CUSTOMER SATISFACTION – ISSUES AND ANALYSIS: A COMPARATIVE STUDY IN TELECOMMUNICATION INDUSTRY

PROJECT TASK TABLE

Sr. No.	Task	Assignee
1	Data Set Collected	Equal contribution by each Team Member Signature Ankita Yadav Oshin Mohabe
2	Data Preprocessing	
3	Exploratory Data Analysis	
4	Text Analytics (Sentiment Analysis) – Algorithm 1	
5	Text Analytics (Sentiment Analysis) – Algorithm 2	
6	Performance Evaluation	
7	Analysis of Experiment Results	
8	Conclusion and Future Work	
9	Presentation	
10	Research Report Writing	

INTRODUCTION

Due to COVID – 19 pandemic, every aspect of the business was disrupted. However, one piece of technology that kept everyone connected is the telecom services. As per the report by Deloitte, network usage has been increased drastically during the COVID-19 as there has been an increase in the remote workforce, and the purpose of using data has also risen. The launch of 5G services has taken place in the United States. Since 2020, the need to use data has expanded by and large, so it becomes essential for telecom companies to understand what their customers need and how to keep them fulfilled.

As per Statista, the top two wireless carrier/operator subscriber share in the U.S. 2020 were AT&T and Verizon, hence we considered these two companies for our analysis. AT&T is one of the oldest companies in the telephone business and has a market value estimated at approximately $209 billion. Verizon provides wireless and wireline services in addition to broadband and information services and has a current market capitalization value of approximately $236 billion.

Wireless digital technology is becoming the primary form of communication. Understanding the issue being faced by the customer and working on it will become a significant aspect to reduce customer churn and improve the overall experience of the customer with the data service provider.

MOTIVATION AND RESEARCH QUESTION

The use of the internet has increased to a great extent especially during the time of COVID-19 when classes are conducted online and work from home culture has increased tremendously. Hence, companies need to understand what features and services are important for the customers. This will lead to improved service offerings, better customer service, and retention of old customers as well as attracting new customers.

We decided to analyze the reviews of AT&T and Verizon since they hold the highest market share in the United States. Analyzing the reviews of the customers shall help to understand what benefits are offered by these companies and what are the drawbacks of them. Once we understand this, we shall be able to highlight the competitive advantage these companies have over others.

The research shall enable the telecom companies to understand where they are lacking and therefore are unable to grab the top spot. Through our research, we shall provide a set of attributes that makes a company stand out from others and highlight key issues faced by the customers that need to be worked upon.

BACKGROUND AND RELATED WORK

DATA COLLECTION PROCESS

We have done a comparative analysis of two telecom companies namely, AT&T and Verizon. We have collected data from Consumer Affairs (https://www.consumeraffairs.com/). Consumer Affairs provides verified reviews of customers of AT&T and Verizon. We collected a total of 13,749 reviews through web scraping. The total of AT&T reviews is 6838 and Verizon reviews is 6911.

PREPROCESSING PROCESS

Through data collection process we collected raw data and through Data Preprocessing we transformed the raw data into an understandable format to do effective and efficient data analysis.

The dataset collected through web scraping was incomplete, inconsistent, and contained few errors. Through data preprocessing we have tried to resolve these issues. Following are the steps undertaken for cleaning the data:

- Converted the Date "object" datatype to "datetime64" datatype.
- Handled missing values by replacing the missing value by a constant.
- Used pandas profiling to get detailed information about the variables.
- Handled duplicated data by deleting the duplicate rows.
- Handled invalid values by deleting the invalid values. The dataset was divided into training and testing dataset. The training dataset of AT&T contains 4587 rows and 4 columns, and the testing dataset contains 38 rows and 4 columns. The training dataset of Verizon consists of 4445 rows and 4 columns and testing dataset consists of 103 rows and 4 columns.

EXPLORATORY DATA ANALYSIS

Exploratory data analysis helps in investigating and summarizing datasets. It also helps in determining the appropriateness of the statistical analysis method being considered for data analysis. Verizon has greater mean and standard deviation when compared with AT&T in terms of rating received over a span of six years. The length of reviews written were analyzed and customers who did not gave any rating wrote the longest review and the customer who rated the telecom companies as 4 wrote the shortest review. The content of review was normalized using the lemmatization technique. The tokens were created, punctuations and stop words were removed from the list of tokens and then collocations of part of speech tagging as adjective-noun and noun-noun were discovered from the content which helped us to find the positive and negative collocations about AT&T and Verizon. Naïve based sentiment analysis was done through which we can conclude that Verizon performed better than AT&T from the year 2015 to the year 2020. After getting labeled data through naïve based sentiment analysis, "Label" and "Rating" variables were compared and we got to know a positive correlation existed among them. Feature

engineering using CountVectorizer technique enabled us to find the positive features of AT&T and Verizon.

METHODOLOGY

We have done Supervised sentiment analysis using word vectors to analyze the reviews as positive or negative. Supervised learning based sentiment analysis comprises of two steps. Step one is letting the machine learn also known as training and step two is testing. In the testing phase, the trained classifier tests the model by predicting the target class of unseen test data to assess the model accuracy.

For supervised learning, labeled dataset is required. Hence, manual labeling of data has been done. The review with the rating 1 and 2 is considered as negative, the review with the rating 4 and 5 are considered positive and remaining reviews with no rating and rating as 3 are manually examined and labeled as either positive or negative. The entire test dataset reviews were labeled manually for both AT&T and Verizon.

There are two machine learning model being used for training the classifier namely, Support Vector Machine and Naïve Bayes.

A Support Vector Machine (SVM) is a supervised machine learning model that uses classification algorithms for two-group classification problems. After giving an SVM model sets of labeled training data for each category, they are able to categorize new text. The objective of the support vector machine algorithm is to find a hyperplane in an N-dimensional space (N — the number of features) that distinctly classifies the data points. Support vector machine is highly preferred as it produces significant accuracy with less computation power.

Naïve Bayes algorithm is a supervised machine learning algorithm, which is based on Bayes theorem with an assumption of independence among predictors and used for solving classification problems. A Multinomial Naïve Bayes classification algorithm tends to be a baseline solution for sentiment analysis task. The basic idea of Naive Bayes technique is to find the probabilities of classes assigned to texts by using the joint probabilities of words and classes. Naïve Bayes Classifier is one of the simple and most effective Classification algorithms which helps in building the fast machine learning models that can make quick predictions.

Performance Evaluation is a process of evaluation the performance of the algorithm. It is done with the help of a confusion matrix and classification report.

A Confusion Matrix is a table that is often used to describe the performance of a classification model on a set of test data for which the true values are known. Its specific table layout allows visualization of the performance of an algorithm.

A Classification Report is used to measure the quality of predictions from a classification algorithm. The report shows the main classification metrics precision, recall and

f1-score on a per-class basis. The metrics are calculated by using true and false positives, true and false negatives.

AT&T

Support Vector Machine

label	0
sentiment	
0	38

	precision	recall	f1-score	support
0	1.00	1.00	1.00	38
accuracy			1.00	38
macro avg	1.00	1.00	1.00	38
weighted avg	1.00	1.00	1.00	38

As we can see from the above table, the confusion matrix shows that there are 38 negative reviews, and all are being correctly classified through support vector machine model. The classification report shows that the accuracy, macro average, and weighted average is 1.00.

Naïve Bayes

label	0
sentiment	
0	33
1	5

	precision	recall	f1-score	support
0	1.00	0.87	0.93	38
1	0.00	0.00	0.00	0
accuracy			0.87	38
macro avg	0.50	0.43	0.46	38
weighted avg	1.00	0.87	0.93	38

As we can see from the above table, the confusion matrix shows that there are 38 negative reviews, and 33 are correctly classified while 5 are incorrectly classified through naïve bayes model. The classification report shows that the accuracy is 0.87, the macro average is 0.46, and weighted average is 0.93.

VERIZON

Support Vector Machine

label	0	1
sentiment		
0	98	2
1	3	0

	precision	recall	f1-score	support
0	0.98	0.97	0.98	101
1	0.00	0.00	0.00	2
accuracy			0.95	103
macro avg	0.49	0.49	0.49	103
weighted avg	0.96	0.95	0.96	103

As we can see from the above table, the confusion matrix shows that there are 101 negative reviews and 2 positive reviews. 98 negative reviews are correctly identified by support vector machine model and rest of the reviews are incorrectly identified. The classification report shows that the accuracy is 0.95, macro average is 0.49, and weighted average is 0.96.

Naïve Bayes

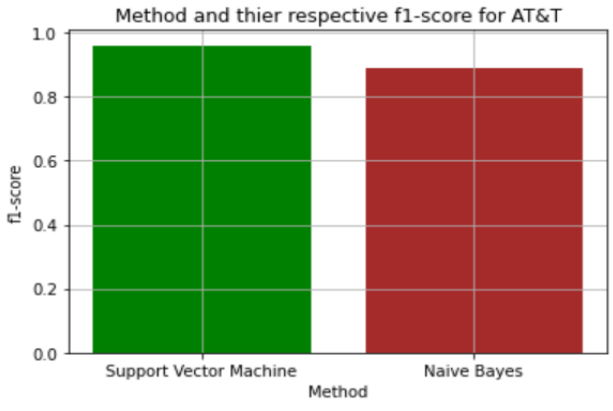
label	0	1
sentiment		
0	85	2
1	16	0

	precision	recall	f1-score	support
0	0.98	0.84	0.90	101
1	0.00	0.00	0.00	2
accuracy			0.83	103
macro avg	0.49	0.42	0.45	103
weighted avg	0.96	0.83	0.89	103

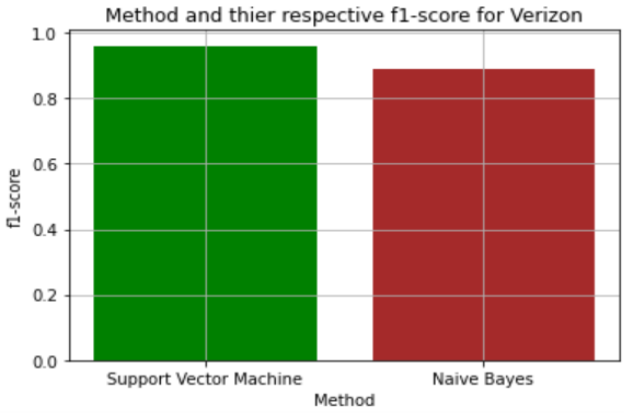
As we can see from the above table, the confusion matrix shows that there are 101 negative reviews and 2 positive reviews. 85 negative reviews are correctly identified by naïve bayes model and rest of the reviews are incorrectly identified. The classification report shows that the accuracy is 0.83, macro average is 0.45, and weighted average is 0.89.

MODEL COMPARISON

By comparing the weighted average of Support Vector Machine and Naïve Bayes model we get the below result:



AT&T



Verizon

The model support vector machine performed better than naïve bayes while analyzing both the companies AT&T and Verizon. The accuracy and weighted average are higher through support vector machine model when compared to the accuracy and weighted average through naïve bayes model. The f1-score weighted average and accuracy is 1.00, 1.00, 0.96, and 0.95 in the case of AT&T and Verizon through support vector machine while the f1-score weighted average and

accuracy is 0.93, 0.87, 0.89, and 0.83 in the case of AT&T and Verizon through naïve bayes. Hence, we can conclude that support vector machine model has better performance when compared to naïve bayes model.

ANALYSIS OF EXPERIMENT RESULTS

Part of methodology that worked/did not worked and why?

The dataset was imbalanced. The negative reviews were at a greater number than the positive reviews. In total there were 7685 negative reviews and 1347 positive reviews. Therefore, both the models were able to predict negative reviews quite accurately however they were not able to correctly predict the positive reviews.

Steps for improvement

The imbalanced dataset issue can be resolved by simply balancing the dataset by oversampling instances of the minority class or undersampling instances of the majority class. The data can be balanced during data preprocessing step, or we can use advanced techniques like SMOTE (Synthetic Minority Over-sampling Technique) that helps create new synthetic instances from minority class.

Utilization of results

The results can be utilized by AT&T and Verizon to understand how their company is performing in terms of services offered by them to their customers. They can build strategies to retain their existing customers and acquire new customers. The companies can work upon themselves to improve on different aspect where they are lacking on the basis of customer reviews.

The results can also be utilized by the competitors to understand why these companies hold a competitive advantage over others. They can understand what features differentiate these companies from others in the market. Telecom industry is growing day by day and to remain in the competition it is very important to understand the services offered by the market leaders and to learn from their mistakes and build better product offerings then what already exists in the market.

Business insights from analysis

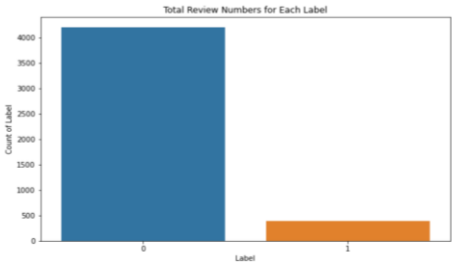
Through the analysis we got to know the most discussed topics in reviews about AT&T and Verizon. There has been a lot of discussion about customer service, internet service, data plan, billing cycle in AT&T and about customer service, data plan, unlimited data, service rep, tech support in Verizon. This shows us that this area should be explored and given proper attention.

There has been negative feedback given too about AT&T such as worst experience, poor service, service issue, poor customer. The negative feedback about Verizon consisted of bad experience, poor customer, fraud department, bad experience, poor service. The analysis shows that both

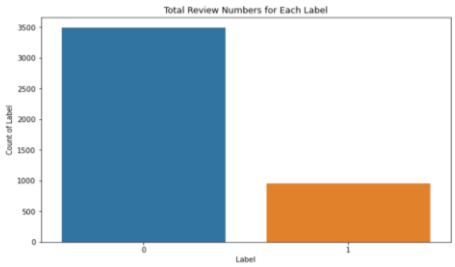
AT&T and Verizon customers have issues related to customer service. The companies should work upon their customer service.

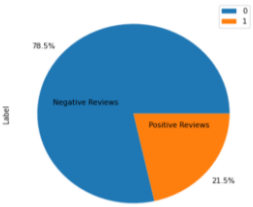
The positive features highlighted of AT&T were easy use, great deals, excellent service while those of Verizon are always reliable, extra mile, reliable network, great plan, signal everywhere, reception great, service reliable, fit budget, rarely dropped, great experience, good selection. The product offerings of Verizon are appreciated more as compared to AT&T.

From 2015 to 2020, Verizon has received fewer negative reviews and more positive reviews when compared to AT&T. This shows that Verizon has performed better as compared to AT&T over a period of six years.



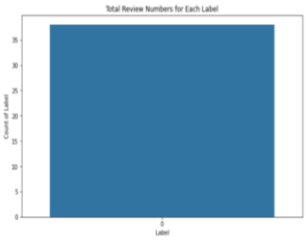
AT&T

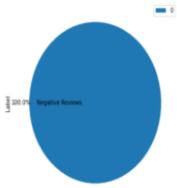




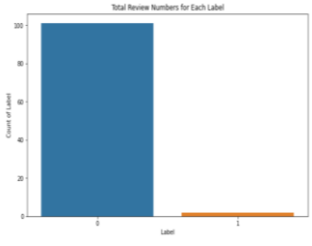
Verizon

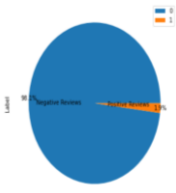
However, in 2021, Verizon has received more negative reviews when compared with AT&T and it has received 2 positive reviews while AT&T has received 0 positive reviews. This shows Verizon should improve their service since the usage is greatly increasing during COVID-19 times.





AT&T





Verizon

CONCLUSION

Sentiment analysis is extremely useful in monitoring feedback, and it allows the companies to quickly gain an overview of the wider public opinion. This helps the companies to strategize and plan for the future.

Analyzing the sentiments shall enable the companies to understand their strengths and weaknesses. After analyzing the entire reviews, we can conclude that from 2015 to 2020, Verizon has provided better overall service than AT&T since the number of positive reviews received by Verizon is higher than AT&T and the number of negative reviews received by Verizon is less as compared to AT&T. Also, through collocations and feature engineering, Verizon had a greater number of positive features when compared to AT&T. In 2021, Verizon has received higher number of negative reviews when compared with AT&T.

Through our analysis, we can also state that support vector machine model performs better than naïve bayes model in predicting the sentiments of the reviews. The f1-score is higher while predicting the reviews through support vector machine model in comparison with naïve bayes model.

Also, the review with lesser rating is bigger in length as compared with reviews with higher in rating.

FUTURE WORK

This research focused on two companies, in future other companies can be included and more in-depth study of the industry can be done.

The dataset did not include any geographical information of different regions of the United States. If the geographical information is available in the future, then analysis can be done on the basis of region and region wise performance of the telecom service provider can be evaluated.

Demographic information of the reviewers is missing. If information such as age of the reviewers is available, then a detailed study can be done on the preference of services by different age group of the users thus helping companies design strategies on the basis of the same.

REFERENCE

- https://www.investopedia.com/ask/answers/070815/what-telecommunications-sector.asp
- https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47
- https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/