International Journal of Latest Engineering and Management Research (IJLEMR)
ISSN: 2455-4847
www.ijlemr.com || Volume 06 - Issue 02 || February 2021 || PP. 06-11

Application of Text Mining to predict Quality of response to Consumer Complaints

Yalavarthi Bharat Chandra[1], Gouru Karthikeya Reddy[2]
[1]*Computer Science and technology, VIT university, Vellore, India*
[2]*Computer Science and technology, VIT university, Vellore, India*

Abstract: Artificial Intelligence can be used by many companies to provide better customer service and higher level of convenience. It can help to process large amount of textual data to gain useful insights. In this paper we applied text mining and machine learning to develop a binary classification model which predicts if a company can give a satisfactory and timely response for a consumer complaint or not. This model can be used to gain insights of consumer complaints and to provide improved customer service by distinguishing complaints based on their difficulty to resolve and understanding which complaints require more attention. Companies can anticipate customer problems and implement proactive measures rather than reactive ones on improving service. We used Consumer Financial Protection Bureau's (CFPB) consumer complaints on financial products dataset for implementing this work. This dataset consists of textual data which cannot be used directly to build machine learning models hence we used text mining techniques like tokenization, stemming, lemmatization and vectorization to make textual data compatible with machine learning models. We built and compared machine learning models based on their performance. We chose four predictive models namely Logistic Regression, Multinomial Naïve Bayes, Gradient Boosting and Adaptive Boosting.

Keywords: Natural Language Processing**,** Text mining**,** Machine Learning**,** Consumer Complaints

1. Introduction

Satisfied customers become devoted buyers when a business is trustworthy and provides good customer service. [1]Customer service is important to a business because it retains customers and extracts more value from them. By providing top-notch customer service, businesses recoup customer acquisition costs and cultivate a loyal following that refers customers, serves as case studies, and provides testimonials and reviews. [2]Resolving consumer complaints is an integral part of Customer service. Consumer complaints are widely spread in their difficulty to resolve, some are quite easy and some extremely hard. Companies also need to gain insights on complaints received to improve their product and the service. Allotting complaints to customer service executives by mapping their experience and difficulty to resolve a complaint can be beneficial for both company and customer. We used Text mining, Natural language processing (NLP) and machine learning to gain useful knowledge from consumer complaints. Text mining is the process of deriving useful knowledge from textual data. This can be achieved by using natural language processing techniques. Natural language processing is a subfield of linguistics, computer science, and artificial intelligence concerned with the interactions between computers and human language, how to program computers to process and analyse large amounts of natural language data. [3]In this work we employ NLP techniques and machine learning models to predict if a company can give a satisfactory and timely response for a consumer complaint or not.

We built four machine learning models using 1) Logistic Regression 2) Multinomial Naïve Bayes 3) Gradient Boosting and 4) AdaBoost algorithms to compare their performances and choose the best one.

2. Dataset Information

The dataset used is made up of real-world data on financial complaints received by various financial companies. The data is recorded and maintained by Consumer Financial Protection Bureau (CFPB) of U.S.A. The CFPB is an agency of the United States government responsible for consumer protection in the financial sector. It logs all the complaints received by all financial companies from their customers. The dataset contains 18 columns and over 9 lakh rows or observations. We removed some columns because of lack of relevance in predicting class label or because of having high correlation with other column, out of 18 columns the following columns are taken into consideration for building the model. Column 7 and 8 in the below list is used to create the class label (Satisfactory response). Satisfactory Response has two possible values namely i)Yes ii)No.

1) Product
2) Sub-product
3) Issue
4) Sub-issue
5) Consumer complaint narrative

6) Company
7) Timely response?
8) Consumer disputed?

1.1 Logistic Regression

It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all the regression analysis, the logistic regression is a type of predictive analysis. Logistic regression is used to describe the data and to explain the relationship between one of the dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. In regression analysis, logistic regression is estimating the parameters of a logistic model (a form of binary regression). [4]

1.2 Multinomial Naïve Bayes

The multinomial Naive Bayes (Multinomial NB) classifier is suitable for classification with discrete features (e.g., word counts for text classification). The multinomial distribution normally requires integer feature counts. However, in practice, fractional counts such as tf-idf may also work. [5]

1.3 AdaBoost Classifier

AdaBoost, short for Adaptive Boosting, is a machine learning meta-algorithm. It is adaptive in the sense that subsequent weak learners are tweaked in favour of those instances misclassified by previous classifiers. AdaBoost is sensitive to noisy data and outliers. In some problems it can be less susceptible to the overfitting problem than other learning algorithms. The individual learners can be weak, but as long as the performance of each one is slightly better than random guessing, the final model can be proven to converge to a strong learner.[6]

1.4 Gradient Boosting

Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function. [7]

3. Related Work

Traditional methods can only give limited insights of customer needs from complaints based on the product features. In [8] Outcome driven Innovation (ODI) is used find true needs of customers. This method can identify latent needs of customers during different phases of product use by customers which cannot be identified by previous methods. The tool implemented in this paper clusters various complaints received from which customer needs can be derived. The knowledge derived can provide advanced information. In [9] text mining was used to extract useful knowledge from online reviews on trip adviser by satisfied and dissatisfied customers . NLP techniques like tokenization, stemming and filtering are used to understand the factors that leads to customer satisfaction and those that lead to customers recommending the service to others. In [10] the methods of NLP and machine learning are employed to understand how hotel industry can leverage information gained from online reviews of hotels. Useful knowledge is extracted by automated evaluation of quality, topic extraction and sentiment analysis. In [11] Several Fundamental text mining techniques are described along with various text pre-processing techniques. Text mining approaches like Information retrieval, Natural language processing, Information extraction from text, Text summarization, supervised learning methods are discussed. Text pre-processing tasks like tokenization, filtering, lemmatization and stemming are described in detail. Machine learning models both supervised and unsupervised like Naïve Bayes, Decision Trees-means and K-nearest neighbours along with their implementation in text mining is shown.

4. Methodology

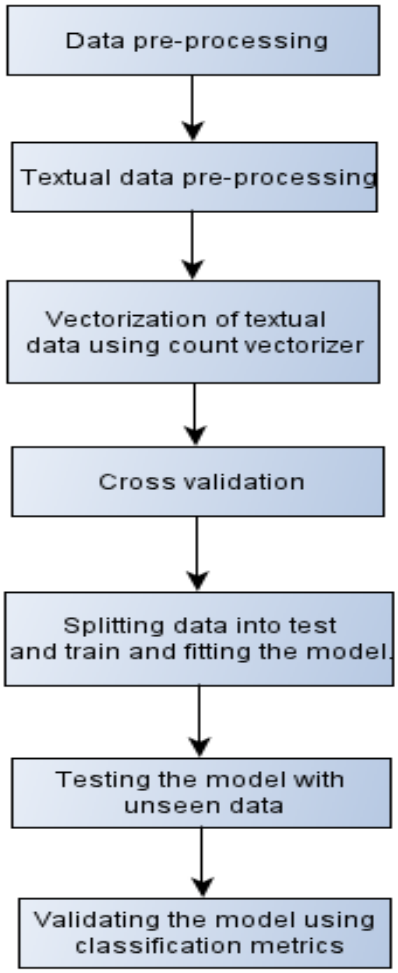

Figure 1: Implementation Methodology

4.1 Data Pre-processing

The textual data is converted into feature vectors which can be used by various machine learning models to train. But null values cannot be converted into feature vectors hence it is important to deal with null values properly.

The ratio of positive to negative samples is 97%:2.9%. Hence this dataset has large difference between count of positive and negative samples. Therefore, this is an imbalanced data. If there are two classes, then the balanced data would mean 50% points for each of the classes. For most machine learning models slight imbalance would not cause any drop in performance. But if ratio of classes cross 90:10 then the dataset needs to be modified and made more balanced to avoid significant performance degradation.

4.2 Handling Imbalanced data

Resampling of the dataset is particularly important for dataset to have a balance between positive and negative samples in binary classification. If an imbalanced data is used, then classifier can totally ignore minority class for the majority one. We need to change the dataset by resampling it to have more balance between positive and negative samples.

We used two methods to resample the dataset they are a) Up sampling b)Down sampling. Adding instances of minority by randomly duplicating them is called Up sampling and randomly deleting instances of majority class is called down sampling.

For the consumer complaints dataset after up sampling of minority class value which is [Satisfied response? =0] and down sampling majority class [Satisfied response? =1] we changed the imbalanced dataset into a much more balanced dataset. The ratio of class values are changed from (0.97-0.029) to (0.64-0.35)

4.3 Textual data pre-processing

Textual data pre-processing techniques like stemming, lemmatization and stop word removal are applied to the dataset. These techniques are important to remove unwanted words, prefixes, suffixes and make the textual data suitable to be converted to vectors. We used NLTK library which has default set of stop words for English language to remove stop words before vectorization.

With stemming, words are reduced to their word stems. A word stem need not be the same root as a dictionary-based morphological root, it just is an equal to or smaller form of the word. Lemmatisation in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.[12].Stemming and lemmatization helps to remove unwanted prefixes and suffixes of a word and make the words suitable for converting them to vectors. Some irrelevant words that appear in the dataset like "XXXX","XX","N/A", and "Closed with explanation ","- are added in the stop words. The vectorizer ignores these stop words while converting the text into sparse matrices. NLTK library has default set of stop words for English language are removed automatically before vectorization.

4.4 Vectorization

Textual data in its raw format cannot be directly fed to machine learning algorithms, because these algorithms work on numerical data. So to use textual data in these models we need to convert the text into feature vectors which holds the internal dependencies in the textual data. We used count vectorizer to convert the textual data present in the dataset to vector form. Count is a function used to convert a collection of text documents to a matrix of token counts. The resultant data after applying count vectorizer is compatible to use in machine learning models.

4.5 Model Building

Feature vectors that are obtained from applying count vectorizer are used to fit and train the models. Data is split between train and test in the ratio of 80:20 respectively. Cross validation is used to test the models ability to predict class label of new unseen data. This helps to identify any problems like over fitting or selection biasand to give an insight on how the model will generalize to an independent dataset.

We now use a machine learning classifier to predict the class label(Satisfactory Response).Four classifiers namely Logistic Regression, Multinomial Naïve Bayes, Gradient boosting, Adaboost are built and compared based on performance and the best model is chosen.

5. Results

The performance of the models that are built needs to evaluated to choose the best one. We used Accuracy, Precision,Recall,F1-score,Area under ROC curve(AUC) as performance metrics. We have provided mathematical expression of all the performance metrics mentioned where TP is True positive, TN is True negative, FP is False positive, FN is False negative.

Accuracy=(TP+TN)/(TP+FP+FN+TN)

Precision=TP/(TP+FP)

Recall=TP/(TP+FN)

F1-score = 2 ∗ (precision ∗ recall) / (precision + recall)

ROC curve(AUC):A receiver operating characteristic curve, or ROC curve, is a graphical plot that illustrates the diagnostic ability of a binary classifier system as its discrimination threshold is varied.The ROC curve is created by plotting the true positive rate (TPR) against the false positive rate (FPR) at various threshold settings. [13]

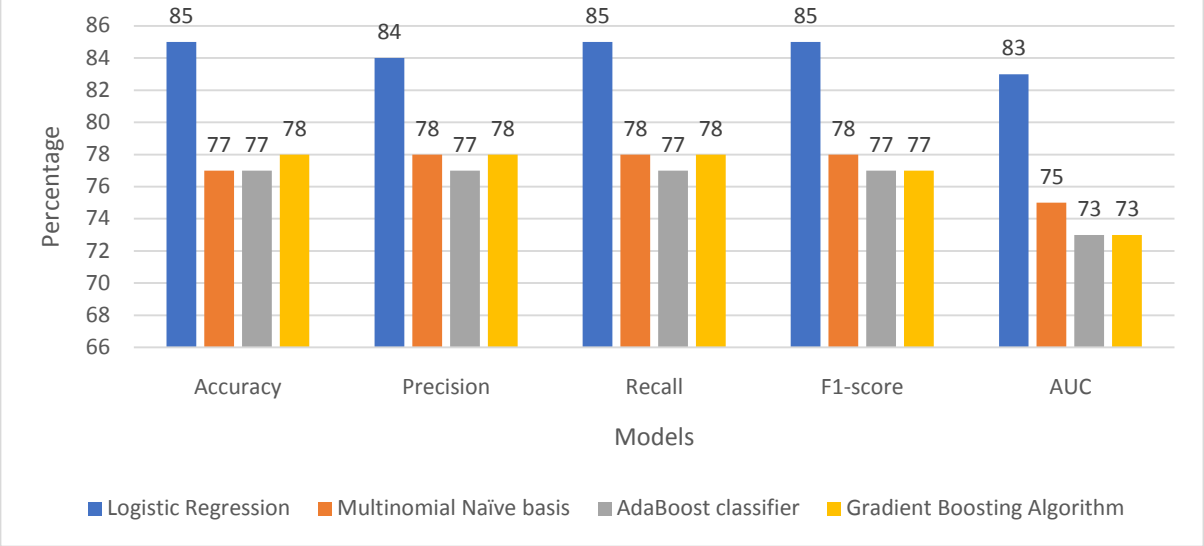

Figure 2: Performance Comparison of models

Comparison of performance of different models based on different performance metrics is shown in Figure 2.

We can infer that Logistic Regression has best performance in terms of all the metrics among all the models tested. Logistic regression model has an accuracy of 85.18% in predicting test data and performs significantly better than other models tested by almost 7% in accuracy. We were able to predict if a satisfactory response can be given to a customer complaint with an high accuracy(85.18%).The model predicts both true positives and true negatives with no significant performance degradation even though the data is highly skewed in favour of true positives. This was achieved due to data balancing techniques which reduced the difference between positive and negative samples.

6. Conclusion

In this paper we applied text mining to a consumer complaints dataset and developed a machine learning model to predict if a company can give a good and timely response for a complaint. As the dataset consists of textual data, to be able to use a machine learning classifier, we had to represent it in the form of feature vectors using vectorization. We have applied Logistic Regression, Multinomial naïve Bayes, AdaBoost and Gradient boosting classifiers of which logistic regression performed the best with an accuracy of 85.18%. Text analytics on consumer complaints gives an insight into various factors involved in handling complaints for a company. The developed model can help companies to distinguish complaints based on their difficulty to resolve which results in efficient management of resources and good customer service. This model can be easily ported to other similar consumer complaints datasets and be used for prediction.

7. References

[1]. Ameritas [online] https://www.ameritasinsight.com/employee-benefits/industry-buzz/why-good-customer-service-is-important, [Accessed on 20th August 2020].

[2]. Swetha Amaresan [online] https://blog.hubspot.com/service/importance-customer-service, [Accessed on 23rd August 2020].

[3]. Wikipedia [online] https://en.wikipedia.org/wiki/Natural_language_processing, [Accessed on 24th August 2020].

[4]. Static Solutions [online] https://www.statisticssolutions.com/what-is-logistic-regression/, [Accessed on 24th August 2020]

[5]. Scikitlearn,[online]https://scikitlearn.org/stable/modules/generated/sklearn.naive_bayes.MultinomialNB.html, [Accessed on 27th August 2020]

[6]. Tejashwini S.G "Fraud Detection in Examination using LBP method" in IJLEMR, Volume 02 - Issue 04 April 2017 PP. 28-35.

[7]. Wikipedia [Online] https://en.wikipedia.org/wiki/Gradient_boosting , [Accessed on 28th August 2020].

[8]. Junegak Joung , Kiwook Jung , Sanghyun Ko and Kwangsoo Kim "Customer Complaints Analysis Using Text Mining and Outcome-Driven Innovation Method for Market-Oriented Product Development" ,in MDPI published on 21st December 2018

[9]. Katerina Berezina ,Anil Bilgihan , Cihan Cobanoglu , Fevzi Okumus "Understanding Satisfied and Dissatisfied Hotel Customers: Text Mining of Online Hotel Reviews", in Journal of Hospitality Marketing and Management vol 25,2016.

[10]. Shawn Mankad, Hyunjeong "Spring" Han, Joel Goh, Srinagesh Gavirneni "Understanding Online Hotel Reviews Through Automated Text Analysis", in The Institute for Operations Research and the Management Sciences vol 8, 5th May 2016.

[11]. Mehdi Allahyari, Seyedamin Pouriyeh, Mehdi Assef, Saied Safaei, Elizabeth D. Trippe, Juan B. Gutierrez, Krys Kochut "A Brief Survey of Text Mining: Classification, Clustering and Extraction Techniques" in arXiv vol1, 10th July 2017.

[12]. Towards Data Science [online], https://towardsdatascience.com/stemming-lemmatizationwhatba782b7c0bd8?gi=7af70d9ec22f#:~:text=With%20stemming%2C%20words%20are%20reduced,off%20the%20ends%20of%20words, [Accessed on 1st September 2020].

[13]. Wikipedia [online], https://en.wikipedia.org/wiki/Receiver_operating_characteristic, [Accessed on 2nd September 2020].