

**Topic: Sentiment Analysis of Tweets on Electric Car-sharing Services, BlueSG**

|  |  |
| --- | --- |
| **Course Code** | ANL488: Business Analytics Applied Project |
| **Title** | Final Report |
| **Seminar Group** | T04 |
| **Submission Date** | 6 November 2022 |
| **Name** | Krystal Chee Jia Qi |
| **PI number** | B1981108 |

**Abstract**

The internet today is now a social platform for online learning, idea exchange and opinion sharing. Social networking services such as Twitter, Facebook, and Instagram are fast gaining popularity because they allow people to share and express their opinions on many topics, engage in discussions with various communities, and post messages all over the world. This study seeks to understand consumer sentiment and attitudes towards electric car sharing by exploring its usage in the context of electric car-sharing company, BlueSG, in Singapore. Sentiment analysis was conducted on social media comments made on BlueSG electric car-sharing service. Users' tweets and comments on social media platforms can reveal its brand reputation and consumer behavior, allowing researchers to better analyze public impressions of EVs and its services such as the renting procedure, usability, parking stations, and so on. According to the findings, the unsupervised lexicon-based approach and LDA topic modeling are effective in revealing positive sentiment about the company as well as concerns about the usability of its electric car sharing services. As a result, BlueSG should prioritize important aspects of its classification problems, such as focusing on reducing incorrect sentiment classification of tweets or implementing marketing strategies to address consumer concerns.

**Table of Contents**

[Chapter 1.0 4](#_bookmark4)

[Chapter 1.1 Introduction 4](#_bookmark4)

[Chapter 1.2 Business Problem](#_bookmark2) 54

[Chapter 1.3 Business Objective 54](#_bookmark3)

[Chapter 1.4 Text Mining Objective](#_bookmark3) 65

[Chapter 1.5 Applications of EVs 65](#_bookmark3)

[Chapter 2.0 Literature Review 86](#_bookmark4)

[Chapter 2.1 Introduction 86](#_bookmark5)

[Chapter 2.2 Text Mining on Consumers’ Preferences for EVs](#_bookmark5) 86

[Chapter 2.3 Consumers’ Willingness and Opinions on EVs 98](#_bookmark5)

[Chapter 2.4 Mining Carsharing Use Patterns 11](#_bookmark5)

[Chapter 2.5 Summary and Limitations 12](#_bookmark5)

[Chapter 3.0 Data Understanding and Preparation 13](#_bookmark4)

[Chapter 3.1 Data Understanding 13](#_bookmark7)

[Chapter 3.2 Data Preparation 15](#_bookmark7)

[Chapter 4.0 Proposed Modeling and Evaluation 18](#_bookmark4)

[Chapter 4.1 Proposed Modeling 18](#_bookmark7)

[Chapter 4.2 Evaluation Metrics 19](#_bookmark7)

[Chapter 5.0 Modeling Results 20](#_bookmark4)

[Chapter 5.1 Sentiment Analysis and Classification 20](#_bookmark7)

[Chapter 5.2 Confusion Matrix Analysis 22](#_bookmark7)

[Chapter 5.3 LDA Topic Modeling 24](#_bookmark7)

[Chapter 6.0 Recommendations and Conclusion 28](#_bookmark4)

[Chapter 6.1 Recommendations 28](#_bookmark7)

[Chapter 6.2 Conclusion 29](#_bookmark7)

[Chapter 7.0 Proposed Schedule 31](#_bookmark4)

[References 32](#_bookmark7)

[Appendix 33](#_bookmark7)

**Chapter 1.0**

**Chapter 1.1 Introduction**

In view of global warming today, the Singapore government has launched the Singapore Green Plan 2030 (SGP30) as part of the country’s efforts to combat climate change. With around 1 million vehicles on Singapore roads (Budget Direct Insurance, 2022), part of the government’s plans surrounding environmental sustainability targets vehicles and their impact on global warming.

In the present technological landscape with vehicular transportation, most of the vehicles on Singapore roads still function with internal combustion engines. Given that the utilization of EVs has significantly less carbon footprint as compared to vehicles with internal combustion engines, the Singapore government has also endorsed the consumption of EVs as part of their SGP30. This is followed by rigorous installation of EV charging outlets all over Singapore, with increasingly developed infrastructure to facilitate more accessible utilization of EVs on Singapore roads.

Combining electric vehicles and car sharing is a smart method to overcome these drawbacks and create a greener mode of transportation because car sharing is better suited to meet people’s short-term travel needs. In Singapore, car-sharing as an alternative mode of vehicular transportation comprises over 600,000 cars out of the vehicles present on the road (Budget Direct Insurance, 2022). Among the car-sharing services in Singapore, BlueSG represents the primary car rental service for EVs.

According to Wang et al. (2012), the idea of car-sharing first emerged in the late 1940s in Europe. Additionally, Singapore is an important car-sharing hub in Asia. Singapore’s car-sharing programs were more focused on giving residents who wanted to use vehicles access to it and mobility than they were on reducing auto ownership. Early initiatives in Singapore used cutting-edge technology and focused more on one-way trips than on electric vehicles, but like in Japan, these initiatives gradually merged to offer more conventional car-sharing services (Brook, 2008).

**Chapter 1.2 Business Problem**

Electric vehicles (EVs) are a developing and modern kind of transportation that provides a cleaner substitute for more environmentally friendly transportation. Singapore is in a unique situation in the EV business because, while the market is relatively small, it benefits significantly from being a small country.

However, there is a lack of willingness to adjust to the difficulties of using an electric vehicle in Singapore because of significant disadvantages like the cost of service, the longevity of the battery, the scarcity of charging locations, and the limited number of EV parking locations, and many more.

Another modern technique of renting cars for shared usage is the car-sharing service. As it is normally rented, booked, and used in specially constructed parking places scattered around Singapore, it is both demanding and alluring for its convenience and alternative to transit. The viability and usefulness of EV sharing systems still strongly rely on the infrastructure at renting or returning stations, despite their rising popularity. It is crucial to be able to accurately forecast demand dynamically before executing any expansion strategies, especially for systems that must quickly expand their station networks.

With that said, the Singapore Green Plan 2030 (SGP30) outlines its plans to have cars with internal combustion engines to be phased out, and all vehicles run on greener energy. It aims to have 60,000 charging outlets in place by 2030, including 40,000 public and 20,000 private ones. In addition to these measures, the government has committed to investing S$30 million in EV-related activities over the next five years, such as increasing the availability of charging services in private homes.

**Chapter 1.3 Business Objective**

Due to the present situation of undeveloped charging infrastructure and minimal driving mileage for EVs, consumers are concerned about the utilization of EVs in personal ownership and car-sharing.

The objective of this project is to strengthen BlueSG’s service by identifying and analyzing the impact of common customer pain points across the business and its one-way and all-electric car-sharing service to find out what needs to be improved and make recommendations based on social media reviews to increase customer retention and better brand awareness.

**Chapter 1.4 Text Mining Objective**

The text mining objective is to perform text mining on social media comments made on the electrical car-sharing service, BlueSG. Users' tweets and comments made on social media platforms can surface service performance insights and consumer behavior so as to better understand the public view or perceptions on certain aspects of EVs and the car sharing service such as renting process, usability, parking stations, etc. Patterns and trends can be obtained to help BlueSG formulate new policies and better service features to better serve users.

**Chapter 1.5 Application of EVs**

The COVID-19 pandemic and the ensuing recession had a significant impact on the global market for all sorts of automobiles. The expectation for global EV sales at the beginning of the year was extremely erratic due to the epidemic. As such, the overall number of new car registrations in 2020 did not significantly increase. As time passed, it became clear that 2020 was a shockingly successful year, with global EV sales increasing by 43% from 2019 and the market share of the electric car sector reaching a record 4-6%. Following that, up to 6.75 million vehicles were sold in 2021, a doubling of the market. In 2021, more electric vehicles were sold in a single week than were in all of 2012. The industry is unmistakably preparing for the ambitious goal of zero emission targets set for 2050, which will be mostly driven by EVs.

A study by Nienhaus (2022) from the Oliver Wyman Forum and the Institute of Transportation studies at the University of California examined 13 services across three continents—North America, Europe, and Asia—and found that the new services would develop at various rates depending on the region. Europe will surpass Asia as the region with the fastest-growing electric car industry, and Asia will expand its market for services like bike and moped sharing. The Asian micro-mobility market is currently comparable to that of Europe and North America put together but will surpass those markets by 2030, thus demonstrating the aggressive expansion of the Asian electric car-sharing industry and how with the rising electrification of transportation services, the sector is anticipated to be more environmentally friendly (Nienhaus, 2022).

Due to the decreased combustion of fossil fuels and pollution emissions, EVs are more environmentally friendly than gasoline-powered vehicles. According to the Ministry of Transport in Singapore, EVs produce 50% fewer carbon emissions than gasoline vehicles. In Singapore, less than 5% of newly registered automobiles are electric, despite this trend.

However, the potential for the transportation sector to become even more sustainable can be unlocked by fusing the trend of car sharing with EVs. Passenger car usage is currently extremely low. They are only used for a single trip for a short period before being inactive. However, car sharing can cut down on pollution because fewer automobiles will be on the road. There will be less demand for parking spaces for cars, which opens up more area for vegetation.

Residents may be reluctant to utilize EVs as Singapore switches from gasoline to electric vehicles. According to the Electric Vehicle Association of Singapore, 30% of respondents from Singapore indicated that they were hesitant to purchase EVs because they were unfamiliar with them. The price of the purchase and access to charging stations were also an issue (EVAS, 2022).

To understand the current research and discussions relevant to electric car sharing services and the various data mining methodologies used, the following chapters will cover literature reviews. Along with the data preparation phase of my research, the proposed methodology I would use to conduct my research, the research findings, and research recommendations for BlueSG.

**Chapter 2.0 Literature Review**

**Chapter 2.1 Introduction**

There is a substantial amount of research and prior studies over the years on the usage and impacts of electric vehicles and the aspects of transportation being adopted as a shared economy model, known as car-sharing. The relevance to this paper is previous works that examine online behaviors, opinions and usage patterns of electric vehicles as well as electric vehicle sharing that will be reviewed in this paper. This portion of the paper refers to the different methods of the various literature reviews for the research proposal.

**Chapter 2.2 Text Mining on Consumers’ Preferences for EVs**

This research paper by Ma et al (2019), examines big data and text mining technology to investigate a vast number of Chinese consumers' online behavior related to EV selection, comparison, purchase inquiry, and comments to examine consumer preferences for EVs.

The most crucial variables affecting consumer response, according to their research, were EV costs, car classification, and engine types. They found that consumers preferred small battery electric vehicle (BEV) models with fast-charging batteries, which are ideal for daily use and commuting to work. Most importantly, it was discovered that EV aesthetics play a considerable effect on consumer decisions in addition to price and technical characteristics.

Since the end of 2016, they have been collaborating with Autohome, China's largest auto website, to gather sufficient data on consumer behavior. To better understand consumer preferences for EVs, behavior data from website visitors' browsing, comparing, sending purchase inquiries, and commenting on the website was gathered. Then, utilizing information about EV performance, they used text mining techniques to analyze a total of 27,930 comments made by EV owners about their concerns and motivations throughout the final buying stage as well as how they used their vehicles.

The results from the analysis identified that in terms of purchase prices, it is an important factor for consumers throughout the entire car purchase process as compared to the performance specifications of the vehicles. In addition, consumers place a higher value on charging time than all-electric range or battery capacity. The analysis also suggested that consumers focused more on aesthetics, including look and interior than on EV brands, services, and other aspects, in addition to the purchase costs and performance criteria.

Therefore, when explored in detail, the information analyzed can impact car-sharing service providers such as updating an EV's look and interior since it is quite appealing to customers. From a business perspective, studying customer preferences for EVs is vital, and this includes figuring out the different aspects that affect consumers' decisions to buy EVs. With BlueSG being a car-sharing service offering EVs, analysis of the various perceptions of the customer can reveal preferences, as well as solutions.

Using a similar text mining method on consumer opinions from social media in my research can help the business reflect its offerings and consumers' needs to achieve its business objective. Therefore, consumers' opinions for and uses of BlueSG's car sharing service will be the focus of this research. With reference to the methods and objectives of this literature review, BlueSG can benefit from gaining valuable insights that can promote environmental initiatives and appeal to consumers such as improvements on technical specifications, aesthetics, or even mobile application improvements.

**Chapter 2.3 Research on Consumers Use Willingness and Opinions on EVs**

An empirical study conducted in Shanghai by Wang & Yan (2015), examines customer attitudes toward EV sharing and their desire to use it to assist company owners in efficiently implementing and growing the new business model. A questionnaire was created to gather data regarding consumer acceptance of EV sharing and the elements that influence customers' decision to use EV sharing to support the all-round development of EVs and assist businesses in operating more efficiently. Operators of EV sharing services are given some recommendations for accelerating the technology's growth in China.

Between May 2014 and November 2014, 410 respondents participated in a survey. Three sections make up the questionnaire. A three-part questionnaire was used. Three questions are included in the first section of the survey regarding various aspects of travel, such as the primary method of travel used daily, the variables considered when selecting the mode of travel, and monthly transportation costs. In the second section, respondents were asked about their expectations of EVS, including their use willingness, what appealed to them about it, and what scenarios would be most appropriate for it to be used in, etc. Personal information, which included gender, age, occupation, degree of education, marital status, and monthly income, was the final section.

Using SPSS to perform a test of parallel lines determined that a multinomial logistic regression was appropriate for various groups. Through this analysis, it demonstrated that factors such as the primary trip mode used daily, monthly transportation expense, the driving range of electric vehicles, gender, age, marital status, and occupation have a significant impact on consumers' use willingness. In other words, people who choose to utilize EVS are typically male, between the ages of 18 and 30, and use the bus and subway as their primary forms of transportation. Otherwise, criteria like occupation, personal monthly income, and the accepted maximum price of EVS have a big impact on how willing individuals are to utilize it. These people place more emphasis on EVS's economy and convenience.

These findings show that operators need a fair price, precise target group positioning, a practical site layout, and easy usage to successfully introduce a new EVS transportation mode. While views and demand regarding the usage of EVs have been improving in Singapore, EVs are primarily seen as being expensive and inconvenient to own at this point. Therefore, this research suggests that investigating the key influencing elements for customer acceptance of EVS in the early stages of development is vital to ensure a successful commercial operation. Nevertheless, it is still unclear how consumers feel about EV sharing services if they would use the service and their preferences in this regard.

Therefore, by performing text mining on social media comments made about BlueSG, the aim of this process is to investigate if relevant opinions will reveal findings similar to that in the literature review on China (in view that this paper’s data focuses on consumer's perceptions of BlueSG, a car-sharing operator in Singapore with a different population characteristic). Using the insights obtained, BlueSG can more effectively operate and expand their service and better position target groups in Singapore.

**Chapter 2.4 Mining Car-sharing Use Patterns**

Another study by Qian et al. (2017) examines the travel habits and activity characteristics of various active member types in a car-sharing system. It presents car-sharing user patterns using data from Chefenxiang, a car-sharing service in Hangzhou, China, and the SPSS Two-step Cluster Analysis method to examine car-sharing user trends. Chinese car-sharing operator, Chefenxiang, provided the monthly rental data for the period of September 2013 to September 2014 that was used in this study. The classification of car-sharing participants with comparable usage patterns was the main topic of this study.

To examine the usage trends of car-sharing, the SPSS Two-Step Cluster Analysis tool was developed. To describe member travel characteristics and car-sharing use behaviors, a set of descriptive variables was chosen. Five clusters—short-term frequent heavy usage pattern, long-term frequent light usage pattern, long-term occasional light usage pattern, and long-term frequent heavy usage pattern—were ultimately identified by calculating the usage count, usage location count, usage time length, and usage time interval of car-sharing members.

Results from the spatio-temporal analysis of these five clusters were that stable users for commute or business purposes were predicted to be long-term frequent heavy members. Students from universities and others who frequent office buildings and commercial districts are drawn to car-sharing in China. Car-sharing was primarily used for short-to-medium distance, transient trips. Without previous knowledge of the categories of user behaviors, the clustering method to mine use patterns was valid.

In comparison to the previous two literature reviews, this research paper aims to point out a thorough understanding of EVS acceptance and the target audience for EVS from the relationship between various types of car-sharing use patterns and inconsistent requests for land use around car-sharing locations. From a business perspective, BlueSG can benefit from the reference of this literature review as we can research travel behaviors and activity features made on the opinions of users in Singapore. Findings are ineluctably constrained by the data provided without car-sharing users' sociodemographic data, vehicle GPS and rental data. However, there is still an increased understanding if we can discover car sharing user behavior and specific characteristics.

**Chapter 2.5 Summary & Limitations**

From the literature reviews above, some takeaways are that analyzing consumers' opinions can reveal insights into how consumers feel about the brand and suggest potential solutions. In this research, a similar text mining technique will be applied to gather consumer feedback from social media on BlueSG. This will allow the company the ability to reflect on its services and better understand the customers' wants. Additionally, examining the primary influencing factors and elements that affect user acceptability of EVS as well as gathering possible consumer characteristics and usage patterns as seen in the studies done in China in my sentimental analysis research for BlueSG.

However, there are several limitations in the studies such that when utilizing consumer online behavior or implementing questionnaires, the data gathered may leave out some crucial elements. For instance, being unable to gather demographic data about EV owners, such as their gender, age, and income, due to concerns about customer privacy. As a result, the analysis cannot take customer heterogeneity into account and cannot track customers' offline behavior. Additionally, users often produce inconsistent data when using unstructured language, making it difficult to extract accurate and consistent data patterns.

**Chapter 3.0 Data Understanding and Preparation**

**Chapter 3.1 Data Understanding**

Data is highly abundant in Twitter. A user's tweets are entirely public and retrievable, unlike other social media networks. Also, quite specific to Twitter’s data is that it allows users to run complicated searches using Twitter's API, such as getting all non-retweeted tweets from a specific person or all recent tweets regarding a particular topic or hashtag.

However,due to a large number of meaningless tweets and the fact that the majority are either devoid of context or too brief to be effective meaning carriers, scraping and analyzing comments posted on Twitter can be challenging. The character limit for a Twitter post might lead to strange writing styles and noises in tweets that would not be present in everyday English, like using abbreviated words, emoticons, Html etc.

Therefore, the data flow diagram gives you a summary and visual representation of the data mining process of my research from the data selection process, data preparation and text pre-processing to the modeling.

|  |
| --- |
| **Selection**  Scraping raw data from Twitter using API  Data Source  Setting up a twitter account and developer credentials to use the Twitter API |
| **Data Preparation**  Data cleaning for filtered and noise free data  Loading data to Jupyter to perform coding |
| **Text pre-processing**  Linguistic Data processing using NLP |
| **Modeling**  Evaluation & Recommendations  Apply sentiment analysis and topic modelling |

***Diagram 1: Text Mining - Data Flow Diagram***

**Chapter 3.2 Data Preparation**

**Selection**

1. Identified appropriate social media data source – Twitter to extract my consumer opinions for my sentiment analysis on BlueSG.
2. Creating and accessing Twitter developer portal to obtain consumer keys and authentication tokens to use API for Twitter scraping.
3. Extracting Twitter comments into a textual csv file dataset using Twitter API with Python. under a set of criteria:

* Pertaining to the keyword “BlueSG”
* Start date and end date for tweets to be scrapped
* Number of tweets to be scraped

**Data preparation**

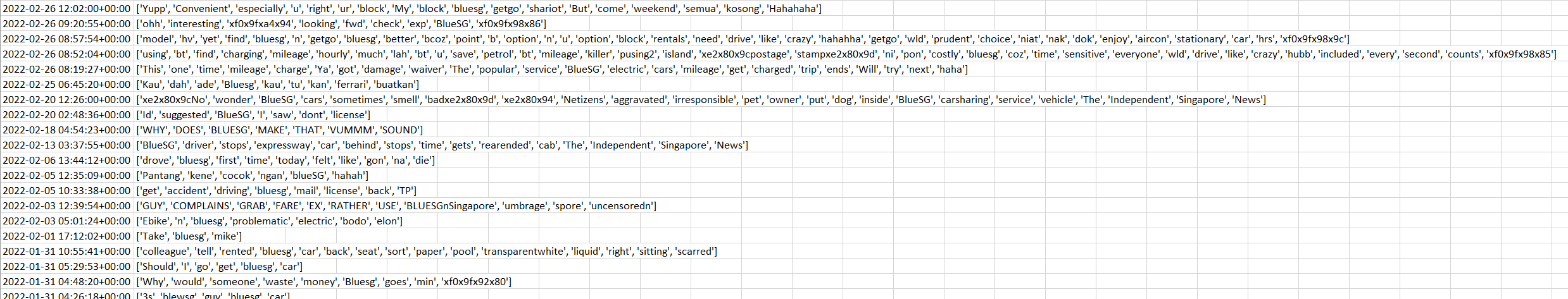
1. Data cleaning and formatting performed in Python includes:

* Decoding HTML
* Removal of links, hash, username, and punctuations
* Removal of desired stop words such as “the”, “a”, “an”, “so”, “what”.
* Tokenizing the tweets

In the data preparation process, it is noted that web scraping Twitter data to obtain textual comments pertaining to BlueSG, tweets with the hash tagged content “BlueSG” were extracted using Python and stored in a csv file. The processing of scraping tweets on Twitter most importantly requires the need for credentials and the installation of packages and libraries such as Snscrape, Tweepy and a tweet-preprocessor package (refer to Appendix A and B).

Firstly, with a Twitter account, an application to become a Twitter developer needs to be submitted to acquire access to the Twitter API which enables programs to access Twitter. Before using Tweepy in Python, consumer keys and authentication tokens that are unique to each account are required to utilize the Twitter API. After which, the desired tweets to be scraped utilize a specific query keyword, “BlueSG”, with the date period (2019-2022) and the number of tweets to scrap which will be coded to have it stored in a csv file.

After which, it can be observed that the rows of data retrieved pertaining to the BlueSG comments made by users on Twitter are in short and brief sentences and mainly in English with a few which may contain unknown characters that could be from emojis or in a different language which may pose as a challenge for my text mining application. Some of these noises include HTML, links, hash, username, punctuations and stop words. Therefore, data cleaning for a data output for sentiment analysis is required. Using python, cleaning libraries are imported that perform the cleaning of the noises as mentioned above (refer to Appendix C and D).

****

**Figure 1. *Extracted dataset from Twitter scraping using Python***

After the data extraction and cleaning are performed, the resulting dataset as seen in Figure 1, contains about 800 rows and 2 columns consisting of the date and tweets scrapped using Python for the sentiment analysis using sentimental analysis approaches.

**Text pre-processing**

During this stage, Natural Language Processing with Python's Affin Package is initialized after data pre-processing to process the text data. With the use of the Natural Language Toolkit (NLTK) for processing the text data, it helps us creates a sentiment dictionary and assigns a text score based on a function that indicates how well the text's words and phrases match the dictionary. This way, we can identify the positive and negative sentiment of the tweet pertaining to BlueSG and conduct modelling techniques and use these algorithms to extract insights from linguistic data through effective built-in machine learning operations.

**Modeling**

1. Sentiment analysis to identify brand reputation based on tweets from consumers (refer to Appendix E and G).
2. Topic Modeling visualisation - Identify intriguing patterns amongst text data by visualizing the relationships between them based on topics generated (refer to Appendix H).
3. Evaluating the modeling results.

**Chapter 4.0 Proposed Methodology and Evaluation**

**Chapter 4.1 Proposed Methodology**

Once the data has been parsed and categorized, the next step is to perform modeling of the Twitter comments. In this research, the aim is to analyze BlueSG's users or potential customer preferences for its electric car-sharing service, namely from Twitter. To reach its aims, sentiment analysis of tweets pertaining to BlueSG using Python library, Affin, will be conducted to identify the key consumer sentiments using BlueSG. Afinn, is a popular lexicon used for sentiment analysis. It contains 3300+ words with a polarity score associated with each word.

Using an unsupervised lexicon-based approach, words in texts are labeled as positive or negative with the help of a dictionary of lexicons that can be created manually or automatically generated. Lexicon is a list of lexical words that are labeled with positive of negative based on the semantic meaning, with a positive or negative sentiment score assigned to each word. This approach compares the words in the tweet to lexicons, a dictionary of sentiment words, to determine the sentiment orientation. However, uncertainty surrounds the best strategy for developing a sentiment lexicon. There are numerous words lists with emotional valence labels. These word lists vary in terms of the words they contain, for example, some do not contain strong obscenities or Internet acronyms like "WTS" and "LMAO."

Consequently, performing an (Latent Dirichlet Allocation) LDA Topic modeling where the automated model analyzes the text data and creates the clusters of the words from that dataset or a combination of documents. It works on finding out the topics in the text and find out the hidden patterns between words relates to those topics. After which, an interactive visualization of the clusters can be created to better understand and interpret individual topics. With reference to the study conducted by Ma et al. (2019), although unstructured textual data is meaningful, when the volume of information may be too great for manual analysis, text mining technology will be employed to analyze the text data gathered.

**Chapter 4.2 Evaluation Metrics**

Sentiment analysis employs the evaluation criteria of performance metrics like Precision, Recall, and F-score as a classification issue. Precision, Recall and F-score for evaluating classifier performance, looks at the actual and predicted results produced by the classifier in a confusion matrix. The metrics can be computed as per formulas given below.

F-score = (2\* Precision \* Recall) / (Precision + Recall)

Precision = True Positive / (True Positive + False Positive)

Recall = True Positive / (True Positive + False Negative)

|  |  |  |
| --- | --- | --- |
| Actuals | Predicted Class | |
| Predicted Positive | Predicted Negative |
| Actual Positive | **True Positive** | **False Negative** |
| Actual Negative | **False Positive** | **True Negative** |

***Table 1: Confusion Matrix Format***

The F-score measure is combination and a harmonic mean of Precision and Recall. Only when one of their values gets high or low does it occasionally become challenging to make a judgment based on Precision and Recall. The validity of the result can then be determined by taking the F measure into account.

To fully evaluate the effectiveness of a model, we must examine both precision and recall as they are often in tension. unfortunately improving precision typically reduces recall. Therefore, there is a need to consider a precision or recall tradeoff.

Additionally, for multi-class issues, averaging measures like macro, micro, and weighted F1-scores are helpful. The best measure should be utilized based on the balance of classes in the dataset.

**Chapter 5.0 Modeling Results**

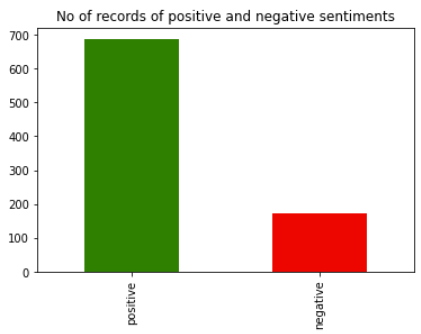
**Chapter 5.1 Sentiment Analysis and Classification**

After conducting the data preparation on the dataset containing the tweets to be use for sentiment analysis. The first objective was to find out the polarity score and sentimental distribution of the tweets. Using Python library, Affin, iteration through all tweets is conducted first to assign the polarity score and sentiment classes of each tweet. As seen in Figure 1 below, The Affin lexicon assigns words with a score that runs between -5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment based on the polarity of the opinion.

****

***Figure 1: Tweets with assigned polarity scores and class***

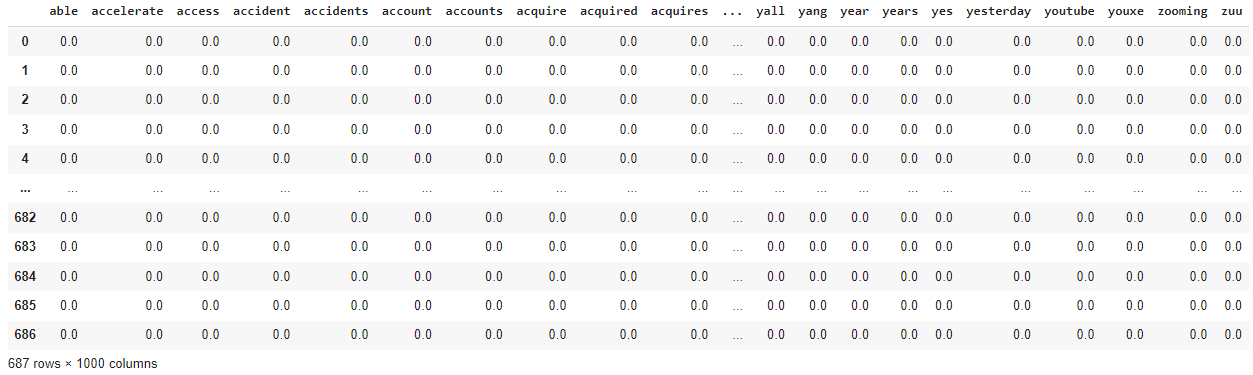
A bar chart is then plotted to get a sense of the distribution of sentiment. The bar charts reveals that there are more positive tweets then negative ones in the dataset. It also shows the general distribution of emotional perception of tweets on BlueSG being overall positive. In the context of sentiment analysis, users or customers of BlueSG are generally satisfied and have a positive brand perception. This also adds values to the brand managers to measure the performance of their brand equity in regional standards. The results are plotted visually in Figure 2 below.

****

***Figure 2: Sentimental Distribution of Tweets***

Once the sentiment for every tweet is achieved, a training and testing data is created by converting the tweet columns to numerical values using Python, the results are as seen in Figure 3. This process is known as text vectorization, the term is used to describe the process of turning text into a vector. It is a key step in natural language processing because no machine learning method, not even computers, can comprehend a text.

Text can be converted into vectors with the aid of the text vectorization technique Term Frequency-inverse document frequency (TF-IDF) vectorizer, which is a well-liked method for conventional machine learning methods. Additionally, this allows us to build a model for sentiment classification to see how good polarity is being identified. The training set data and test data were split in the ratio 80:20, ensuring at least 75% of data goes into training.

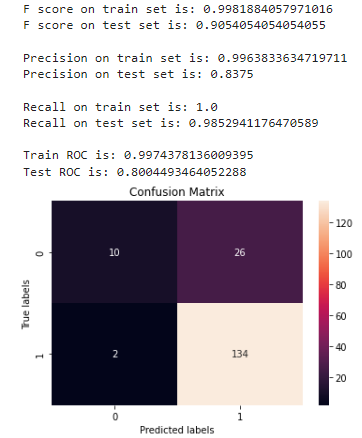
****

***Figure 3: Vectorization of tweets for training dataset***

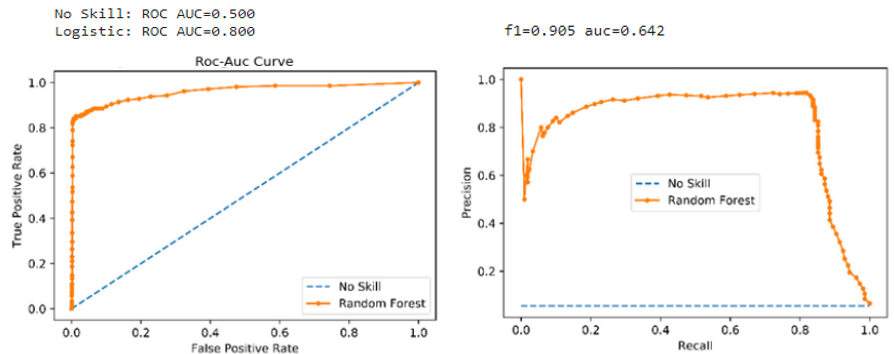
**Chapter 5.2 Confusion Matrix Analysis**

Performing text vectorization allows for the data from the testing set to gauge the accuracy of classification. The resulting confusion matrix in Figure 4, is a method of measuring performance that demonstrates how well the model categorizes each tweet and any potential areas for inaccuracy. It also measures precision, recall, f-score and ROC-AUC curve of the model. A confusion matrix helps us gain an insight into how correct our model classifications were in and how they hold up against the actual values. The x-axis are the predicted labels (0,1) and the y-axis are the actual labels (0,1). 0,1 represents the different classes respectively so label 0 is negative sentiment, 1 is positive sentiment.

As per the model prediction, there are 2 instances that are wrongly classified as negative and 26 instances that are wrongly classified as positive out of a sample size of 172 for test set. The rest have been classified correctly for its sentiments for actual outcomes against the predicted outcome. Having 0 instances would mean that the model is performing perfectly. The goal of this confusion matrix would be to maximize the values of the True Positive and True Negative sentiments. The best performance in terms of correct classification would be the positive sentiment with 134 being correctly classified, followed by 10 being negative sentiment correctly classified.

****

***Figure 4: Confusion Matrix with Performance Metrics***



***Figure 5: ROC-AUC Curve and Precision-Recall Curve***

`

The ROC-AUC curve and Precision-Recall curve for the model was also plotted to provide a summary of the model's performance and accuracy. These two curves are the most common measures to demonstrate the performance of a classification model. The modes showed a f-score of 0.905 or 91% for the polarity prediction and the area under the ROC is 0.8 indicating a good precision and recall performance. This means the model is very accurate and has a high sensitivity and specificity with high prediction of the sentiment of tweets.

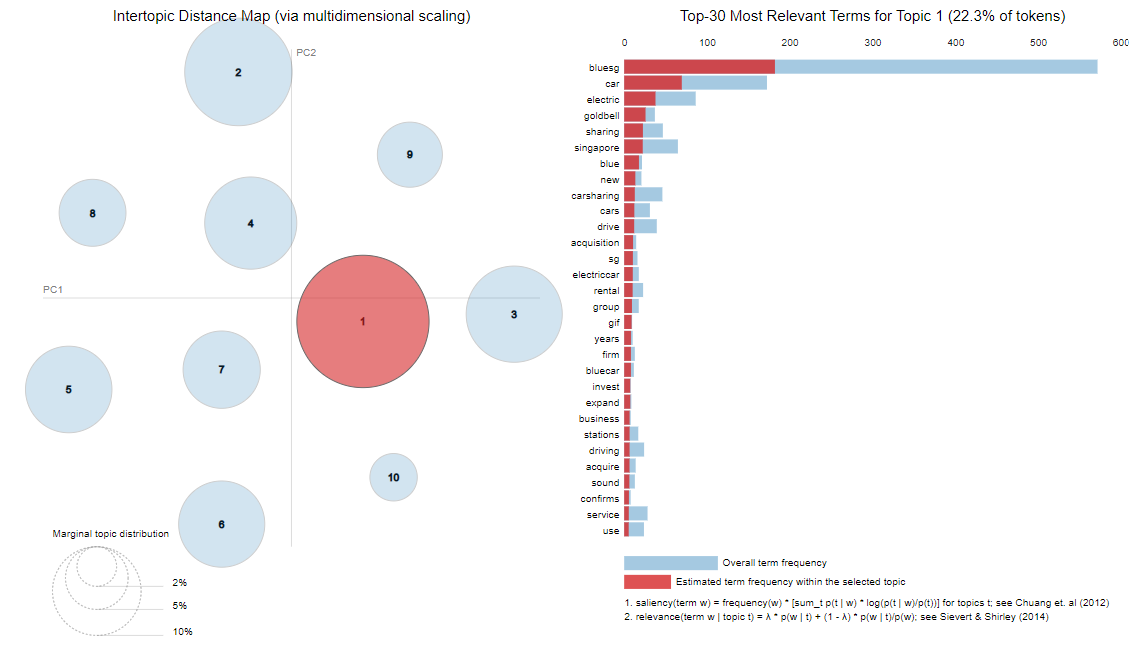
Additionally, we can better comprehend model evaluation by also observing the precision and recall metrics, which take classification accuracy one step further. Precision shows what percentage of the positive predictions made is correct, meaning, of all tweets classified as positive, how many of them are actually positive. Positive prediction is the main focus of precision as it shows the proportion of accurate forecasts. The precision on my test set is 0.837, in other words, when it predicts that a tweet is positive, it is correct around 84% of the time. Recall shows what percentage of actual positive predictions were predicted, meaning how much of the actual tweets that are positive were also predicted positive. The recall on my test set is 0.985, in other words, it correctly identifies 99% of all positive tweets.

Additionally, we can see that there is higher recall performance then precision performance. Recall can be viewed as a measure of quantity, whereas precision can be viewed as a measure of quality. Therefore, higher recall indicates that an algorithm provides the majority of the relevant results, while higher precision indicates that an algorithm delivers more relevant results than irrelevant ones.

**Chapter 5.3 LDA Topic Modeling**

Lastly, I employ an LDA Topic modeling where the automated model analyzes the text data and creates the clusters of the words from that dataset or a combination of documents. Because it automatically categories words without using a predetermined list of labels, this is unsupervised learning. It works on finding out the topics in the text and find out the hidden patterns between words relates to those topics. If you feed the model data, it will give you different sets of words, and each set of words describes the topic. After which, I created an interactive visualization of the clusters of words and numbers to better understand and interpret individual topics.

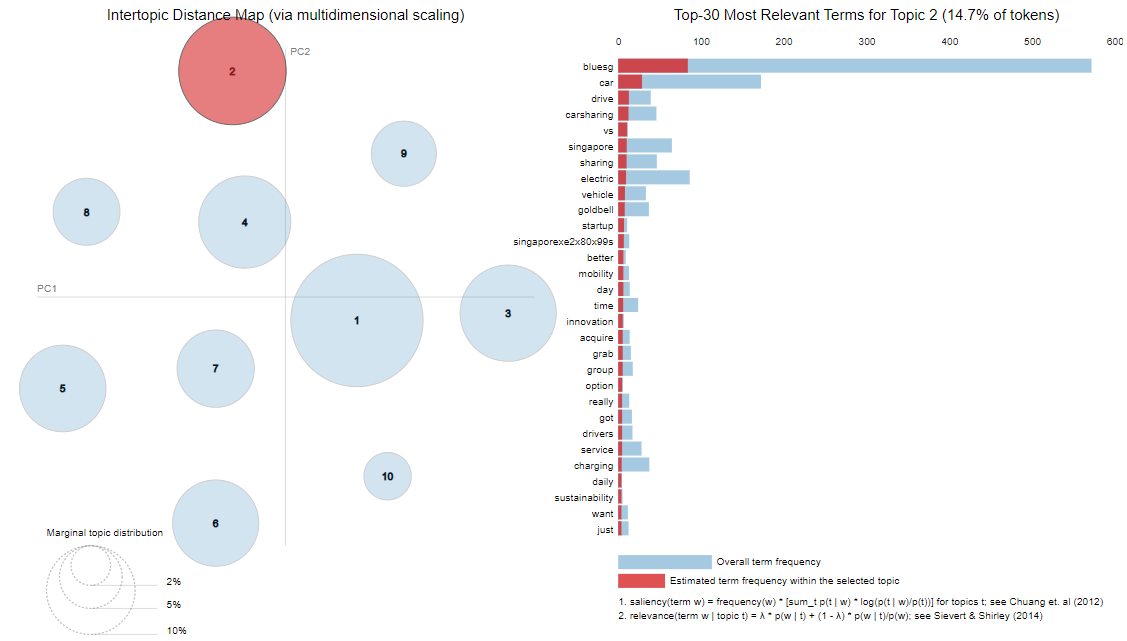
Topic modeling using Python library, pyLDAvis, works on finding out the topics in the text and find out the hidden patterns between words relates to those topics. The resulting LDA topic model as seen in Figure 6 is built with 10 different topics where each topic is a combination of keywords, and each keyword contributes a certain weightage to the topic.



***Figure 6: Intertopic Distance Map - Topic 1***

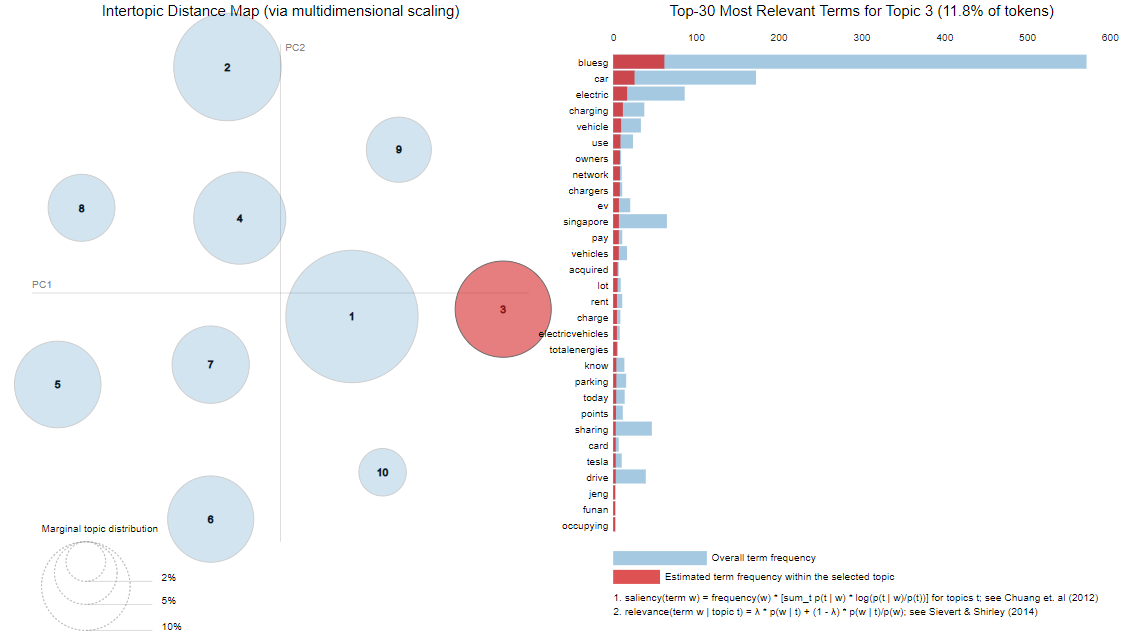
The Intertopic Distance Map on the left half of the chart represents each topic as a bubble, the size correlates to the prevalence of its topic within the text document. Blue bars represent the overall frequency of each word in the data. Red bars give the estimated number of times a given term was generated by a given topic.

For instance, from Figure 6, showing topic 1, there are about 550 of the word ‘bluesg’, and this term is used close to 200 times within topic 1. The word with the longest red bar is the word that is used the most by the tweets belonging to that topic. The further away the bubble are from each other the more different they are.



***Figure 6: Intertopic Distance Map - Topic 1***

***Figure 7: Intertopic Distance Map - Topic 2***



***Figure 8: Intertopic Distance Map - Topic 3***

Among the 10 topics generated, I narrowed down just the top 3 topics and its top terms to evaluate some of the concerns generated from tweets about BlueSG. I created a table to distinguish them between different topics and its respective top terms as seen in Table 2. Some of the top terms of topic 1 include goldbell, acquisition and rental, invest or expand associated with BlueSG car sharing service. And from topic 2 and 3, some of the frequent terms include mobility, charging/chargers, sustainability, pay, rental and parking points. This tells me that firstly, on Singapore’s Goldbell group acquisition of BlueSG, seems that many users are showing interest or concerns on the potential business changes after this acquisition. Secondly, there is concerns on its rental, its sustainable mobility, charging, payment and parking issue of BlueSG.

|  |  |  |
| --- | --- | --- |
| **Categories** | | |
| **Topic 1 – Politics/Business** | **Topic 2 - Aspects** | **Topic 3 - Service** |
| Goldbell | Mobility | Chargers |
| Acquisition | Time | Pay |
| Rental | Innovation | Rent |
| Invest | Charging | Parking |
| Expand | Sustainability | Points |

***Table 2: Topic Modeling Categories***

**Chapter 6.0 Recommendations & Conclusion**

**Chapter 6.1 Recommendations**

Recommendations based on confusion matrix

From the confusion matrix and the performance metrics analysis previously, BlueSG should consider a precision-recall tradeoff in terms of the classification objective or problem. In terms of understanding accuracy, to fully evaluate the effectiveness of a model, we must examine both precision and recall because they are often in tension. For instance, unfortunately improving precision typically reduces recall. Therefore, there is a precision-recall tradeoff. Hence, BlueSG should decide which is more important for its classification problem such as focusing on achieving higher precision meaning reducing incorrect classification of the sentiment of tweets. When building a model to see perception, the model should classify new tweets also from various social media platforms and identify their sentiment classes correctly, this can allow them to improve customer service, determine better marketing strategy and improve their product & service message, to better improve their customer retention and loyalty.

Recommendations based on topic modeling

Furthermore, according to the topic modeling research, some of the top terms of Topic 1 include Singapore's Goldbell group acquisition of BlueSG and concerns about BlueSG's rental, sustainable mobility, charging, payment, and parking issues. One strategy that could tackle concerns over payment and rental prices for BlueSG would be to adopt a dynamic pricing strategy to ensure customers get the best deal or pricing at any given point in time instead of packages or membership. Another strategy would be to suggest BlueSG focus on areas where it can enhance its offerings, such as the rental process on the app by having a more user-friendly interface, better marketing on its sustainability initiatives to encourage more sustainable transportation use, and better improving their charging and parking infrastructure.

The sentiment analysis gives BlueSG a general sense of the sentiment distribution of the reviews and brand reputation through the combined examination of these two approaches. And how topic modeling enables them to better comprehend the most common themes and terms found in tweets and adjust their strategies accordingly. We may also conduct a competitor analysis and analyze BlueSG's rivals, such as GetGo or Tribe Car, in order to learn more about customer preferences and valuable information that BlueSG might not learn from tweets that are exclusively about BlueSG. Finding and contrasting important market indicators that assist BlueSG in distinguishing its goods and services from those of its rivals is the main goal of competitive market research. Additionally, an efficient sales and marketing plan that makes its business stand out from the competition is built on a solid foundation that is created by thorough market research.

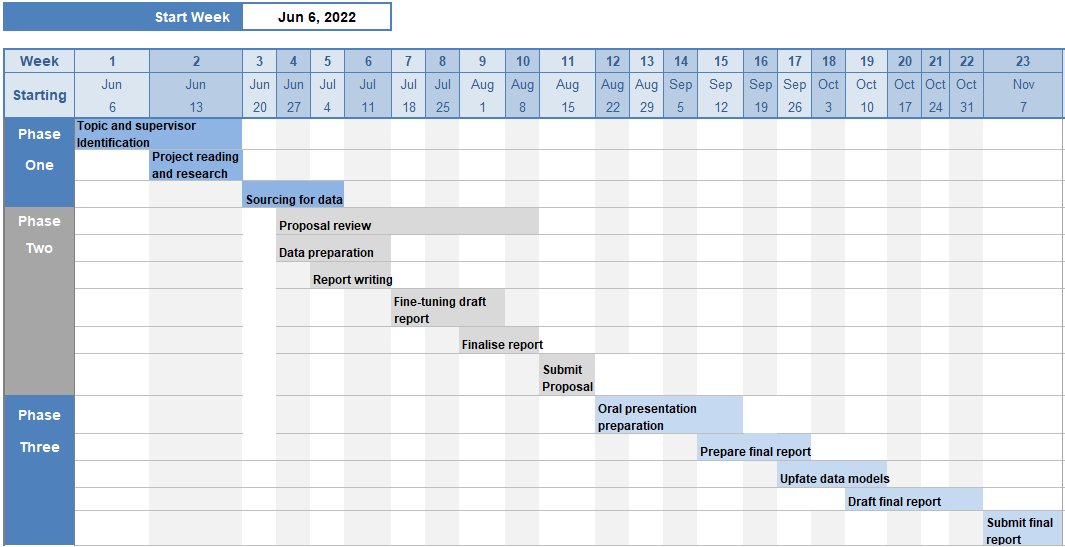
**Chapter 6.2 Conclusion**

With the availability of the internet and access to social media platforms and information, many of the social media platforms available now include sentiment analysis as an important feature to indicate positive or negative sentiment. It has developed into a way for businesses to understand their current customers, attract new ones, avert PR crises, and maintain their credibility. In its current state, text analysis provides a nice way to see the most basic interpretation of the emotion expressed in a post or tweet. In the worst-case scenario, the analytic tool misinterprets or is unable to interpret the emotion and the tone, such as when sarcasm is used.

Therefore, the limitation of this study is that natural language processing can make it challenging to analyze sentiment or feelings since machines need to be trained and educated to do so in the same way that the human brain does. Additionally, it might be challenging for sentiment analysis techniques to determine the genuine context of what the response is actually saying when negative sentiment is expressed by employing backhanded compliments. This can frequently lead to a higher volume of falsely positive sentiment tweets. It is critical to have a comprehensive and thorough methodology that allows for the consideration of pertinent context. For example, knowing that a specific user is often sarcastic or having a bigger sample of natural language data that provides hints to identify whether or not a sentence is ironic, would be required. The latest techniques include handling emoticons, spelling variations, and negation detection. None of these methods were utilized in the current application of my list, which could have been beneficial.

However, interpreting natural language data will be an issue for the next 2-3 decades. It is a difficult problem, because sarcasm and other sorts of sarcastic language are naturally difficult for algorithms to identify when seen in isolation. Additionally, sentiment is likely to fluctuate over time as a result of a person's mood, world events, and so on, it can be vital to examine data in terms of time.

**Chapter 7.0 Proposed Schedule**



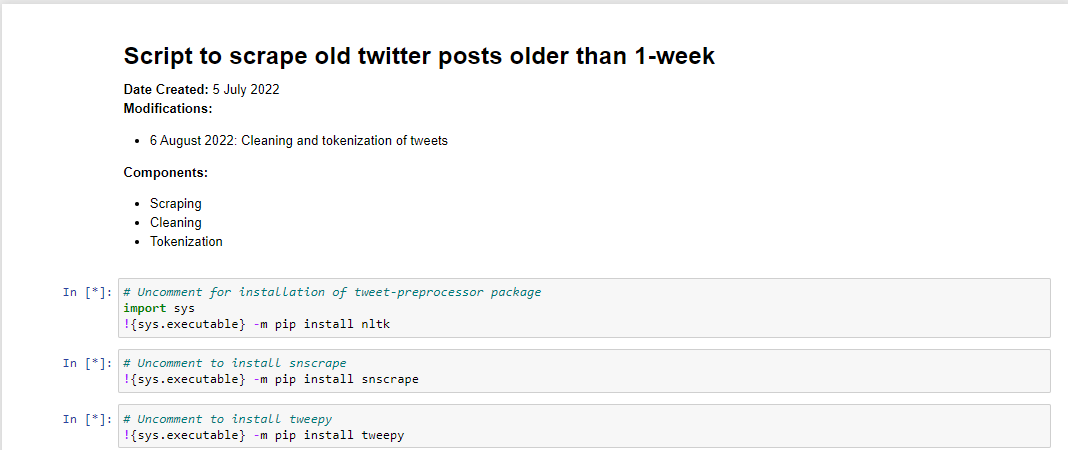
***Figure 2: Anl488 Project Schedule Timeline***

**(Total Word Count including Abstract: 6,649)**

**References**

1. EVAS. (2022). *Singapore study reveals strong interest in EV adoption despite challenges and misconceptions.* <https://drive.google.com/file/d/1359dpE-lzbtQsT_7wSn34Wlg7hkuOodV/view>
2. Ma, S.-C., Fan, Y., Guo, J.-F., Xu, J.-H., & Zhu, J. (2019). Analysing online behaviour to determine Chinese consumers’ preferences for electric vehicles. Journal of Cleaner Production. <https://doi:10.1016/j.jclepro.2019.04.374>
3. Nienhaus, A. (2022). *How Urban Mobility Will Change by 2030. Oliver Wyman Forum*. <https://www.oliverwymanforum.com/content/dam/oliver-wyman/ow-forum/mobility/2022/Value%20Pool%20Report.pdf>
4. Qian, C., Li, W., Ding, M., Hui, Y., Xu, Q., & Yang, D. (2017). Mining Carsharing Use Patterns from Rental Data: A Case Study of Chefenxiang in Hangzhou, China. *Transportation Research Procedia, 25, 2583–2602*. <https://doi:10.1016/j.trpro.2017.05.307>
5. Virta.global*. The Global Electric Vehicle Market In 2022 – Virta*. (2022). Virta.global. <https://www.virta.global/en/global-electric-vehicle-market>
6. Wang, M., Martin, E. W., & Shaheen, S. A. (2012). Carsharing in Shanghai, China. *Transportation Research Record: Journal of the Transportation Research Board, 2319(1), 86–95.* <https://doi:10.3141/2319-10>
7. Wang, N., & Yan, R. (2015). Research on Consumers’ Use Willingness and Opinions of Electric Vehicle Sharing: An Empirical Study in Shanghai. *Sustainability, 8(1), 7*. <https://doi:10.3390/su8010007>
8. Brook, D. (2008). *Innovative Carsharing Program in Singapore Ends*. <http://carsharingus.blogspot.com/2008/03/innovative-carsharingprogram-in.html>.

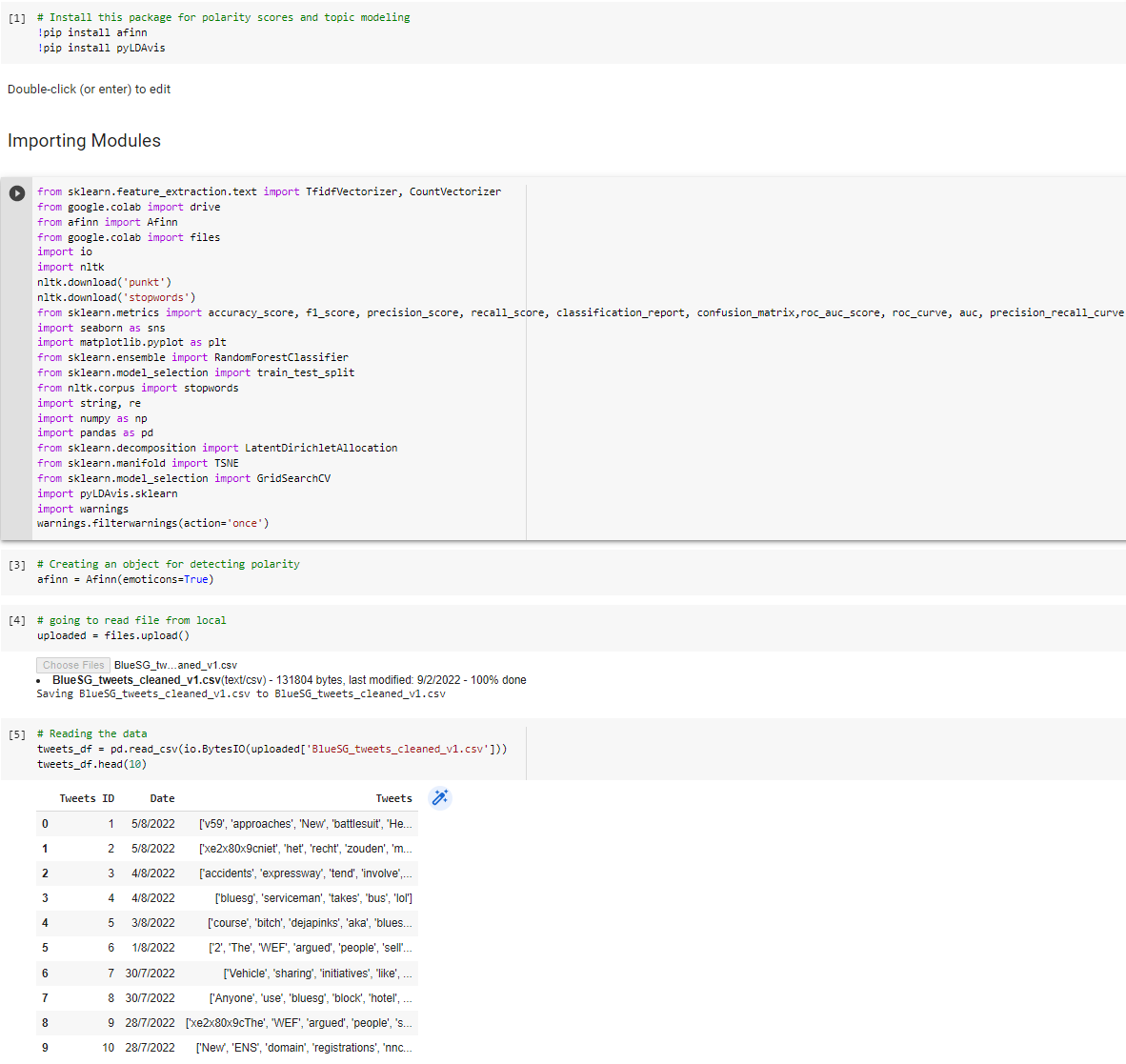
**Appendix**



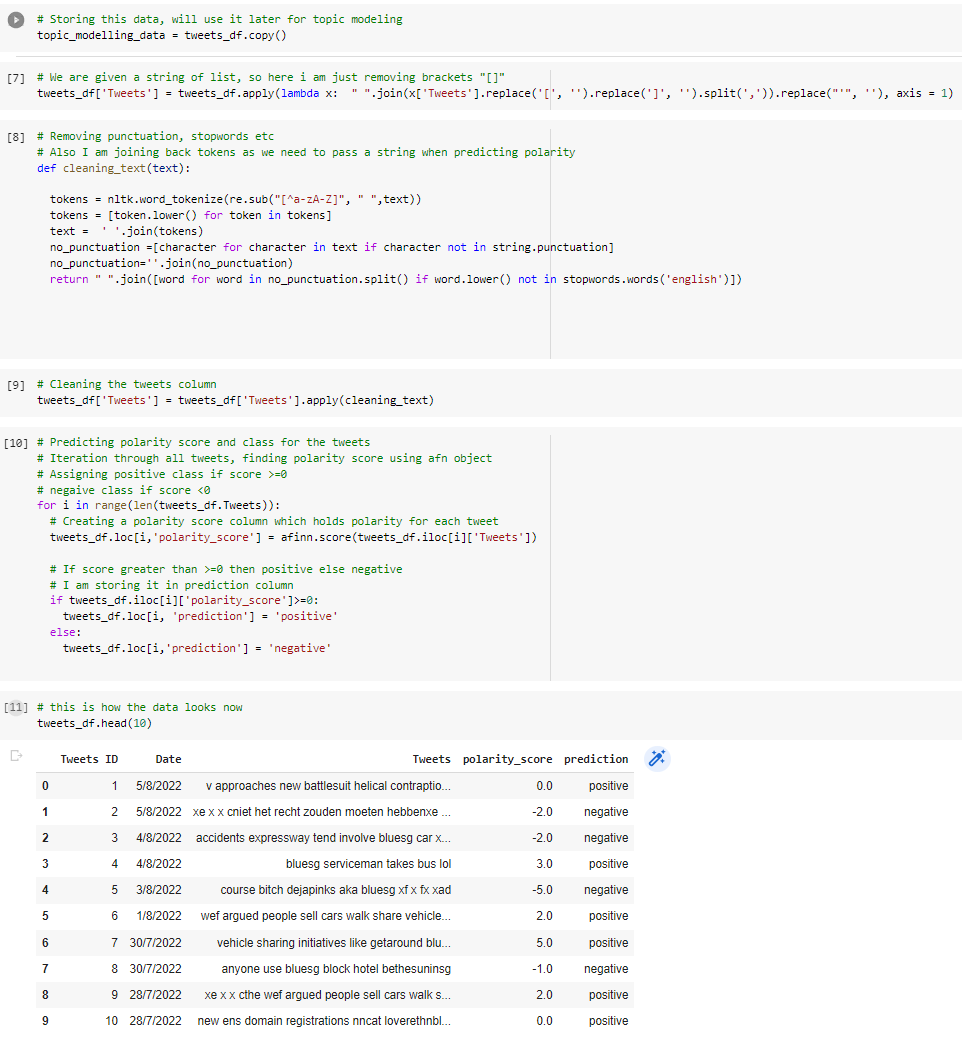
**Appendix A: Importing packages for Twitter scraping using Python**



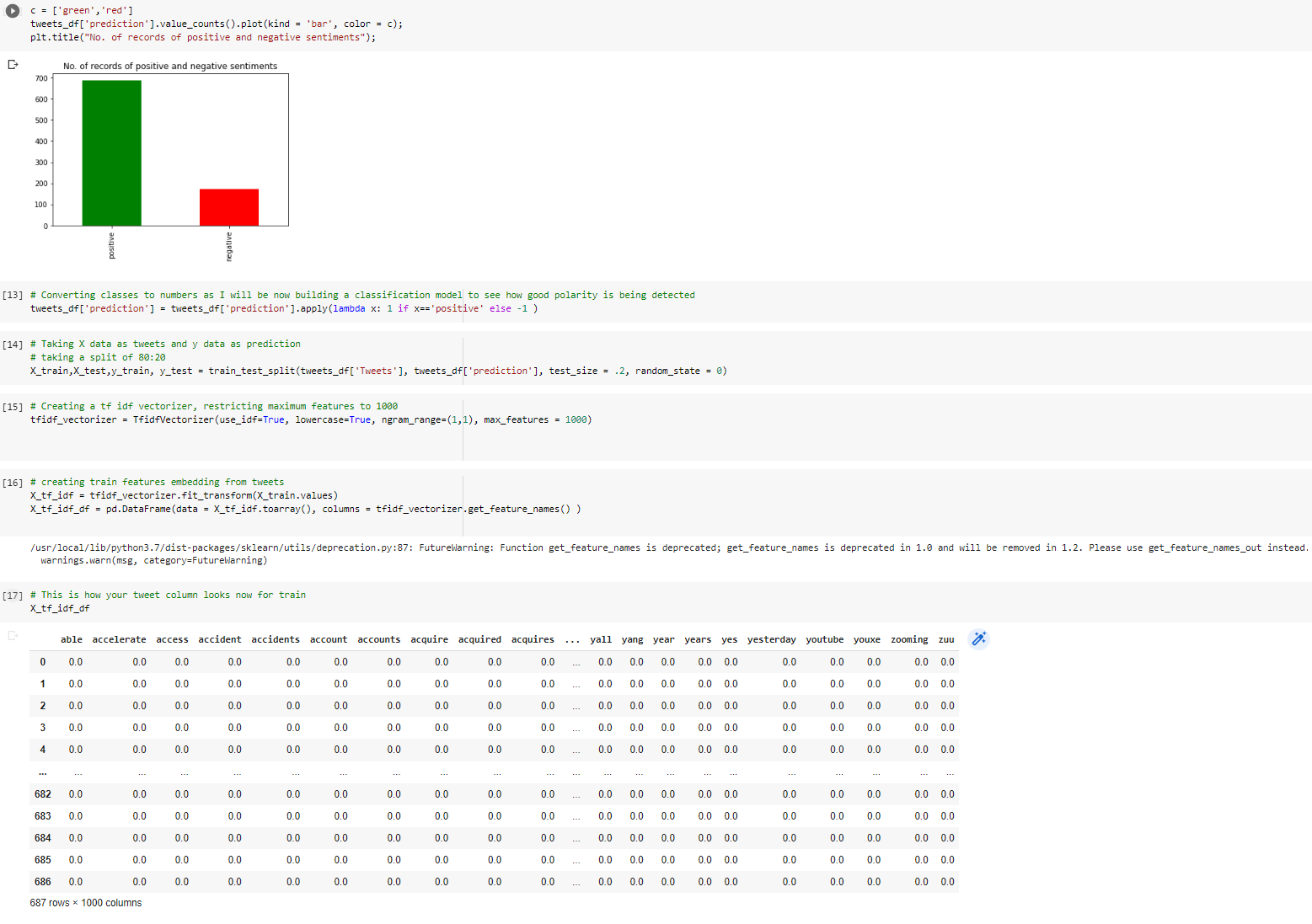
**Appendix B: Twitter scraping using Python**



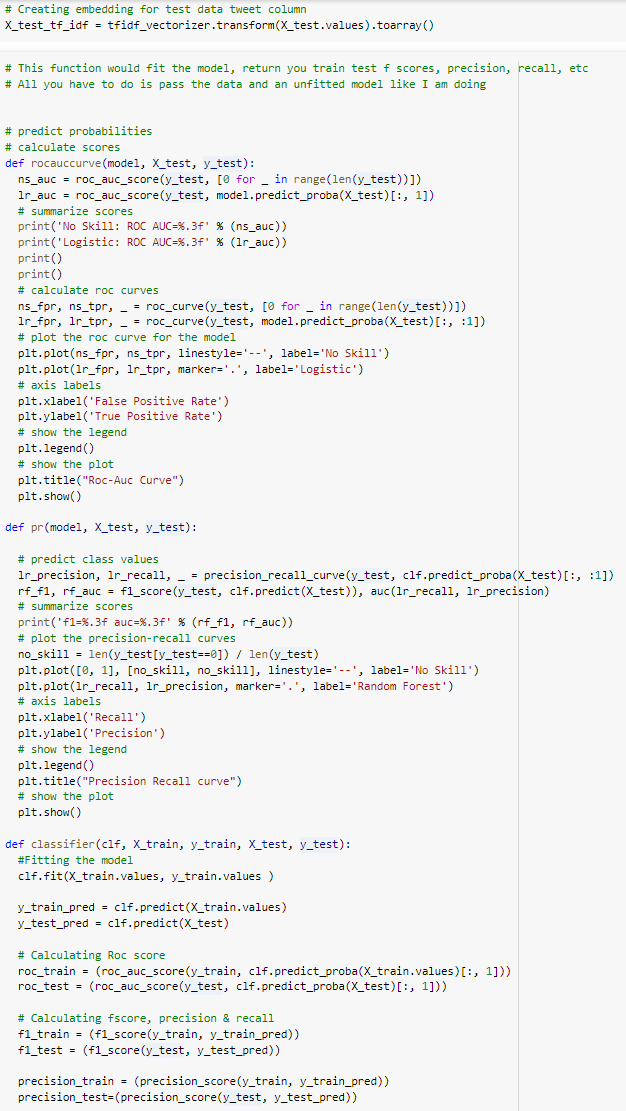
**Appendix C: Import python modules and reading data**

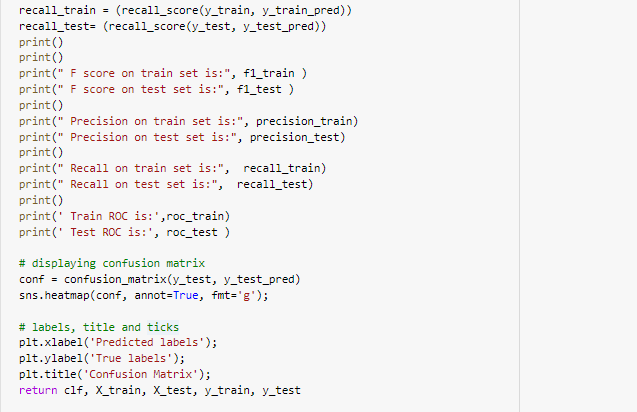


**Appendix D: Performing data cleaning**

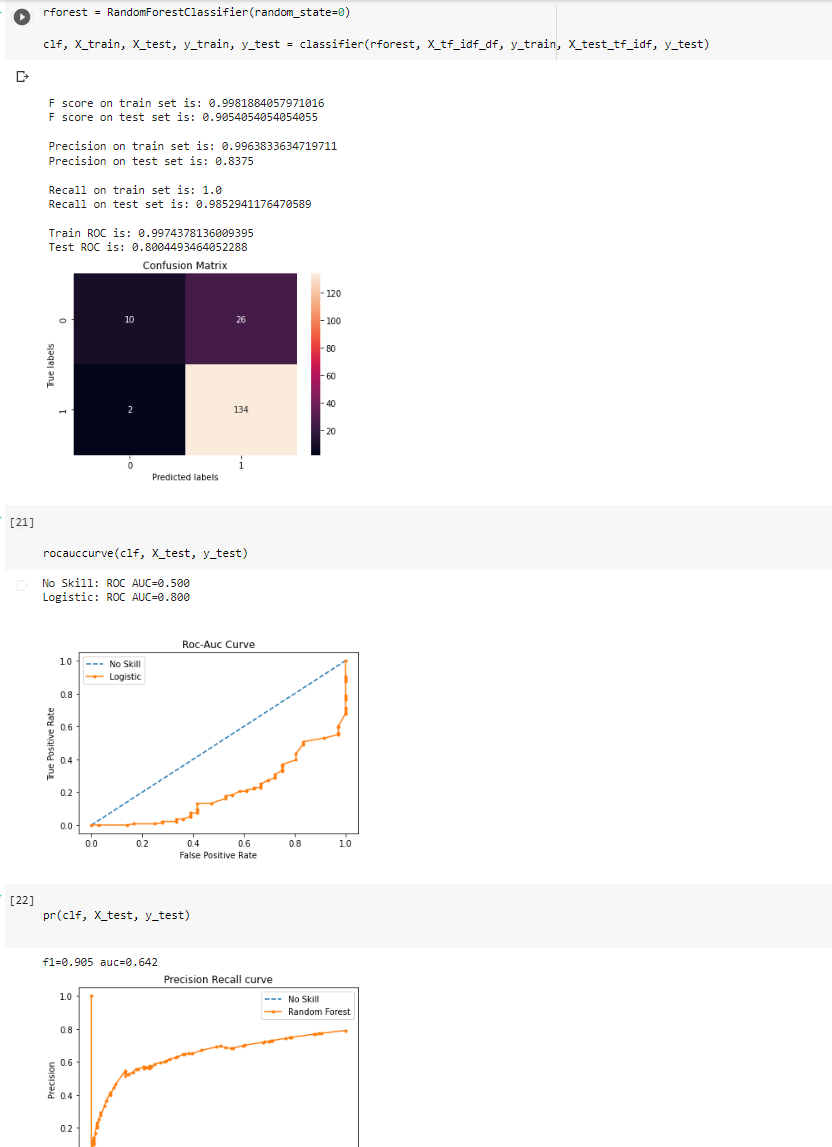


**Appendix E: Plotting sentiment distribution and creating test and train model**



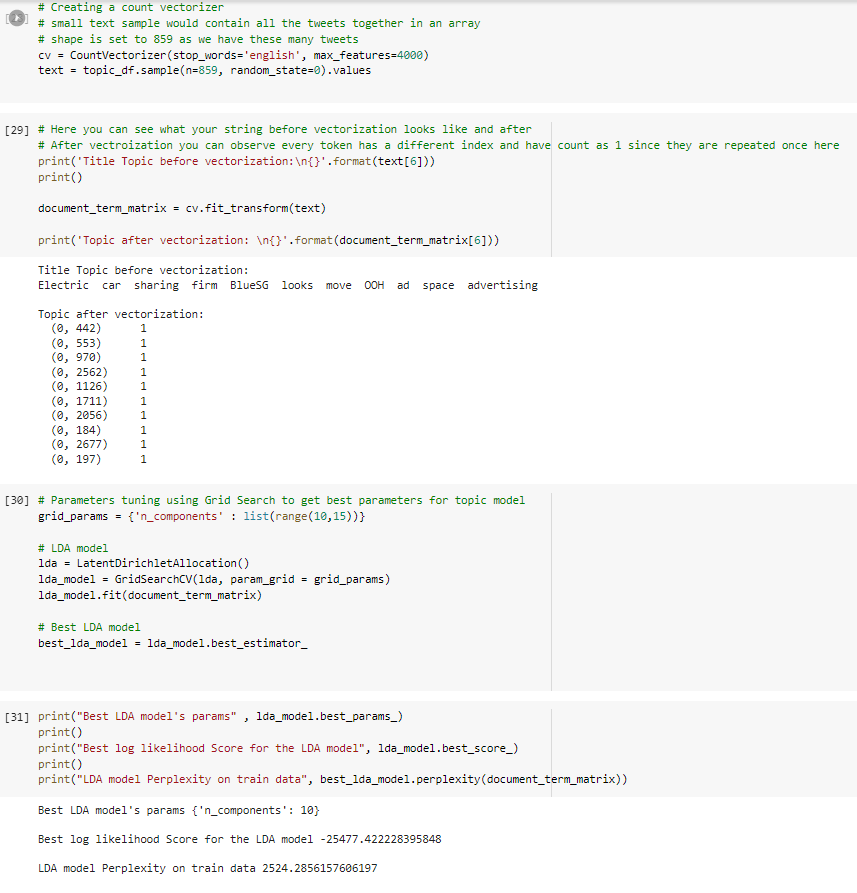


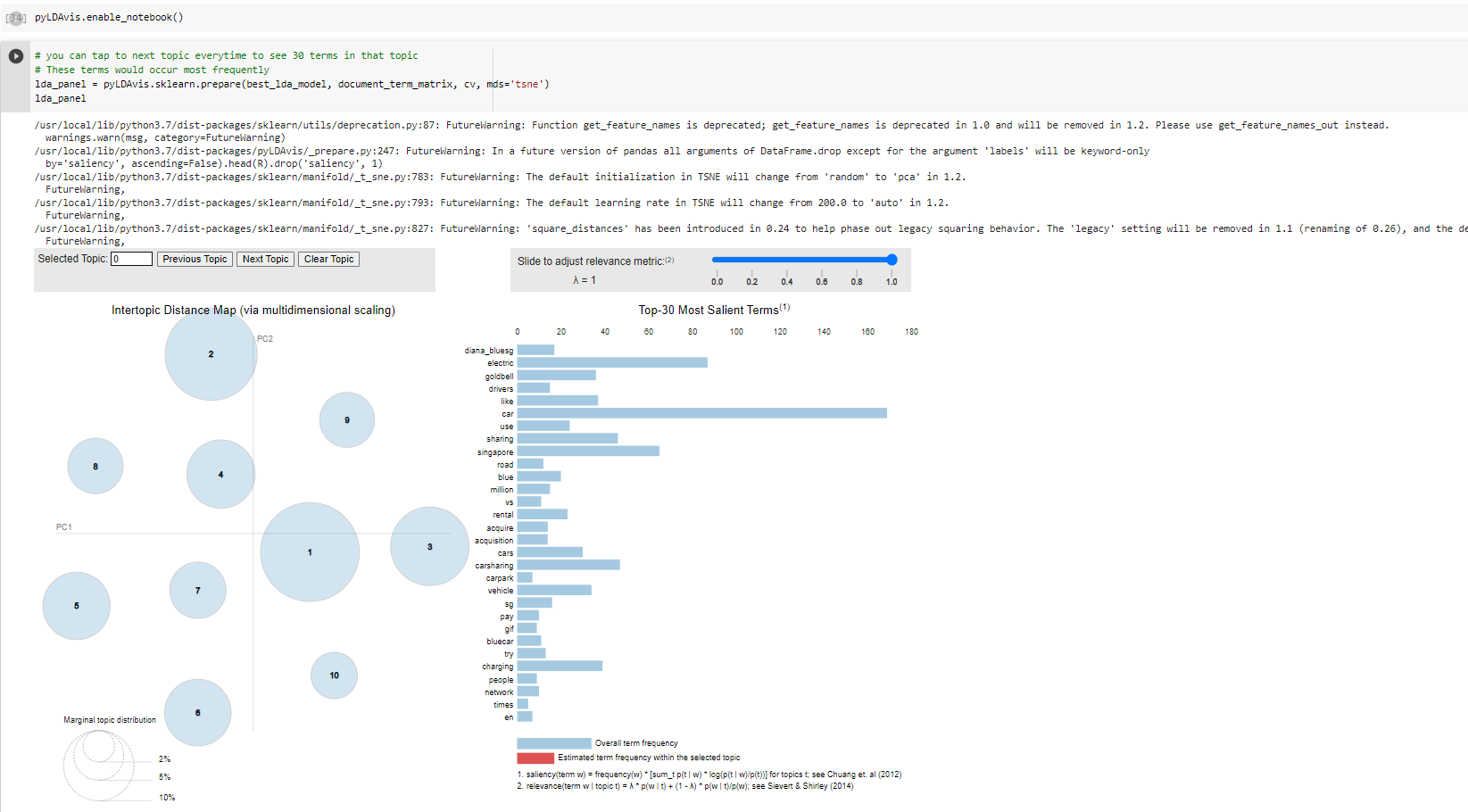
**Appendix F: Creating performance models**



**Appendix G: Confusion Matrix, ROC AUC Curve and Precision-Recall Curve**







**Appendix H: Plotting Topic Modeling Visualization**