IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

Customer Review Analysis and Identifying Spam Reviews

Nalwala Fardeen[1], Meraj Ansari[2], Shaikh Hanzala[3], Shaikh Nadim[4], Prof. Ashfaq Shaikh[5]
Students, Department of Information Technology[1,2,3,4]
Assistant Professor, Department of Information Technology[5]
M. H. Saboo Siddik College of Engineering, Mumbai, Maharashtra, India

Abstract: *Gathering customer feedback is essential for any organization, especially in the airline service sector. Surveys are one of the most common ways to collect customer feedback and measure customer satisfaction. However, creating surveys and managing survey responses can be a challenging and time-consuming task. This is where twitter come's into picture, where everybody can share their opinion, this allows airlines to understand how customers feel about their services, identify areas of improvement, and make necessary changes to improve their services. By using this, airlines can gain valuable insights into customer preferences that can help them create more personalized experiences for their customers. Additionally, it can help airlines stay ahead of the competition by understanding what customers want and providing them with better services than their competitors. To achieve a better overall rating from their consumers, some businesses employ comment spam to downgrade the rankings of their competitor firms based on the categories of their items. Thus, one of the tasks to boost sentimental analysis's authenticity is to analysed these spammers patterns and identify them as genuine or fake. Thus, in this project we focus on the reviews which is given by the people on twitter for an airline company and for every individual review. We basically classify them into spam or not spam thus we will use different algorithms like Support Vector Machine,Naïve Bayes, Random Forest and choose the optimal one after comparison of each algorithm.*

Keywords: Machine Learning, Data Analysis, Support Vector Machine, Airline Service

I. INTRODUCTION

Identifying spam reviews is important because [3] not all online reviews are truthful and trustworthy. Detecting spam reviews can help businesses improve their services based on genuine feedback from customers. Spam reviews can be created by spammers who want to downgrade [4] a service, or who want to harm the reputation of their competitors. Manual analysis of reviews can also be used to detect fake reviews, but this approach is time-consuming and may not be effective for large volumes of data.

Machine learning methods involve building a [1] spam collection from crawled reviews and using algorithms to identify patterns in the data that distinguish between real and fake reviews. Sentiment analysis techniques involve identifying [2] opinions or sentiments expressed in the text of the review and using these sentiments to determine whether the review is genuine or fake. To detect fake online reviews using machine learning, one method is to up-sample minority class using synthetic data generation to increase the number of samples from the minority class in the data set [5] Another method is to extract user-behaviour features from reviews and reviewers' information such as the number of words per review, average rating per user, etc., then apply outlier tests such as z-score test or Isolation Forest method to identify spam users who post many fake reviews. However, these methods may have trouble finding fake reviews but can correctly classify real ones.

II. LITERATURE SURVEY

Twitter data serves as a good source to gather feedback tweets and perform a sentiment analysis. This approach starts off with pre-processing techniques used to [6] clean the tweets and then representing these tweets as vectors using a deep learning concept to do a phrase-level analysis. This gives data scientists and Airline companies a broader perspective about the feelings and opinions of their customers. In this paper, they go through several tweet pre-

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

processing techniques followed by the application of seven different machine learning classification algorithms that are used to determine the sentiment within the tweets.

Twitter sentiment classification about airline services. In our experiments, six individual classification approaches, and the proposed ensemble approach were all trained and tested using the same dataset of 12864 tweets, in which 10 fold evaluation is used to [7] validate the classifiers. The results show that the proposed ensemble approach outperforms these individual classifiers in this airline service Twitter dataset. Based on our observations, the ensemble approach could improve the overall accuracy in twitter sentiment classification for other services as well.

When it comes to decision-making, the Internet plays an important role all over the world. Many people use blogs, social media and other online platforms to share their thoughts and views via the internet. As a result, the internet is filled with irrelevant relevant information. So, it creates a big challenge to retrieve the desired ones from the internet [8] by scanning each document. The new, improved Adaboost technique for sentiment analysis is shown in the proposed study methodology. To find the best algorithm for the system, a variety of machine learning techniques have been used. Based on the accuracy of the algorithms and the confusion matrix, performance analysis has been carried out. These airlines fly across the same geographic region in the USA, which places them in the top spot when choosing a carrier

In this study, the LSTM network model provided a deep learning-based sentiment analysis method. They examine social media tweet data for the airline sector, which [9] exhibits improved performance in the training set. Also, the Bidirectional LSTM network can improve accuracy (Bi-LSTM)

The classification of the labelled data from six different airline firms is the goal of this study. Although the data is unbalanced and there are different [10] numbers of classes in the data from each airline company, they still wish to perform NLP stages and utilize different ML algorithms

To find word sentiments, an airline dataset is used as the primary input, and the data is then analyzed and preprocessed. In the suggested work, Nave Bayes is [11] used to take the entire tweet in order to determine the number of words in it as well as their polarity, hence resolving the unigram problem of the existing work where a single word tweet is taken. The suggested system can categorize confusing tweets and neutralize them based on their gravitas. It uses the unigram and n- gram techniques to categorize tweets as positive or negative based on their average value

III. PROPOSED WORK

3.1 Overview

This project aims to create a customer review analysis for an Airlines company that can perform sentiment analysis on customer reviews and identify spam reviews. The sentiment analysis is done using the text blob library, which is a popular natural language processing library. The text classification for identifying spam reviews is done using machine learning algorithms and natural language processing. For that we had taken the dataset of US Airline from Kaggle containing more than 10 columns and more than 10,000 rows and trained our machine learning model. After analyzing our data we found out that there is no label as spam and not spam in our dataset so we added manually and our data is also imbalanced to address imbalanced data, the SMOTE library is used to generate synthetic samples of the minority class to balance the data. This can improve the performance of the machine learning algorithms used to make predictions about whether a review is spam or not. The model evaluation is done using confusion matrix, precision score, and other evaluation metrics to measure the accuracy and effectiveness of the model. The model is then saved using the pickle library, which allows the model to be saved and loaded for later use. Flask is used as the server-side framework to handle the back-end processing and provide the data for the UI. The end goal is to provide a user-friendly, accessible, and effective dashboard for analyzing customer reviews and identifying spam reviews for the Airlines company.

3.2 Data Collection

The US Airline dataset used in this project was sourced from Kaggle, a popular platform for hosting and sharing datasets. The dataset contains reviews from various airlines operating in the United States, and includes over 10,000 rows of data with more than 10 columns of information. The Us Airline dataset is a valuable resource for this project, as it provides a large and diverse set of customer reviews for training and testing machine learning models for review analysis and spam detection. With over 10,000 reviews and more than 10 columns of information, the dataset is well-

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

suited for developing and evaluating machine learning model to accurately classify customer reviews and identify spam.

For sentiment analysis, the data was collected from Twitter using snscrape. Snscrape is a Python package that allows users to scrape data from social media platforms like Twitter. Using snscrape, we collected a large data of tweets related to a specific topic. The tweets were pre-processed to remove any URLs, hashtags, and mentions, and were tokenized to split them into individual words and show result in our dashboard.

3.3 Data Pre-processing

Data cleaning is basically the process of removing errors and anomalies or replacing observed value with true data value in order to gain greater values in analytics. In order to train a machine learning model to predict whether a given message is spam or not, we would need to have a labelled dataset where each message is already labelled as either spam or not spam, but in our dataset there is no such label so we add this label manually on the basis of this keyword like 'offer', 'free', 'discount', 'win', 'winner', 'limited', 'time', 'today', 'buy', 'purchase', 'sale', 'click', 'act', 'now', 'money', 'cash' if this keyword is present there are higher chances of spam tweets.

```
def label_spam(tweet):
    keywords = ['offer', 'free', 'discount', 'win', 'winner', 'limited', 'time', 'today', 'buy', 'purchase', 'sale', 'click', 'act', 'now', 'money', 'cash']

    for keyword in keywords:
        if keyword in tweet:
            return 'spam'

    return 'not spam'

df['label'] = df['text'].apply(label_spam)
df.head()
```

Figure 1: Adding new label spam and not spam

The process of cleaning a dataset includes removing irrelevant values as well as columns that don't add anything useful to it in order to create a beneficial result and aid in performance improvement by increasing the classifier's and the classification process's accuracy. Also, the existing raw data is high-dimensional and completely unstructured, therefore it needs to be cleaned and modified in accordance with the processing requirements in order to produce a usable dataset. In contrast to splitting the string or texts into an initial list of tokens, which is equivalent to a sub-part of a sentence, tokenizing is a procedure used to extract only the key words needed for subsequent processing. Prior to the actions that will help in classification and result-giving, it is one of the most crucial techniques for simplifying the information.

Stop words are eliminated in this step of the process. Another function of it is to remove punctuation that is ineffective for identifying if a particular tweet is positive, negative, or neutral.

Bag of words is used to transform the text data into a form that the text classification algorithm can understand. The information is transformed into a matrix, where each row describes the number of occurrences of a word in the document and each column the number of essential terms

```
vectorizer = CountVectorizer()
X = vectorizer.fit_transform(X)
print(X.toarray())
```

```
[[0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 ...
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]
 [0 0 0 ... 0 0 0]]
```

Figure 2: Bag of words

3.4 Data Sampling

We analyzed our data to see if it is balanced before moving on to the Classification. Because of the high dimensionality of our dataset and the extreme class imbalances in Spam and not spam, so classifier's accuracy will degrade. We used the package SMOTE to[12] resample our data in order to resolve this issue. Synthetic Minority Oversampling

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

Technology is referred to as SMOTE. It is one of the various strategies that can be used to address the issue of class inequality. SMOTE creates new minority instances by synthesized them with actual minority instances.

```
df['label'].value_counts()
```

```
not spam    8611
spam        2902
Name: label, dtype: int64
```

Figure 3: Sampling

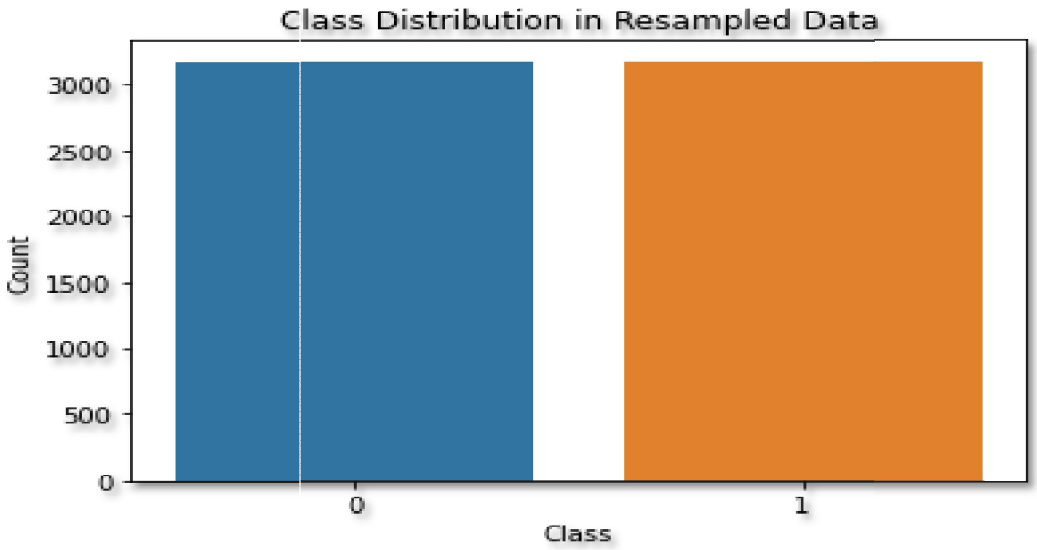

Figure 4: Balanced data

IV. CLASSIFICATION

4.1 Naïve Bayes

Naive Bayes is a classifier technique used for building classifiers: Models assigns class labels to instances, represented feature values as vectors, in which class[13] labels are extracted from some finite set. For example, the fruit which is round may be considered as an orange if its colour is orange, round, and it is about 4 in radius. Naive Bayes classifier independently considers each of these features to find the probability whether the fruit is an orange, regardless of any possible relationships between the features like roundness, colour and diameter.

4.2 Support Vector Machine

Support Vector Machine is a supervised machine learning algorithm that attempts to create a hyperplane in an n-dimensional feature space that successfully classifies all its data points. Support vectors are the datapoints that are closest to the hyperplane, and n is the number of features. The margin, or the distance between the points that are closest to the other class points, is maximized when SVM creates a hyperplane. Support vector machines are particularly effective at classifying data[14].

4.3 Random Forest Classifier

In order to obtain an even more accurate result, a sizable number of decision trees that seem to be generally uncorrelated are combined [15] While one tree is incorrect, the other trees could be correct, thus their total impact is always correct. This is an excellent accuracy. Even with no hyper tuning parameter, Random Forest typically yields the right outcome. The more uncorrelated trees there are in the model, the more accurate our results will be.

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

V. RESULTS

To evaluate the performance of the classification algorithms on the label, we calculated various performance metrics such as precision, recall, F1score, and accuracy. Additionally, we generated confusion matrices for all labels for the three classifiers including Naive Bayes, Support Vector Machine, Random Forest Classifier. The confusion matrix diagonal element show the amount of samples that were classified correctly. Based on our analysis, we found that the Random Forest algorithm outperformed the other algorithms.

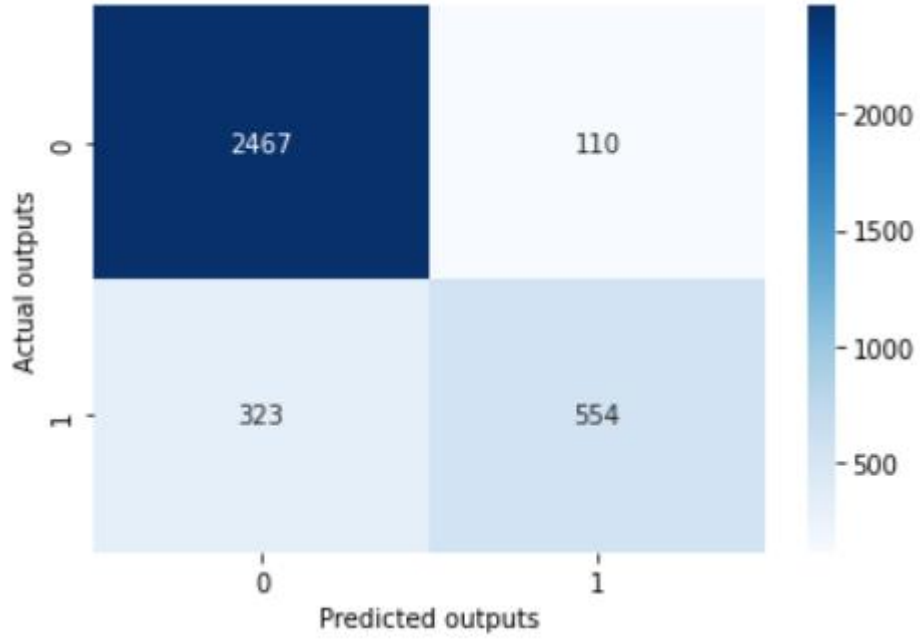

Figure 5: Naïve Bayes

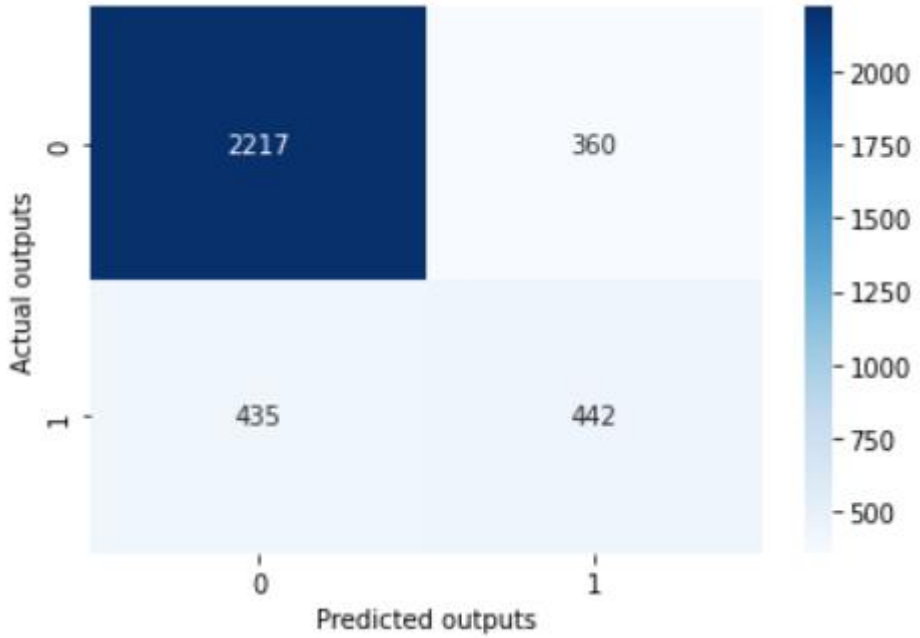

Figure 6: Support Vector Machine

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

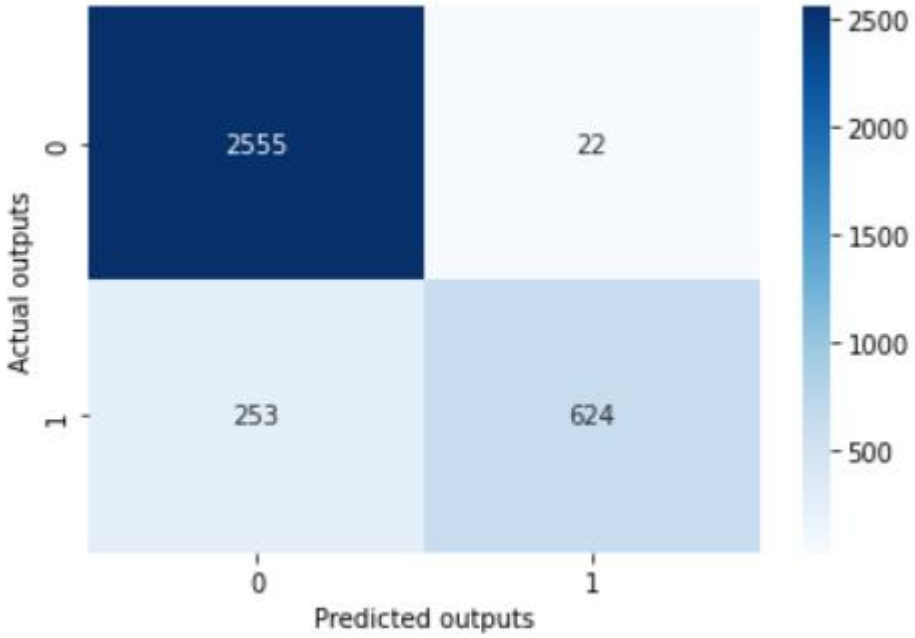

Figure 7: Random Forest Classifier

Table 1: Summary of result

Performance metrics	Naïve Bayes	Support Vector Machine	Random Forest Classifier
Accuracy	87.46%	76.98%	92%
Precision	0.88	0.84	0.91
Recall	0.96	0.86	0.99
F1 score	0.92	0.85	0.95

VI. CONCLUSION

In conclusion, the analysis of customer reviews for an airline company is a crucial task that can provide valuable insights into the customer experience and identify any potential issues that need to be addressed. In this project, we analyzed a dataset of customer reviews for an airline company and implemented various natural language processing techniques to identify spam reviews. The results of our analysis showed that a combination of machine learning algorithms and feature engineering can effectively identify spam reviews with a high degree of accuracy. One of the key findings of our analysis was that certain words and phrases, such as "scam" and "fake," were strongly correlated with spam reviews. Additionally, we found that the use of punctuation and capitalization, as well as the presence of certain special characters, also played a role in identifying spam reviews. We also identified several key areas where the airline company could improve the customer experience, including issues with flight delays and cancellations, as well as poor customer service. By addressing these issues, the airline company can improve the customer satisfaction.This experiment demonstrates the efficiency of natural language processing altogether the techniques in identifying and analyzing customer reviews. By utilizing these techniques, businesses can gain a deeper understanding of their customers' experiences and take action to improve their products and services.

REFERENCES

[1]. https://www.researchgate.net/publication/220815566_Learning_to_Identify_Review_Spam
[2]. https://research.aimultiple.com/fake-review-detection/
[3]. https://journalofbigdata.springeropen.com/articles/10.1186/s40537-015-0029-9
[4]. https://www.sciencedirect.com/science/article/pii/S0969698921003374
[5]. https://scoredata.com/how-to-detect-fake-online-reviews-using-machine-learning-2/

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
Volume 3, Issue 2, March 2023
Impact Factor: 7.301

[6]. A. Rane and A. Kumar, "Sentiment Classification System of Twitter Data for US Airline Service Analysis," 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), Tokyo, Japan, 2018, pp. 769-773, doi: 10.1109/COMPSAC.2018.00114.

[7]. Y. Wan and Q. Gao, "An Ensemble Sentiment Classification System of Twitter Data for Airline Services Analysis," 2015 IEEE International Conference on Data Mining Workshop (ICDMW), Atlantic City, NJ, USA, 2015, pp. 1318-1325, doi: 10.1109/ICDMW.2015.7.

[8]. https://www.researchgate.net/publication/341070490_Sentiment_Analysis_of_US_Airline_Twitter_Data_using_New_Adaboost_Approach.

[9]. R. Monika, S. Deivalakshmi and B. Janet, "Sentiment Analysis of US Airlines Tweets Using LSTM/RNN," 2019 IEEE 9th International Conference on Advanced Computing (IACC), Tiruchirappalli, India, 2019, pp. 92-95, doi: 10.1109/IACC48062.2019.8971592.

[10]. C. Baydogan and B. Alatas, "Detection of Customer Satisfaction on Unbalanced and Multi-Class Data Using Machine Learning Algorithms," 2019 1st International Informatics and Software Engineering Conference (UBMYK), Ankara, Turkey, 2019, pp. 1-5, doi: 10.1109/UBMYK48245.2019.8965631.

[11]. N. K. Sharma, S. Rahamatkar and S. Sharma, "Classification of Airline Tweet Using Naïve-Bayes Classifier for Sentiment Analysis," 2019 International Conference on Information Technology (ICIT), Bhubaneswar, India, 2019, pp. 70-75, doi: 10.1109/ICIT48102.2019.00019.

[12]. https://rikunert.com/smote_explained.

[13]. M. Trupthi, S. Pabboju and G. Narasimha, "Sentiment Analysis on Twitter Using Streaming API," 2017 IEEE 7th International Advance Computing Conference (IACC), Hyderabad, India, 2017, pp. 915-919, doi: 10.1109/IACC.2017.0186.

[14]. https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47

[15]. https://towardsdatascience.com/understanding-random-forest-58381e0602d2