www.ijecs.in
International Journal Of Engineering And Computer Science
Volume 12 Issue 09 September 2023, Page No. 25801-25808
ISSN: 2319-7242 DOI: 10.18535/ijecs/v12i09.4757

Analyzing Customer Satisfaction Level on Telecommunication Usage Using Big Data

J. Palimote[1], C. Aloy-Okwelle[2], O. T. Olise [3]
Department of Computer Science, Rivers State University, Port Harcourt, Nigeria[1,2,3]

Abstract:
Telecom network often encounters large number of tweets based on the user experience for a network. This huge amount of raw data can be used for industrial or business purpose by organizing them according to our requirement and processing. The aim of this paper is to address the social media review challenges in telecom companies. The methods include; extracting tweets, analyzing them and segregating them into various categories to help the company understand the concerns of their customer. This can help save millions and prevent customer churn. In other to build a robust model, the dataset was pre-processed by checking and removing Nan values. After the pre-processing, stage the cleaned data was tokenized. In tokenization process, each word was divided into tokens, for easy training. After the tokenization process, we performed an exploratory data analysis on the dataset to understand the pattern of the dataset. After the explorative data analysis stage, we trained a random forest classifier to predict/classify the customer's satisfaction into positive, negative, and neural.

Keywords: Big Data, Telecommunication, Machine Learning, Random Forest classifier

1. Introduction

Big data can be used to analyze customer satisfaction levels in the telecommunications industry by collecting and analyzing large amounts of data from various sources such as customer surveys, call center logs, and social media. This data can be used to identify patterns and trends in customer usage, preferences, and satisfaction. Machine learning algorithms can also be applied to this data to predict customer behavior and identify potential issues or areas for improvement. Big data can be used to track and monitor the performance of telecommunications networks and identify potential bottlenecks or areas for optimization. In order to analyze customer satisfaction levels on telecommunication usage using big data, its essential to gather a large amount of data on customer usage and behavior. This can include data on calls made, text messages sent, internet usage, and any other relevant information. Once this data has been collected, various big data tools and techniques, such as machine learning, is used to analyze and understand patterns and trends in the data. This can help identify areas where customers are experiencing issues or dissatisfaction, as well as areas where the service is performing well. Additionally, this data can be used to develop and implement strategies for improving customer satisfaction and addressing any issues that are identified.

2. Related Work

[5] Provided a methodology for telecom companies to target different-value customers by appropriate offers and services. This methodology was implemented and tested using a dataset that contains about 127 million records for training and testing supplied by Syriatel Corporation. Firstly, customers were segmented based on the new approach (Time- frequency- monetary) TFM (TFM where: Time (T): total of calls duration and

Internet sessions in a certain period of time. Frequency (F): use services frequently within a certain period. Monetary (M): The money spent during a certain period.) and the level of loyalty was defined for each segment or group. Secondly, the loyalty level descriptors were taken as categories, choosing the best behavioral features for customers, their demographic information such as age, gender, and the services they share. Thirdly, several classification algorithms were applied based on the descriptor and the chosen features to build different predictive models that were used to classify new users by loyalty. Finally, those models were evaluated based on several criteria and derive the rules of loyalty prediction. After that by analyzing these rules, the loyalty reasons at each level were discovered to target them the most appropriate offers and services.
[4] Developed their own predictive models. The "Logistic regression model Design". The model used customer data to predict customer retention in a telecommunications company. The model predicts customer retention based on billing, value-added services, and SMS service issues with. The system has a 95.5% accuracy.
[1] Focused on using machine learning to determine the value of customers in the hospitality sectors. The engaged customers by introducing the dynamic loyalty program. Their results show
that automated learning processes performed well in identifying the customers with greater value in specific promotions. There work contributed in deepening the practical and theoretical understanding of automated learning in the value chain of customer loyalty, in a structure that uses a dynamic model for customer engagement.
[2] Provides a retrospect on how telecom operators have been striving, before the era of big data,
in analyzing large volumes of data in business operation. Their work examined the driving forces of big data analytics in the telecom domain. They explored the benefits and the associated challenges.
[4] Explored ways of integrating big data insights with automated and assisted processes related to key customer touch points to ultimately improve the customer experience. They provided an insight on how the Alcatel-Lucent and Bell Labs helps the CSPs improve their business performance, using a unique methodology. This can be used to improve the Net Promoter Score (NPS) and a higher customer value.

3. DESIGN METHODOLOGY

This session discusses the technique used and the methods used in the data collection. The system architecture of the proposed system can be seen in Figure 1.

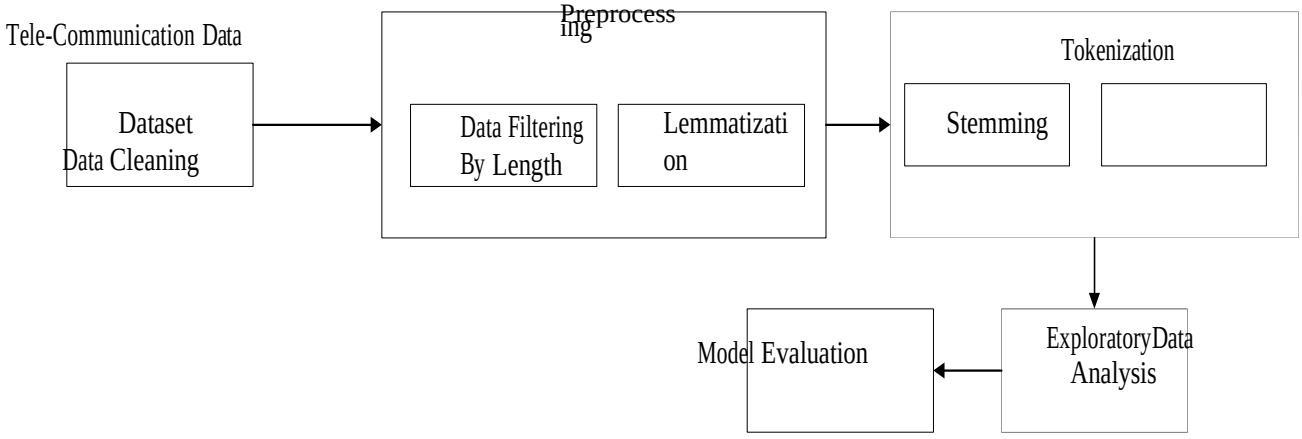

Figure 1. Architecture design of the proposed system

Dataset (Tweets): The dataset contains thousands of tweets from different customers concerning their satisfactions towards Telecommunication Company. The dataset is an unstructured one. It was collected via web scrapping using beautiful soup technique.

Pre-processing: This involves the processing of the legal case document, so that it would be fit for a training a deep learning model. The following are the process of tokenization:

i. **Data Cleaning and Noise Removal:** One of the key steps in processing language data is to remove noise so that the machine can easily detect the patterns in the data. Text data contains a lot of noise, in the form of hashtags, punctuation and numbers.
ii. **Filtering by length:** It is useful to remove unwanted words in a sentence. Words that are usually less than two characters in length do not represent a special meaning in a sentence. However, these words have characteristics that are not defeated in the previous pretreatment process. Therefore, in this process, only the necessary words are split by limiting the length of the words
iii. **Transform cases:** In this process, each character in a word is converted to lowercase.

Tokenization: Tokenization is the process of splitting the text into smaller pieces called tokens. Words, numbers and punctuation marks and others can be considered as tokens. Each sentence in the document was separated by words. To achieve this, the nltk (Natural Language Tool Kit) package in Python programming language was used.

i. **Stemming and Lemmatization:** The aim of stemming is to inflectional forms to a common base form. For grammatical reasons, text documents are going to use different forms of a word such as take, taken, taking.

Exploratory Data Analysis (EDA): This was used in performing visualization of the data. This was used in analyzing the data using charts and graphs.

Model Evaluation: The Random Forest Regression algorithm is used in order that to find the output from the training dataset, it uses multiple decision trees. The mathematical representation of the random forest can be seen in Table 1.

Table 1. Mathematical Representation of Random Forest Model

Impurity	Task	Formula	Description
Gini Impurity	Classifications	$\sum_{i=1}^{C} f_i(1-f_i)$	f_i is the frequency of label I at a node and C is the number of unique labels
Entropy	Classifications	$\sum_{i=1}^{C} -f_i \log(f_i)$	f_i is the frequency of label I at a node and C is the number of unique labels
Variance/Mean Square Error	Regressions	$\frac{1}{N}\sum_{i=1}^{N}(y_i - \mu)^2$	y_i is the label for an instance, N is the number of instances and μ is the mean
Variance/Mean Absolute Error	Regressions	$\frac{1}{N}\sum_{i=1}^{N}\lvert y_i - \mu \rvert$	y_i is the label for an instance, N is the number of instances and μ is the mean

4. Implementation/Result

This system presents a model for analyzing customers' behavior on telecommunication using big data. The system starts by acquiring unstructured data, to address the social media review challenges for telecommunication companies and then analyzing the data, and segregating them into various categories to help the company understand the concerns of their customers. In other to build a robust model, the tweet dataset was pre-processed by checking and removing NaN values. The cleaned data can be seen in Figure 2. After the stage of pre-processing, the cleaned data was tokenized. By tokenization, we divided each word into tokens, for easy training and understanding, this is seen in Figure 2. After the tokenization process, we performed an exploratory data analysis on the dataset in order to understand the dataset's pattern. This is displayed in Figures 3, 4, and 5. After the explorative data analysis, we trained a random forest classifier to predict/classify the customer's satisfaction into positive, negative, and neural. The result of the random Forest model can be found in Figures 6 and 7.

	_c0	Category	Label	Text	words	words_nostopwords
0	0	Poor service	4	vzwsupport give me a working phone without hav...	[vzwsupport, give, me, a, working, phone, with...	[give, work, without, jump, hurdle, customer, ...
1	1	Poor service	4	verizon my daughter and i both have verizon an...	[verizon, my, daughter, and, i, both, have, ve...	[daughter, keep, fail, talk, , suggestions, ri...
2	2	Poor service	4	verizon customer service is the worst i dread...	[verizon, , customer, service, is, the, worst,...	[, customer, service, bad, dread, ever, contac...
3	3	Happy Customer	2	i love having verizon i get service just about...	[i, love, having, verizon, i, get, service, ju...	[love, get, service, anywhere, vzwnow]

Figure 2. Cleaned Data

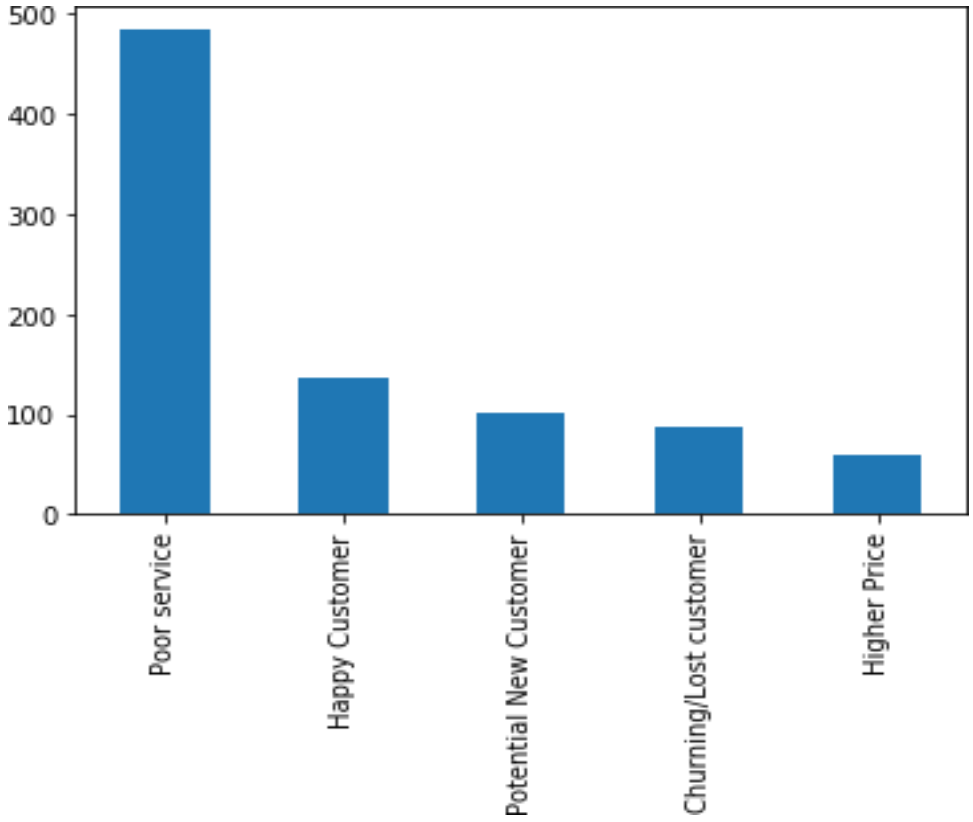

Figure 3. Histogram of the customer's satisfactory

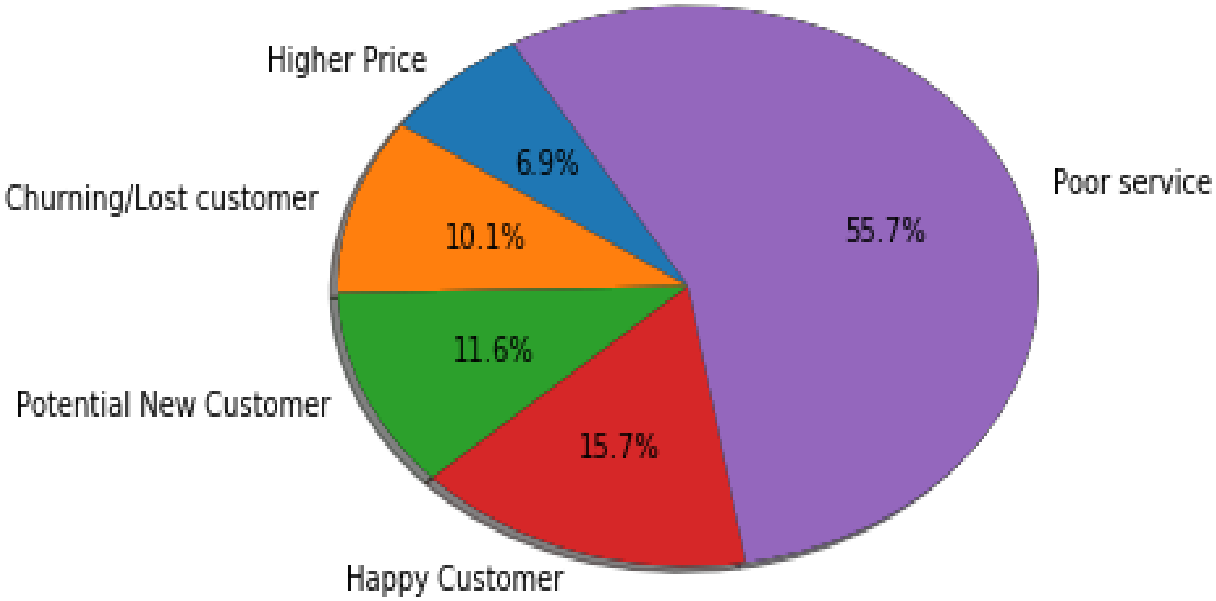

Figure 4. Pie Chart of Customer's reaction towards Tele communication

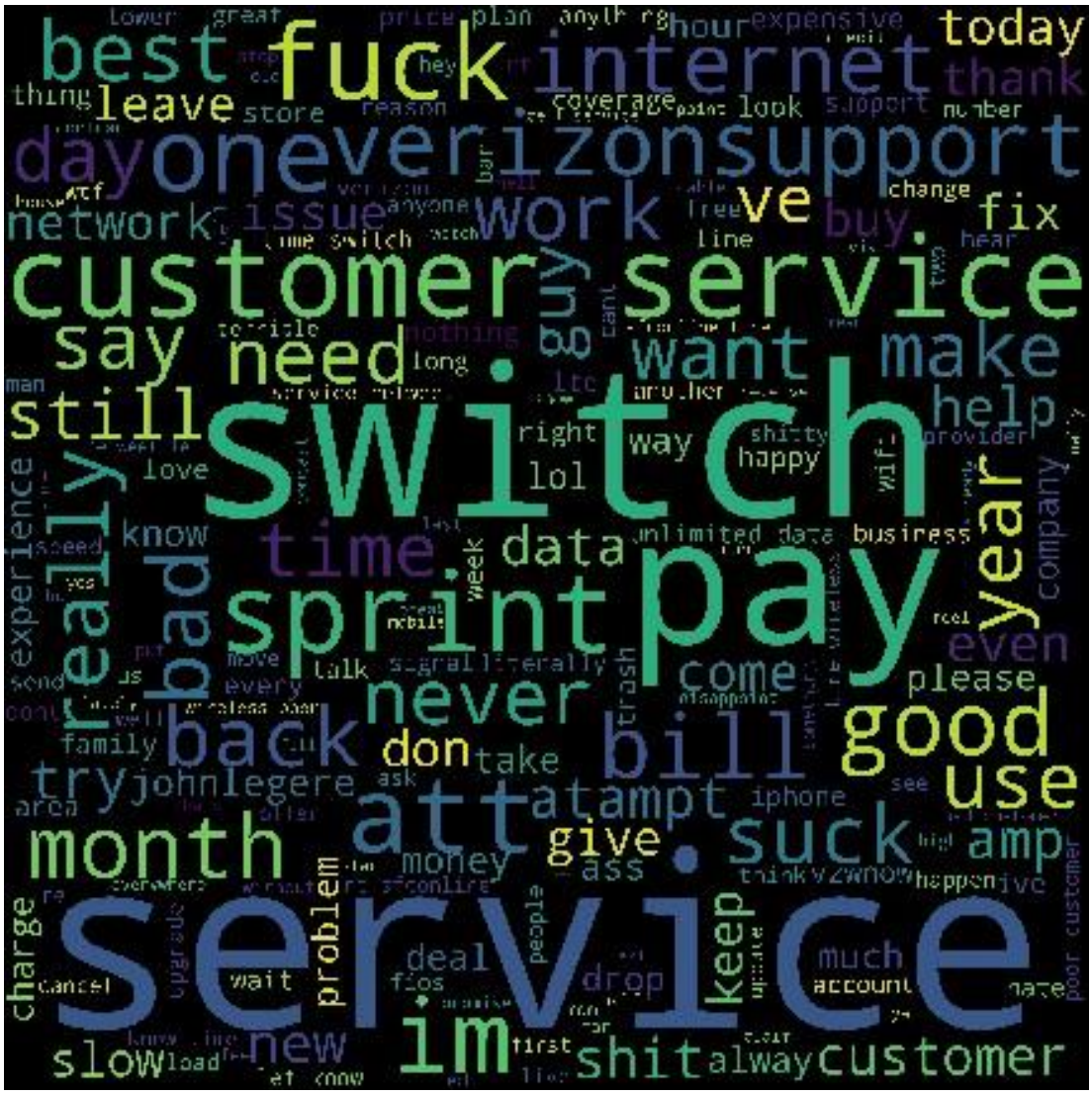

Figure 5. Word Count

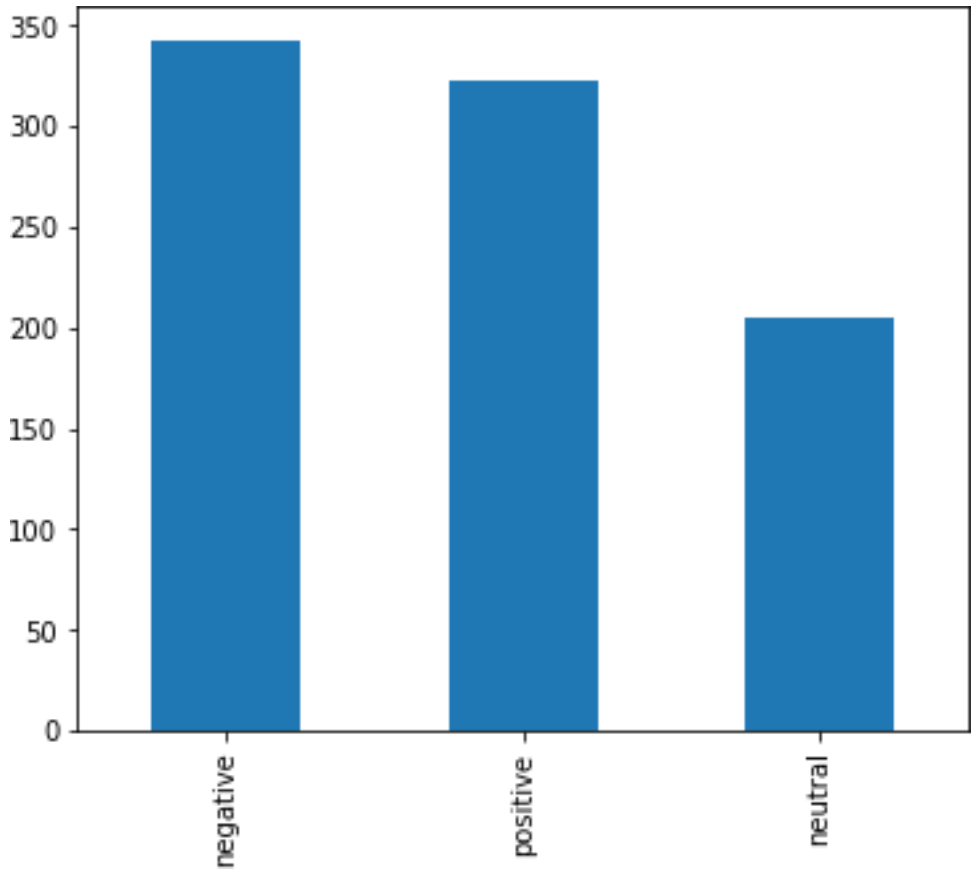

Figure 6: Histogram of the total comments made by customers

label	features	rawPrediction	probability	prediction
1.0	(11000,[111,1516,...	[0.0,1.9654302924...	[0.0,0.0982715146...	4.0
1.0	(11000,[139,157,1...	[0.0,1.9654302924...	[0.0,0.0982715146...	4.0
1.0	(11000,[139,157,1...	[0.0,1.9654302924...	[0.0,0.0982715146...	4.0
1.0	(11000,[139,157,1...	[0.0,1.9654302924...	[0.0,0.0982715146...	4.0
1.0	(11000,[222,520,8...	[0.0,2.4916510573...	[0.0,0.1245825528...	4.0

only showing top 5 rows

Figure 7. Predicted Features of the Random Forest

5. Conclusion and Recommendation

This paper presents a model for the prediction and analysis of customer satisfaction in telecommunication. In this paper, we examined a large amount of Verizon page tweets extracted from Twitter for a period of 4 days using the hashtag #Verizon, classifying the tweets into different labels based on the user experience for the network. We also analyzed the tweets based on their categories. Visualization of the dataset was carried out to give a graphical representation of the user experience. The random forest algorithm was used for classification. We can further recommend the following.

1. From our analysis we can say that Verizon must focus on their service (which includes customer service and Network issues). By focusing on these issues they can improve their customer satisfaction and avoid churn (which is also visible in the plots)

2. Using real time tweets to analyze the problem with the Verizon network and can provide a quick solution to the problem and avoid network traffic. This can provide quality service to the customer.

3. Location based data can be used to offer good deals to customers and design campaigns to combat churn.

Reference

1. Aluri A, Price BS, McIntyre N. (2019). "Using machine learning to concrete value through dynamic customer engagement in a brand loyalty program". J Hosp Tour Res. 2019;43 (1):78–100.
2. Chung-Min Chen(2016) "Use cases and challenges in telecom big data analytics"Published online by Cambridge University Press.APSIPA Transactions on Signal and Information Processing , Volume 5 , 2016 , e19.DOI: https://doi.org/10.1017/ATSIP.2016.20
3. Jeffrey S., Yves T'J., Raluca D., Peter S., Laurent P. (2014). "Using big data to improve customer experience and business performance" Bell labs technical journal 18 (4), 3-17, 2014
4. Oladapo K, Omotosho O, Adeduro O. (2018). Predictive analytics for increased loyalty and customer retention in telecommunication industry. Int J Comput Appl. 2018;975:8887
5. Wissam N, Ramez A,, Kamal S, & Shadi B.(2020)"Predictive analytics using big data for increased customer loyalty: Syriatel Telecom Company case study" J Big Data (2020) 7:29 https://doi.org/10.1186/s40537-020-00290-0 pp 2-24