters: The process of data pre-processing is case sensitive hence the program will assume “review” and “Review” as two very different words. In data pre-processing this is the very first phase which is to change the uppercase letter to a lowercase letter. However, it very essential to consider these two words are the same features, otherwise the classifier will impact potential differences in feelings between these two terms. Consider these three sentences: “Phone is better”, “Amazing phone”, and “worst Phone”. The word "phone" is used in the first two of the statements, which are both positive and in the third sentence, which is a negative term. The classifier will presuppose that sentences without the word "phone" are far more probable to be negative than those with it. If the uppercase letters had been swapped out for lowercase ones, the classifier would certainly have been able to infer that the word "phone" is not particularly important in determining whether the phrase is positive. Since the data were obtained through Twitter, the following step is much more crucial. Because social media members frequently utilize capital letters even when it is not necessary, this initial processing phase is likely to have a greater influence on data from social media than on traditional "classical" data.[3]. **Remove URLs and user references:** The Twitter application permits user messages to contain hashtags, references, and URLs. However, client URLs and references are not essential for the content analysis of the text. In order to reduce the overall number of attributes gathered from the corpus, this pre-processing phase employs a regular expression to identify and replace each URL with "URL" and each user reference with "AT_USER" [2]. It's worth noting that during tokenization, hashtags are retained, as they often contain a word of significance in the analysis, with only the '#' symbol being removed.

Remove digits: Since for analyzing the data digits will not play a vital role so they could be eliminated from the sentences. It may be noted that letters and numbers are occasionally mingled together, and doing away with them may make it possible to link two qualities that the classifier may have previously separated. Consider, some text may comprise “realme”, while others “realme7”.

Remove stop words: Although these terms are more often employed in the language, they are crucial for pre-processing. Stop words are frequently eliminated from the collection in natural language processing. Removing stop words helps reduce the number of features gathered from the samples. To aid in comprehension, a compilation of every one of the stop words employed in the content is displayed in Fig. 4.

```
{'o', 'it', 'we', "don't", 'your', 'that', 'until', 'this', 'had', 'needn', "shan't", 'am', 'between', "doesn't", 'haven', 'int
o', 'their', 'ours', "you'll", "shouldn't", 'her', 'most', 'having', 'to', 'only', 'why', 's', 'because', "mightn't", "should'v
e", 'being', 'mustn', 'will', 'there', 'hadn', 'itself', 'with', "mustn't", "needn't", 'himself', 'where', 'do', 'yourself', 'o
wn', "you're", 'is', 'y', 'nor', "weren't", "you'd", 'ain', 'by', 'they', "it's", 'our', 're', 'be', 'or', 'didn', 'was', 'her
e', 'each', 'a', 'some', "couldn't", 'out', 'his', 'doesn', 'of', 'were', 'myself', 'from', 'then', 'you', 't', 'wasn', 'whic
h', 'couldn', 'again', 'should', 'over', 'theirs', 'an', "won't", 'after', 'ma', 'isn', 'shouldn', 'how', 'my', 'themselves',
'don', 'll', 'them', 'those', 've', 'have', "haven't", 'few', 'been', 'who', "hasn't", 'does', 'these', 'yourselves', 'doing',
'hasn', 'weren', 'off', 'further', 'above', 'when', 'he', "hadn't", 'wouldn', "you've", 'the', 'very', 'yours', 'aren', 'hers',
'under', 'while', 'm', 'for', 'its', 'before', 'and', 'no', 'she', 'at', 'can', 'during', "aren't", 'but', 'mightn', 'other',
'what', 'just', 'up', 'same', 'as', 'd', "wasn't", 'both', 'herself', 'are', 'through', 'against', 'such', 'once', 'i', 'in',
'won', 'if', 'me', 'too', 'than', 'not', 'more', 'did', "that'll", 'him', "she's", "didn't", 'shan', 'has', 'so', 'all', "would
n't", 'down', 'on', 'about', 'below', 'ourselves', 'now', 'whom', "isn't", 'any'}
```

Figure 3 Stop Words

Remove repeated letters: When users wish to emphasize an expression, users of the Twitter program frequently recite certain letters countless times. Think about the subsequent tweets that use the word "love" many times, such as "I loooovvvee that!" The first two instances of duplicated letters will be retained. Thus, "loooovvveee" for the preceding example will now be "loovvee" [3].

To detect the POS tags: from the samples we have used the NLTK method pos_tag. Lemmatize: In this phase of data pre-processing it will eliminate the plurals and conjunctions. For instance, in this instance, "phone" and "phones" would be regarded as two distinct properties. It is possible to lemmatize the unigrams to enhance the attribute elimination technique. The WordNet implementations featured in the NLTK package will serve as the foundation for the lemmatization [3]. Finally, the results of all the phases have been combined in one file and the results are stored in the excel sheet on the local machine, for this purpose we have made use of the python library called openpyxl.

Tokenize: Each word in this text is given its own space, which allows the token to be created by dividing the language into equal halves. Separating punctuated tokens from the word is used in this research's tokenization process. The experiments would use the NLTK word_tokenize method for tokenizing its samples. Since the in the fig.5 we have shown the results got after the tokenization phase.

```
['sorry', 'for', 'the', 'experience', '.', "we'll", 'surely', 'help', 'you', 'out', '.', 'could', 'you', 'please', 'confirm',
'if', "you're", 'referring', 'to', 'the', 'product', "'", 'realme', 'c11', '2021', '(', 'cool', 'blue', ',', '64', 'gb', ')',
"'", 'so', 'that', 'we', 'can', 'look', 'into', 'it', 'and', 'assist', 'you', 'accordingly', '?', 'awaiting', 'your', 'respon
se', '.', '(', '1/2', ')']
['hi', 'ravinder', ',', 'we', 'would', 'like', 'to', 'inform', 'you', 'that', 'realme', 'c11', '2021', 'does', 'not', 'suppor
t', 'clone', 'app', 'feature', '.', 'thanks', 'for', 'understanding', '.', 'further', ',', 'if', 'you', 'have', 'any', 'quer
y', 'feel', 'free', 'to', 'contact', 'us', 'again', '.']
['however', ',', 'the', 'product', 'realme', 'c11', '2021', '(', 'cool', 'blue', ',', '32', 'gb', ')', 'will', 'be', 'deliver
ed', 'on', 'october', '18', ',', '2021', 'and', 'the', 'product', 'realme', 'c11', '2021', '(', 'cool', 'grey', ',', '32', 'g
b', ')', 'on', 'october', '14', ',', '2021.your', 'patience', 'is', 'much', 'appreciated', '.', '^', 'jm', '(', '2/2', ')']
['we', 'understand', 'your', 'concern', 'about', 'the', 'delivery', 'of', 'your', 'order', '.', 'we', 'would', 'like', 'to',
'inform', 'you', 'that', 'your', 'product', '"', 'realme', 'c11', '2021', '(', 'cool', 'blue', ',', '64', 'gb', ')', '"', 'i
s', 'moving', 'as', 'per', 'the', 'schedule', 'and', 'will', 'be', 'delivered', 'by', 'october', '10', ',', '2021', '.', 'app
reciate', 'your', 'patience', '.', '(', '1/2', ')']
['need', 'a', 'phone', 'and', "you're", 'on', 'a', 'budget', '🤔', 'get', 'the', 'at', 'any', 'supermarkets', 'for', 'only',
'ksh', '9,490', '/', '-', '.', 'offer', 'valid', 'toll', '19th', 'september', '.']
['very', 'pocket', 'friendly']
['the', 'best', 'smartphone', 'deal', 'out', 'here']
```

FIGURE 4 RESULT OF TOKENIZATION

Detect POS tags: To clarify the significance of a word, consider its part of speech. Although the meaning of the word "like" in the phrases "I like that" and "I am not like you" may seem to make sense to the reader as having distinct connotations, the classifier will treat them as having the exact same meaning when calculating the bag of words [3].

Fake review detection

What is a fake review? A fake review is considered a type of opinion spamming [13]. Online reviews are becoming more influential and of higher quality as the internet expands daily. Reviews have the power to persuade consumers across a wide range of businesses, but they are especially significant in the world of online shopping, where customers' comments and reviews of goods and services are frequently the easiest, if not the only, method to determine whether to purchase them [14]. Authentic users write reviews with genuine intentions, aiming to share their positive or negative experiences with the product. This information can be valuable for other potential users [15]. Fake review detection using sentiment analysis can be achieved by analyzing the sentiment of the text and looking for patterns that are indicative of fake reviews. The rise of online platforms and the growing influence of user-generated content have led to an increase in the predominance of fake reviews. A basic approach to fake review detection is to analyze reviews manually. This approach is based on the premise that humans can detect when other humans behave in fraudulent ways [9]. Fake reviews can be harmful to businesses and consumers alike, as they mislead potential customers and mislead market dynamics. Detecting fake reviews has become a critical task in maintaining the integrity of online review systems. However for detecting the fake review we have used the nltk libraries and other built in packages. Here the sample is considered which is reviews contains both fake and genuine and after feeding it through the fake detection model we have received the results that are classified as fake reviews and genuine reviews. In this work we have used Naive Bayes and Logistic Regression classifiers for classification tasks. Naive Bayes: A probabilistic machine learning classifier rooted in Bayes' theorem, it operates on the assumption that the presence of a specific feature in a class is unrelated to the presence of other features. In spite of its simplicity, Naive Bayes has been proven to be effective in many text classification tasks. Logistic Regression: It is a statistical model which models the relationship between a set of features and a binary outcome using the logistic function, which maps the input features to a probability value between 0 and 1. Support Vector Machine: This is yet another and simple classifier widely used in the detection of fake reviews. Naive Bayes, Logistic Regression and Support Vector Machine can be effectively used for classification tasks, including sentiment analysis, spam detection, and fake review detection. Here we have considered the sample reviews and then the classification of the same into fake and genuine.

```
Pseudo code to detect fake review
for each review in all_reviews:
    if is_fake_review(review):
        print("Fake Review: " + review)
    else:
        print("Genuine Review: " + review)
function is_fake_review(review):
     fake_score = 0
    if review.rating > MAX_RATING or review.rating <
MIN_RATING:
        fake_score += 1
if contains_generic_phrases(review.text):
        fake_score += 1
if
has_suspicious_history(review.reviewer):
        fake_score += 1
    return fake_score >= FAKE_THRESHOLD
function contains_generic_phrases(text):
    generic_phrases = ["great product", "terrible service", "amazing
experience"]
    for phrase in generic_phrases:
        if phrase in text:
            return True
    return False
function has_suspicious_history(reviewer):
    if reviewer.num_reviews > MAX_REVIEWS_THRESHOLD and
reviewer.review_frequency < MIN_REVIEW_FREQUENCY:
        return True
return False
```

Reviews = ["This phone is incredible, I would highly suggest it to anyone!", "I don't believe how awful this phone is, it's a complete waste of money.", "This is a scam, don't waste your money on this phone.", "I was not sure about this phone at first, but it turned out to be great!", "This phone is a total fraud, I regret buying it."] Results: Genuine review: This product is amazing, I would highly recommend it to anyone! Fake review: I can't believe how terrible this product is, it's a complete waste of money. Fake review: This is a scam, don't waste your money on this product. Genuine review: I wasn't sure about this product at first, but it turned out to be great! Fake review: This product is a total fraud, I regret buying it. Additionally, the outcomes depicted in figures 5, 6, 7, and 8 pertain to the evaluation of the Naïve Bayes classifier, Logistic Regression, and Support Vector Machine. This assessment is based on performance metrics including Accuracy, Precision, Recall, and F1-score, all of which are visually represented in the corresponding graphs.

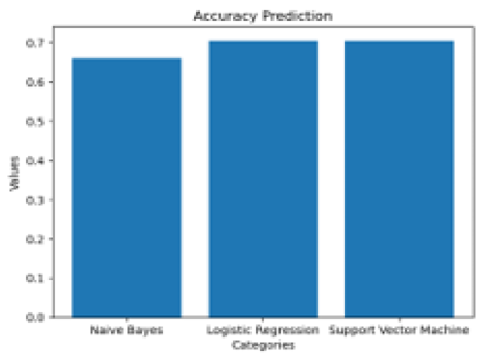

Figure 6 Accuracy Outcome Comparison

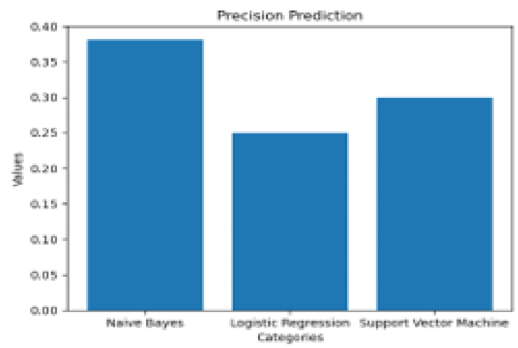

Figure 7 Precision Result Comparison

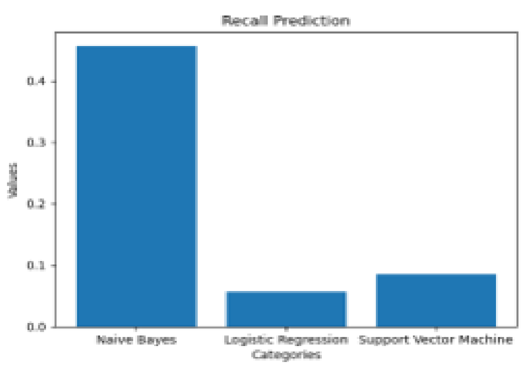

Figure 8 Recall Prediction Comparison

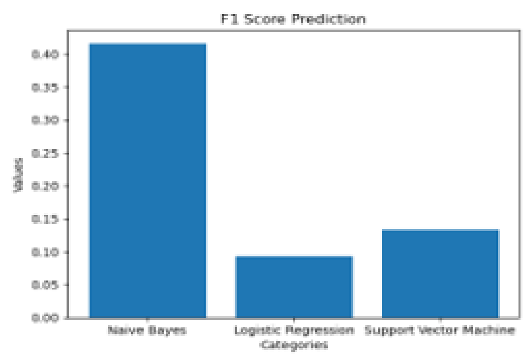

Figure 9 F1 Score Outcome Comparison

TABLE 1 RESULTS FOR DETECTION OF FAKE REVIEWS

Performance Metrics / ML Classifier	Accuracy	Precession	Recall	F1-Score
Naïve Bayes	0.6590	0.3809	0.4571	0.4155
Logistic Regression	0.7045	0.25	0.0571	0.0930
Support Vector Machine	0.7045	0.3	0.0857	0.1333

Conclusion

This paper establishes that a crucial preliminary step in the classification task involves text pre-processing and feature extraction. This process significantly influences the quality of the final outcome. The results we have got are quiet promising scores but it can still improve in the future. The experimental results for fake review detection using Naïve Bayes, Logistic Regression, and Support Vector Machine indicate that Naïve Bayes performs reasonably well when considering all parameters. On the other hand, SVM excels in terms of accuracy and precision, as depicted in the table shown above. For other sentiment analysis problems this approach can be used and extended. However the classifier accuracy can be improved by considering more feature combinations. In this work we have focused on exploiting the results of the limited parameter. In future more features could be taken into consideration so that the prediction should be more accurate.

COMPETING INTERESTS

The authors have no compting interest to declare.

Author's Affiliation

Shrinivas Biradar

Reseach Scholar Dept. of Computer Science and Engineering, SJC Institute of Technology, Chickaballapur: shrinivas.biradar067@gmail.com

Dr. G T Raju
Professor Dept. of Information Science and Engineering, SJC Institute of Technology, Chickaballapur: gtraju1990@yahoo.com

Divakar K M
Reseach Scholar Dept. of Computer Science and Engineering, SJC Institute of Technology, Chickaballapur: dnshgowda@gmail.com

HOW TO CITE THIS ARTICLE:

Biradar, S., Raju, G. T., & Divakar, K. M. (2024). Twitter Data Pre-Processing and Detection of Fake Reviews. *Seybold Report Journal*, *19*(1), 72-86. DOI: 10.5110/77. 1097

REFERENCES

[1] Tajinder Singh and Madhu Kumari, "Role of Text Pre-Processing in Twitter Sentiment Analysis", Twelgth International Multi-Conference on Information Processing-2016.

[2] Muhammad Abdullah Hadi, Nizirwan Anwar, Budi Tjahjono, Lina, Binastya Anggara Sekti, Yunita Fauzi Achmad and Yulhendri, "Go-Food Sentiment Analysis Using Twitter Data, Compared the Performance of the Random Forest Classifier with That of the Linear Support Vector Classifier",MIMSE-I-C 2022, ACSR 102, pp. 3-3, 2022.

[3] Pierre FICAMOS and Yan LIU, "A Topic based Approach for Sentiment Analysis on Twitter Data", International Journal of Advanced Computer Science and Applications,Vol. 7, No. 12, 2016.

[4] Yuxing Qi and Zahratu Shabrina, "Sentiment Analysis using Twitter data: a complete application of lexicon- and machine-learning-based approach" Spinger-2023.

[5] M. Hamza and S.Gupta, "A Comparison of Sentiment Analysis Classifiers on Twitter Data using Machine Learning".

[6] Akrivi Krouska, Christos Troussas and Maria Virvou, "Thr Effect of preprocessing techniques on Twitter Sentiment Analysis.

[7] Huu-Thanh Duong and Tram-Anh Nguyen-Thi, "A review: preprocessing techniques and data augmentation for sentiment analysis", Springer-2020.

[8] Kyoungok Kim, "An improved semisupervised dimensionality reduction using feature weighting: Application to sentiment analysis", Elsevier-2023.

[9] Joni Salminena,b, Chandrashekhar Kandpalc, Ahmed Mohamed Kameld, Soon-gyo Junga and Bernard J. Jansena, "Creating and detecting fake reviews of online products", Elsevier-2022.

[10] Nikhil Chandra Sai Ram, Gowtham Vakati, Jagadesh Varma Nadimpalli, Yash Sah, Sai Karthik Datla, "Fake Reviews Detection Using Supervised Machine Learning", International Journal for Research in Applied Science & Engineering Technology, Volume 10 Issue V May 2022.

[11] Shilpa yadav, Dr.Gulbakshee Dharmela and Khushali Mistry, "Fake Review Detection Using Machine Learning Techniques", JETIR, Volume 8, Issue 4 April 2021.

[12] Mohammed Ennaouri and Ahmed Zellou, "Fake Reviews Detection through Machine learning Classifiers: A Systematic Literature Review", Springer-2021.

[13] Huy Le and Ben Kim "Detection Of Fake Reviews On Social Media Using Machine Learning Classifiers" Issues in Information Systems Volume 21, Issue 1, pp. 185-194, 2020.

[14] Michael Crawford, Taghi M. Khoshgoftaar,Joseph D. Prusa, Aron N.Richter and Hamzah, "Survey of Review Sapm detection Using Machine Learning Techniques" , Springer-2015

[15] Gourav Bathla, Pardeep Singh, Rahul Kumar Singh, Erik Cambria and Rajeev Tiwari, "Intelligent fake reviews detection based on aspect extraction and analysis using deep learning", Springer Nature 2022.

[16] D. V. Lindberg and H. K. H. Lee, "Optimization under constraints by applying an asymmetric entropy measure," *J. Comput. Graph. Statist.*, vol. 24, no. 2, pp. 379–393, Jun. 2015, doi: 10.1080/10618600.2014.901225.

[17] B. Rieder, *Engines of Order: A Mechanology of Algorithmic Techniques*. Amsterdam, Netherlands: Amsterdam Univ. Press, 2020.

[18] I. Boglaev, "A numerical method for solving nonlinear integro-differential equations of Fredholm type," *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.