Consumer Purchase Intention Prediction

***Synopsis Report submitted in partial fulfillment***

***of the requirement for the degree of***

**B. E.(Computer Engineering)**

Submitted By

Naresh Alwala

Sumith Pevekar

Aayush Pandey

Under the Guidance of

## Prof. Prakash Parmar

Department of Computer Engineering

A picture containing qr code

Description automatically generated

Vidyalankar Institute of Technology

Wadala(E), Mumbai 400 037

## University of Mumbai

# 2021-22

### CERTIFICATE OF APPROVAL

**For**

**Project Synopsis**

This is to Certify that

Naresh Alwala

Sumith Pevekar

Aayush Pandey

Have successfully carried out Project Synopsis work entitled

## Consumer Purchase Intention Prediction

in partial fulfillment of degree course in

Computer Engineering

As laid down by University of Mumbai during the academic year

2020-21

Under the Guidance of

## Prof. Prakash Parmar

Signature of Guide Head of Department

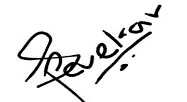
Examiner 1 Examiner 2 Principal

**I. Declaration**

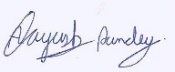
We declare that this written submission represents our ideas in our own words and where others' ideas or words have been included, we have adequately cited and referenced the original sources. We also declare that we have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in our submission. We understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

Name of student Roll No. Signature



1) Naresh Alwala 18102B0021

2) Sumith Pevekar 18102B0031

3) Aayush Pandey 18102B0032

Date:

II. Acknowledgements

Firstly, we would like to express our sincere gratitude to several individuals and organization for supporting us to work on such an interesting and challenging project. We wish to express our sincere gratitude to our supervisor, Professor Prakash Parmar, for his enthusiasm, patience, insightful comments, practical advice and unceasing ideas that have helped us tremendously in our research on the project. His immense knowledge, profound experience and professional expertise in data analysis and algorithms has enabled us to understand the different machine learning algorithms for this project.

We also wish to express our sincere thanks to Vidyalankar Institute of Technology for giving us the opportunity to research and develop the project of our interest, and also aided us with all the necessary resources required to work on this project.

III. Abstract

Recently, there has been a significant rise in the ecommerce industry and more specifically in people buying products online. There has been a lot of research being done on figuring out the buying patterns of a user and more importantly the factors which determine whether the user will buy the product or not. In this study, we will be researching on whether it is possible to identify and predict the purchase intention of a user for a product and target that user towards the product with a personalized advertisement or a deal. Further, we wish to develop a software that will help the businesses identify potential customers for their products by estimating their purchase intention in measurable terms from their tweets and user profile data on twitter. After, performing the initial analysis on the twitter data we have found that customers that show purchase intentions have some common keywords in their tweets. We need to apply some text analytical models to tweets data in order to predict purchase intention of users based on their tweets.

IV. Table of Contents

|  |  |  |
| --- | --- | --- |
| **Sr No** | **Description** | **Page No** |
| 1 | Introduction | 8 |
| 2 | Aim and Objective | 8 |
| 3 | Literature Survey | 9 |
| 4 | Problem Statement | 10 |
| 5 | Proposed Solution | 10 |
| 6 | Methodology | 11 |
| 6.1 | Data Collection | 11 |
| 6.2 | Exploratory data analysis | 11 |
| 6.3 | Data preparation | 11 |
| 6.4 | Text preprocessing | 12 |
| 6.5 | Text visualization | 12 |
| 6.6 | Data training | 13 |
| 6.7 | Tweet scraping | 14 |
| 6.8 | User Interface | 15 |
| 7 | Result and Outcomes | 16 |
| 8 | Conclusion | 17 |
| 9 | Future Scope | 18 |
| 10 | Hardware and software | 19 |
| 11 | References | 21 |

1. Introduction

There have been several research studies for analyzing the insights of online consumers buying behavior. However, only a few have addressed the customers buying intention for products. We want to develop a machine learning approach that will identify potential customers for a product by estimating the purchase intention in measurable terms from tweets on twitter. We have used a text analytical machine learning approach because although text analytics can be performed manually, it is inefficient. By using text mining and natural language processing algorithms it will be much faster and efficient to find patterns and trends. In a way we can say that Purchase Intention detection task is close to the task of identifying wishes in product reviews.

1. Aim and Objective

Purchase intentions are frequently measured and used by marketing managers as an input for decisions about new and existing products and services. Up till now many companies still use customer survey forms in which they ask questions like how likely you are to buy a product in a given time frame and using that information they calculate the purchase intention. We want to see if we can use Twitter tweets to train a model to identify tweets which show purchase intention for a product.

1. Literature Survey

There have been several research studies for analyzing the insights of online consumers buying behavior. However, only a few have addressed the customers buying intention for products. These include suggestions for a product or a desire to buy a product. They used linguistic rules to detect these two kinds of wishes. Although rule-based approaches for identifying the wishes are effective, but their coverage is not satisfactory, and they can’t be extended easily. Purchase Intention detection task is close to the task of identifying wishes in product reviews. Here we don’t use the rule-based approach, but we present a machine learning approach with generic features extracted from the tweets.

Past studies have shown that it is possible to apply Natural Language Processing (NLP) and Named Entity Recognition (NER) to tweets. However, applying NER to tweets is very difficult because people often use abbreviations or (deliberate) misspelled words and grammatical errors in tweets. Other studies used these techniques to apply sentiment analysis to tweets. The first studies used product or movie reviews because these reviews are either positive or negative. These studies merely analyze the sentiment of a tweet about a product after the author has bought it.

Moreover, research articles that we found online like *Identifying Purchase Intentions by Extracting Information from Tweets* ( February 8, 2017, RADBOUD U NIVERSITY NIJMEGEN) and *Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process* (The Berkeley Institute of Design) investigate if an artificial intelligence approach can predict (from existing user created content on twitter) if someone is a potential customer for a specific company or product and identify users at different stages of the decision process of buying a given product. Further looking at research reports online of different universities and researcher, give us an insight of the impact of social network marketing on consumer purchase intention and how it is affected by the mediating role of consumer engagement.

Some of NLP libraries for preprocessing techniques commonly used for twitter data are the TweetNLP library (a tokenizer, a part-of-speech tagger, hierarchical word clusters, and a dependency parser for tweets), unigrams, bigrams and stemming. There are also some dictionary-based approaches such as using the textBlob library (TextBlob is a Python (2 and 3) library for processing textual data. It provides a consistent API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, and more).

The common machine learning algorithms that we are planning to use for text analysis are Random Forest, Naive Bayes, Support Vector Machine and K-nearest neighbor algorithm.

1. Problem Statement

Marketing managers regularly assess and use purchase intentions as a factor in making decisions regarding new and existing products and services. Many organizations still utilize customer survey questionnaires that ask questions like "How likely are you to buy a product in a certain time frame?" and then compute the purchase intention based on that information. We want to explore if we can train a model to recognize tweets that indicate a product purchase intention using Twitter tweets.

1. Proposed Solution

There has been several researches made by different institution on the same field by using NLP sentiment analysis that merely finds the sentiment o f a tweet. However, there are several limitation related to it like grammatical errors and complex sentence formations. We are planning to use the dataset of about 3200 tweets that has been scraped from the twitter using web crawler software and official twitter API. After scraping out the tweets we will manually annotate the tweets to classify them as purchase intent tweet or not and then store it csv formatted file. Once the dataset is ready we will preprocess it using different NLP libraries to tokenize each tweet and also perform stemming to reduce each word to its word stem. After the preprocessing task, we will train the text analytical models like SVM, Random Forest, etc algorithms to predict the class of the tweets that it will read from the twitter and finally we will also prepare a dashboard to show the list of all potential users and provide graphical overview.

1. Methodology
2. Data Collection

The tweets containing products name were gathered from the twitter using Twitter API to create the dataset. There was no publicly available dataset which met our purpose. So, we collected data that showed the purchase intention for products.

1. Exploratory Data Analysis

Exploratory data analysis, often known as E.D.A., is a key phase in exploring and investigating diverse data sets and summarizing their significant properties, sometimes employing various data visualization approaches. It makes it easier for a data analyst to see trends, spot abnormalities, test theories, and draw conclusions about the best approach to monitor data sources to generate more precise answers

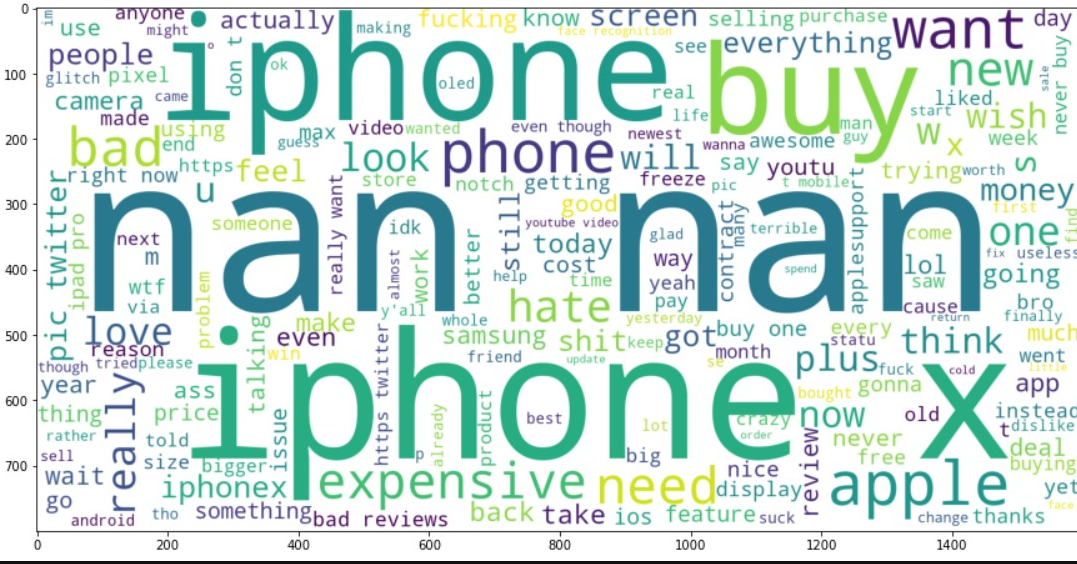
1. Data Preparation

Data pre-processing is a technique for converting raw data into a usable and efficient format. Training a model with a dataset which is not pre-processed can lead to poor results, therefore Data pre-processing is an important step in machine learning algorithms. In this we have removed unwanted columns to reduce dimensionality. Further we have removed the tuples having undefined class label and normalized the class labels.

1. Text Preprocessing

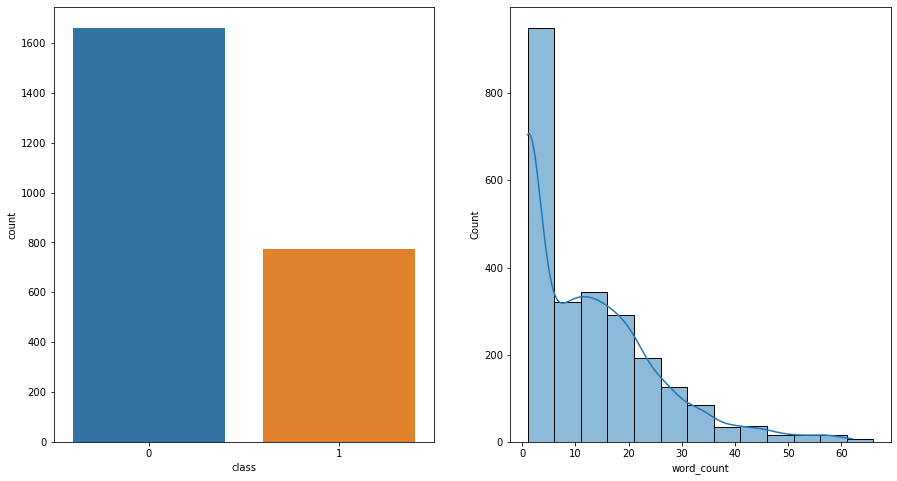
Simply said, pre-processing your text implies converting it into a format that is predictable and analyzable for your task. General tweets contain various symbols like ‘@’, etc. Such symbols are removed before tokenizing and all bad symbols are removed. Stop words are also removed from the dataset. Next, we have tokenized the text since neural networks work by performing computation on numbers, passing in a bunch of words won't work. This data can now be used for training the model as well as visualization.



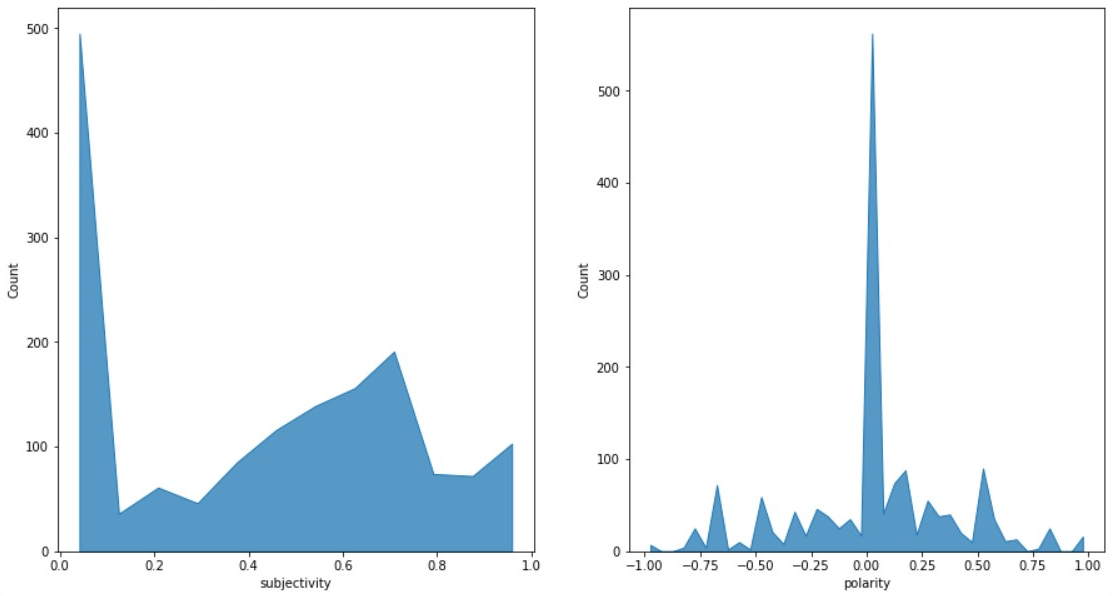


1. Text Visualization

Text visualization is the process of extracting information from the raw text and perform different analysis to derive important structure and pattern out of the information. This is where modern natural language processing (NLP) tools come in. They can capture prevailing moods about a particular topic or product (sentiment analysis), identify key topics from texts (summarization/classification), or amazingly even answer context-dependent questions (like Siri or Google Assistant). Their development has provided access to consistent, powerful, and scalable text analysis tools for individuals and organizations.

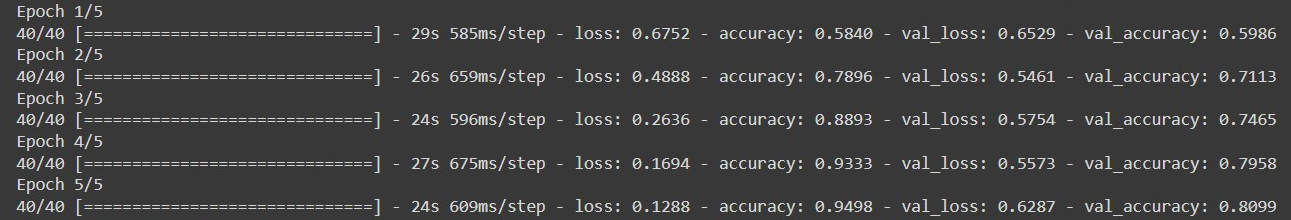


For this project we used seaborn library along with Word Cloud to visualize the sentiment polarity as well as subjectivity of tweets in dataset. We used Text Blob library to measure the sentiment and subjectivity count of each tweet.



1. Data Training

Machine Learning algorithms learn from data. They find relationships, develop understanding, make decisions, and evaluate their confidence from the training data they’re given. And the better the training data is, the better the model performs. We also normalized the data using Minmax scaler to evenly distribute the datapoint values around mean value. Finally, after normalization we split the dataset into training and testing in the ratio of 4:1 respectively and used the Keras Long Short-Term Memory (LSTM) model which is a recurring neural network. For LSTM model we added 3 layers using softmax activation at output layer.



1. Tweet Scraping

The Twitter API is a set of programmatic endpoints that can be used to understand or build the conversation on Twitter.

This API allows you to find and retrieve, engage with, or create a variety of different resources including the following:

* Tweets
* Users
* Spaces
* Direct Messages
* Lists
* Trends
* Media
* Places

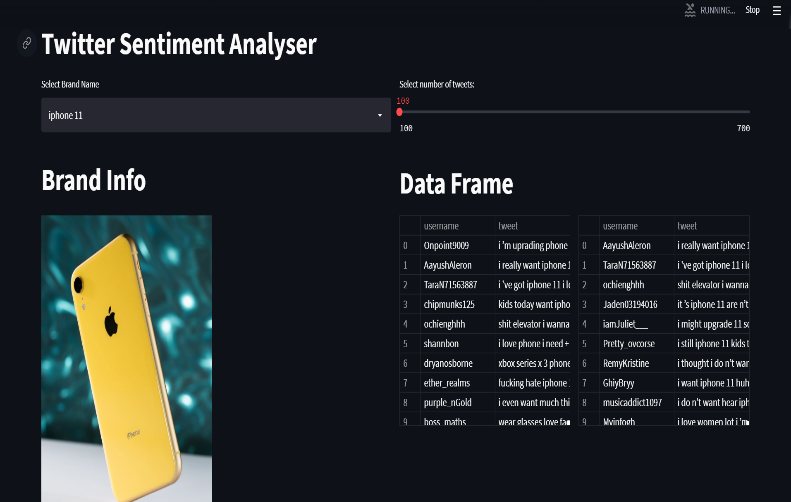
We have used Tweepy to scrape the tweets using Twitter API.

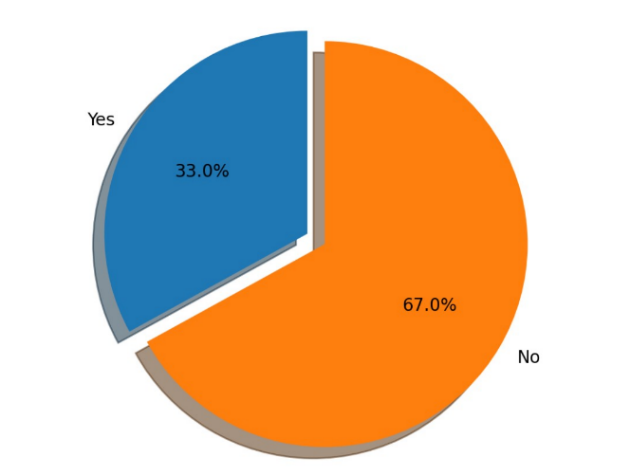
[Tweepy](https://github.com/tweepy/tweepy) is an open-source Python package that gives us a very convenient way to access the Twitter API with Python. Tweepy includes a set of classes and methods that represent Twitter’s models and API endpoints, and it transparently handles various implementation details, such as:

* Data encoding and decoding
* HTTP requests
* Results pagination
* OAuth authentication
* Rate limits
* Streams

Almost all the functionality provided by Twitter API can be used through Tweepy.

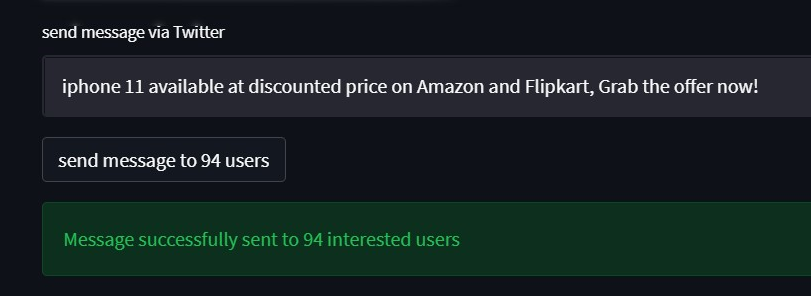
1. User interface
2. Firstly, the user will open the DASHBOARD and see the status of the product through charts and the relevant tweets for the product in a list.



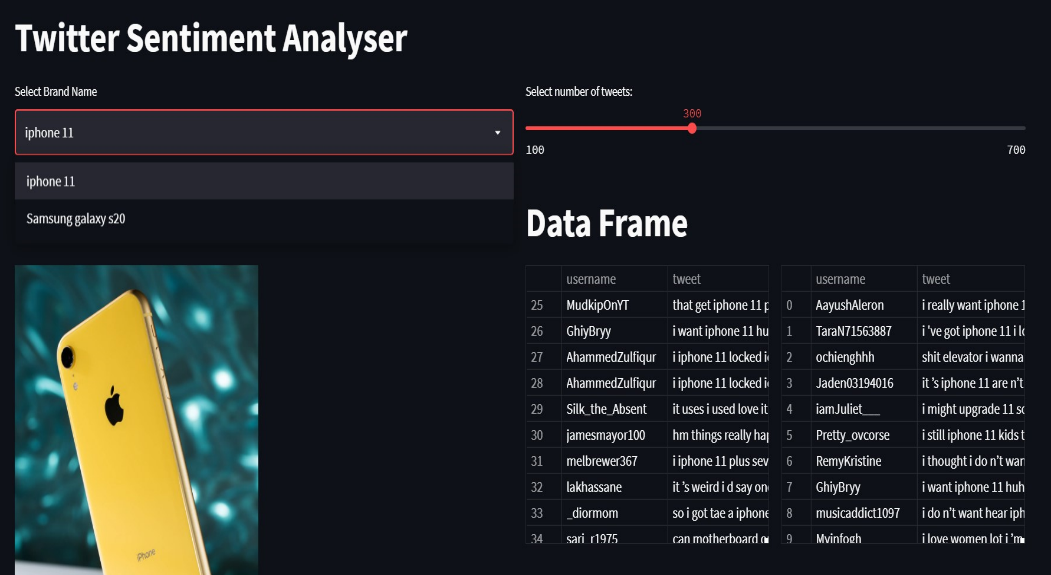
1. Secondly, the user can select the model of smartphone he/she needs to analyze from the dropdown menu and select the number of tweets according (default 100 tweets) from the slider. The time complexity of algorithm increases proportionally with the number of tweets and get the analysis report along with intended user’s live location on map.

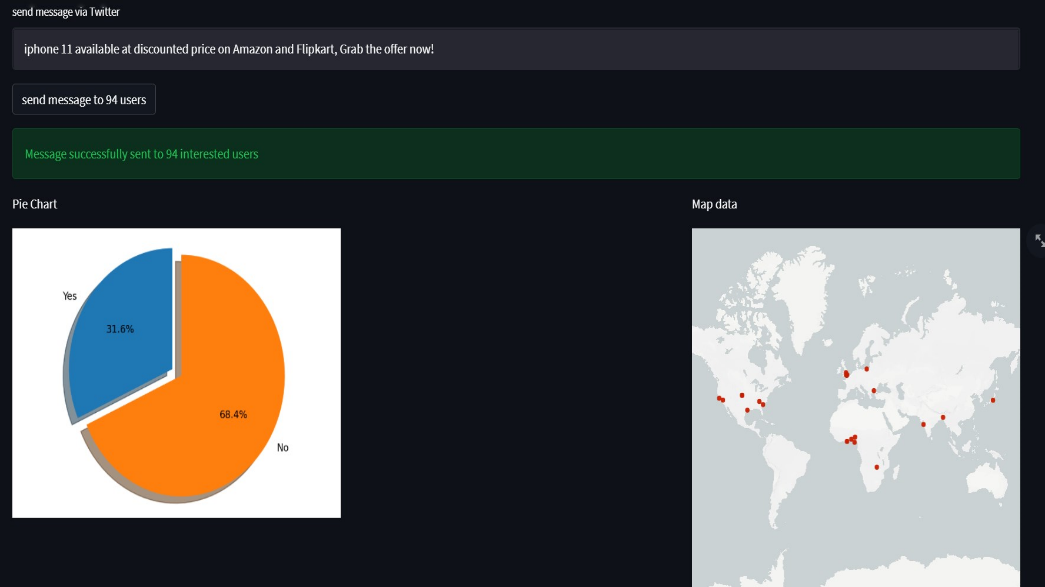


1. Thirdly, after analysis and scraping the intended users user can also send a direct message to the intended users via Twitter using the send button below.



Images of user interface:





1. Result

Since we are using a machine learning (artificial intelligence) based approach we needed to set an accuracy standard for our model and evaluate the results by matching the desired standard.

To evaluate our models, we used the following techniques:

1. Accuracy
2. Precision

3. Recall

4. F-Measure

After evaluating our model here are the following results that we have gotten: After performing all the Data preprocessing this is the results that we got:

* + SVM Model

|  |  |
| --- | --- |
| SVM Model | |
| Accuracy | 0.8066 |
| Recall | 0.7343 |
| Precision | 0.7230 |
| F-Score | 0.7286 |

* + Bi-LSTM Model

|  |  |
| --- | --- |
| Bi-LSTM | |
| Accuracy | 0.7560 |
| Recall | 0.7366 |
| Precision | 0.7572 |
| F-Score | 0.7461 |
|  |  |

* + LSTM Model

|  |  |
| --- | --- |
| LSTM | |
| Accuracy | 0.8354 |
| Recall | 0.8354 |
| Precision | 0.8343 |
| F-Score | 0.8347 |

8. Conclusion

Our results were quite promising since we had created our own dataset and were building the model from scratch. We had to create our own dataset because there does not exist a publicly available dataset for purchase intention based on twitter tweets.

The 2 major problems that we faced were:

1. The imbalance class problem: Since our dataset was manually annotated by us, we had about 2000 positive tweets and 1200 negative tweets. Due to this we were getting a very low True Negative Rate and our model was not accurately predicting the negative class.
2. Limited annotated data: Since we had to manual annotate each tweet in the dataset and this process takes a lot of time, we were only able to annotate about 3200 tweets.

We were not able to get more than 90% accuracy because of the two problems highlighted above. To achieve even 90% accuracy with an imbalance class data and such a small dataset is a victory.

Finally, the user will be able to analyze the tweets and get the list of intended users along with their live location and send them any offer or advertisement to them via Twitter.

## 9. Future Scope

Till date we have successfully created our own database for the model using the web crawler software to extract the tweets from twitter and annotated the data by classifying the tweets as a purchase intent tweet or not. We had to create our own dataset because there does not exist a publicly available dataset for purchase intention based on twitter tweets. Also, we have implemented 2 deep learning models with an average accuracy of 76.9%. The future of trained sentiment analysis will delve deeper beyond the concept of classifying the customer into different segments. As a result, these models are becoming necessary for the businesses to survive in a competitive market.

10. Hardware and software requirement

1. Hardware:-
2. Processor - Intel(R) Core(TM) i5-8265U CPU @ 1.60GHz 1.80 GHz
3. Motherboard – ASRock EPC612D8A
4. RAM – 8 GB SATA DDR4 2133 MHz
5. 1 TB Hard Disk (7200 RPM) + 256 GB SSD
6. GPU – Intel UHD (2GB)
7. Software:-
8. Jupyter lab
9. TensorFlow
10. Scikit learn
11. Keras
12. Python

11. References

1. Websites:
   1. <https://scikit-learn.org/stable/>
   2. <https://github.com/jonbakerfish/TweetScraper>
   3. <https://monkeylearn.com/blog/sentiment-analysis-of-twitter/>
2. Relevant Papers
   1. Identifying Purchase Intentions by Extracting Information from Tweets, February 8, 2017, RADBOUD U NIVERSITY NIJMEGEN, BACHELOR ’S THESIS IN ARTIFICIAL INTELLIGENCE.
   2. Tweetalyst: Using Twitter Data to Analyze Consumer Decision Process, The Berkeley Institute of Design.
   3. Mining Purchase Intent by Dublin City University, School of Computing, ADAPT Centre, Dublin, Irelan