VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi: 590 014, Karnataka.

A Project Phase-1 Report on

"User Interest Based Social Media Data Retrieval System"

Submitted in partial fulfilment for the award of degree of
Bachelor of Engineering
In
Computer Science & Engineering

Submitted by

ANUDEEP BABU K (3SL19CS006)

BASANA GOWDA (3SL19CS011)

RAGHAVENDRA D (3SL19CS035)

VINODKUMAR SANKRANTHI (3SL19CS055)

Under the guidance

PROF. SURESH PATEL

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

H.K.E. Society's S.L.N. COLLEGE OF ENGINEERING

(Affiliated to VTU - Belagavi, Affiliated to AICTE – New Delhi, Accredited by NAAC)
Y-Camp, Raichur-584 135, Karnataka

2022-23

VISVESVARAYA TECHNOLOGICAL UNIVERSITY

"Jnana Sangama", Belagavi: 590 014, Karnataka.

H.K.E. Society's S.L.N. COLLEGE OF ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

CERTIFICATE

Certified that the project phase-1 report entitled **"User Interest Based Social Media Data Retrieval System"** is a bonafide work carried out by ANUDEEP BABU K (3SL19CS006), BASANA GOWDA (3SL19CS011), RAGHAVENDRA D (3SL19CS035), VINODKUMAR SANKRANTHI (3SL19CS055) has been successfully presented at **H.K.E.SOCIETY's S.L.N. COLLEGE OF ENGINEERING** in partial fulfilment for the award of degree of Bachelor of Engineering in Computer Science Engineering of Visvesvaraya Technological University, Belagavi during the academic year **2022- 23**. It is certified that all the corrections / suggestions indicated for the internal assessments have been incorporated in the project phase-1 report deposited in the department. This project report has been approved as it satisfies the academic requirements in respect of project work prescribed for the Bachelor of Engineering Degree.

Signature of the Guide	Signature of the Coordinator	Signature of the HOD
Prof. Suresh Patel	**Prof. Sujatha J.**	**Prof. Sumangala Itagi**

ACKNOWLEDGEMENT

This satisfaction and euphoria that accompany the successful completion of any task would be but incomplete without the mention of the people who made it with constant guidance and encouragement and crowned our effort with success.

We would like to express our gratitude to our Principal DR. R. BASAWARAJA for providing congenial environment and surroundings to work

A heartly thanks to our beloved HOD Prof. SUMANGALA I, for her encouragement and support.

We express our sincere thanks to project guide Prof. SURESH PATEL for his constant encouragement and support throughout our course 'especially for the useful suggestions given during the course of the project period.

We also thank coordinator Prof. SUJATHA J. and all the staff members of department of Computer Science and Engineering and all those who have directly or helped us with valuable suggestions in the successful completion of this project.

Last but not the least we would like to thank our beloved parents for their blessing, love, and encouragement to successfully complete the task by meeting all the requirements.

ANUDEEP BABU K (3SL19CS006)

BASANA GOWDA (3SL19CS011)

RAGHAVENDRA D (3SL19CS035)

VINODKUMAR SANKRANTHI (3SL19CS055)

TABLE OF CONTENTS

Sl.NO	TOPICS	PAGE NO
1	Abstract	1
2	Introduction	2
3	Literature Survey	3
4	Objectives	6
5	Existing System	7
6	Proposed System	9
7	Problem Statement	12
8	System Architecture	13
9	H/W & S/W Requirements	19
10	Conclusion	20
11	References	21

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 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 applying various text analytical models to tweets data, we have found that it is indeed possible to predict if a user have shown purchase intention towards a product or not, and after doing some analysis we have found that people who had initially shown purchase intention towards the product have in most cases also bought the product.

Digital marketing is taken into account the well-liked method comparing to traditional marketing. It can be used by both researchers and academicians for social media marketing and to predict the customers purchase intention. The Proposed work revolves around some valuable information and processes in accordance to the behavior of customer during the online purchase. Business owners, scientists, researchers all post their ads, details on the Web so that they can be linked to owners quickly and easily by web scrap searching on searchable product websites to gain a lot of data from websites. Hence, customer price and rating of product evaluation and prediction has become an important research area. The analysis is done by Support Vector Machine (SVM- Linear) to gather several information and provide variation analysis. The major goal remains to investigate and analyze the extracted dataset using ML oriented algorithms with best accuracy possible. The analysis has a proper path to sentimental analysis of parameters in accordance to the ratings and price of the product to find proper accurate calculations.

1.INTRODUCTION

Our project is a web application that predicts the likelihood/certainty that a customer will buy a product that he is interested in based on his social media posts such as Twitter tweets and user profile data. This will help the company/business target a particular customer more efficiently and boost their sales.

First, we search for Twitter tweets of potential customers wanting to buy a product. And based on those tweets we estimate/predict the likelihood that the customer will buy the product. We then make a model by gathering tweets from users who have already expressed intention to buy the product using their tweet history and if possible, their web search history as well and then training the text analytical model based on those tweets. Using the model, we input potential customers who have tweeted about the product but have not bought it. And based on the training data the model estimates a prediction/likelihood of whether the customer will buy it or not. We have limited the scope of our work to only mobile phones. Our model predicts the consumer intention for the latest upcoming mobile phones. We will be testing it on the latest iPhone X variants and check with its accuracies.

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.

2.LITERATURE REVIEW

[1] Our inference out of first paper:

- Explains about Social platforms like Twitter and Facebook allow users to share and publish information with their users. In addition to this, users use them to answer very precise and highly contextualized queries, or queries for which the relevant content has not been authored yet, e.g., asking about a conference event using its hashtag on Twitter.
- Even if the profile is private, we can still derive some attributes. Also, when we tried as an external user to extract profiles from different online social network e.g. Twitter and Facebook , either no data or minimal data was made available
- Data extraction is a field that is concerned with grabbing information from different web resources including websites, online databases and services. It is necessary to find tools for data extraction because of the dynamic nature of the World Wide Web
- This creates some difficulties for end users and application programs when it comes to finding useful data.
- They are stricter in terms of having well-formed documents i.e. the documents' contents should conform to their syntax rules. This feature helps the parsers of search engines to interact with the web pages' contents more efficiently.
- The common machine learning algorithms that are used for text analysis are Linear Regression, Random Forest, Naive Bayes and Support Vector Machine. We will be looking at these models later in detail.
- 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. Studies on identification of wishes from texts, specifically Ramanand et al. (Ramanand, Bhavsar, and Pedanekar 2010) consider the task of identifying 'buy' wishes from product reviews
- These wishes 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.
- These studies merely analyze the sentiment of a tweet about a product after the author has bought it. We will however be extracting features from tweets to find whether the user has shown purchase intention towards the product or not.
- . The first studies used product or movie reviews because these reviews are either positive or negative.

[2] Our inference out of second paper:

- The common machine learning algorithms that are used for text analysis are Linear Regression, Random Forest, Naive Bayes and Support Vector Machine. Applying Named Entity Recognition (NER) and Natural Language Processing (NLP) to tweets is very difficult because people often use abbreviations or deliberately misspelled words and grammatical errors in tweets.
- Some preprocessing techniques commanly used for twitter data are the sentiment140 API (Sentiment140 allows you to discover the sentiment of a brand, product, or topic on Twitter), 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).
- Further looking at research reports like The Impact of Social Network Marketing on Consumer Purchase Intention in India: Consumer Engagement as a Mediator (Asian Journal of Business and Accounting 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. Based on UGT theory (Uses and Gratification Theory).

- Social platforms like Twitter and Facebook allow users to share and publish information with their friends and often with general public. In addition to this, users use them to answer very precise and highly contextualized queries, or queries for which the relevant content has not been authored yet, e.g., asking about a conference event using its hashtag on Twitter. We refer to such a process of finding information as Social search, and we define it as follows.
- Social platforms like Twitter and Facebook allow users to share and publish information with their friends and often with general public. In addition to this, users use them to answer very precise and highly contextualized queries, or queries for which the relevant content has not been authored yet, e.g., asking about a conference event using its hashtag on Twitter. We refer to such a process of finding information as Social search, and we define it as follows.
- The social search is associated to platforms that are defined as search engines specifically dedicated to social data management such as Facebook. The main ingredient to perform a social search is the user interactions.
- Social platforms allow users to provide, publish and spread information, e.g., commenting or tweeting about an event. In such a context, a huge quantity of information is created in social media.
- They do not suppose any additional structure, including the social network that surrounds users. This does not reflect the real behaviors of users, since they normally ask friends for recommendations before acting

3.OBJECTIVES

- To create Web application we need to create various Dashboards.
- To update or to test the data we need to have the annoted Data.
- The annoted data is the collect the twitter data in the form of csv file.
- based on the training data, the model will estimate a prediction/likelihood of whether the customer will buy it or not.
- We are using the following text analytical models:
    1. Support Vector Machine (SVM)
    2. Naive Bayes
    3. Logistic Regression
    4. Decision Tree
    5. Neural Network
- Positive reviews and comments generated by customers on social networking platforms can help improve business sales and profitability.
- This social media data retrival system enables companies to reach out to new and existing information about there products.
- By erasing the physical and spatial boundaries between people, social networking websites can provide instant reachability.
- Organizations and businesses can use social networking to build a following and review their product globally.

4.EXISTING SYSTEM

The popularity of online social networking sites has increased the amount of personal data which is distributed on the net. This is supported by the fact that social networking sites have overtaken email in terms of usage. Online social networking sites contain user profiles which consist of personal data. Those profiles are semi-structur

ed and the profile data or structure may change in an unpredictable way. This fits in nicely with the way online social networks operate. Social network profiles change all the time not just in structure but content as well. More research needs to be done into the extraction from semi-structured pages in terms of online social networking profiles.

The motivation for this paper is that as far as the paper authors' know there has been little research associated with automated extraction methods from semi-structured web pages from online social networks. Our goal is to extract the relevant profile data so it can be mined in the future to find attributes that can cause the profile owner to be vulnerable to social engineering attacks. In terms of online social networks our research will allow us in the future to investigate the friends of a profile and see if any of the friends have profiles on other online social networks.

This links into the transitivity concept where e.g. A and B are friends on one online social network but B and C are also friends on another online social network. The question is: will A and C become friends and if so how great will the strength of their friendship be. The question posed fits with theory about weak ties and how they can provide an alternative information source to the ones associated with the strong ties.

Data extraction is a field that is concerned with grabbing information from different web resources including websites, online databases and services. It is necessary to find tools for data extraction because of the dynamic nature of the World Wide Web. This creates some difficulties for end users and application programs when it comes to finding useful data

Automated data extraction is just starting to be used in research about online social networking. Table 1 below illustrates some of the data extraction techniques used to extract attributes from online social networking profiles. It shows some data extraction methods ranging from manual through to automated methods. Our methodology outlines our approach to extracting attributes and a list of top friends from a MySpace profile. MySpace was chosen

because it allows a rich source of data to be derived from profiles without the need to be a member of MySpace. Even if the profile is private, we can still derive some attributes. Also, when we tried as an external user to extract profiles from different online social network e.g. Facebook and Friendster, either no data or minimal data was made available.

These collaborative tasks that make users more active in generating content are among the most important facators for the increasingly growing quantity of available data. In such a context, a crucial problem is to enable users to find relevant information with respect to their interests and needs.

Even if the profile is private, we can still derive some attributes. Also, when we tried as an external user to extract profiles from different online social network e.g. Facebook and Friendster, either no data or minimal data was made available.

Our process will comprise of the following components:

- Stage 1- Data Pre-processing involves analysis of the HTML structure of a given profile. The HTML content is parsed and a vector of tokens is produced. The extracted tokens help in the design of the tables in the repository. The reason of data pre-processing is because MySpace profiles have different structures and therefore different tokens.
- Stage 2- Specify the URL address of a profile. All social network profiles come with a unique profile URL address. The algorithm for extraction of the personal details involved developing and expanding the library provided.
- Stage 3- Check the user's stopping criteria. The user can specify whether they want to stop the extraction by the number of friends extracted (e.g. the first 100 friends) or by the level (e.g. 1 which is just the top friends of the specified profile) extracted
- Stage 5- Extract the relevant personal details from the profile and insert them into the repository. The repository used PostgreSQL 8.1.4. The repository structure has to be designed for the Breadth First Search algorithm.
- Stage 6- Extract list of friends and their profile addresses then insert them into the repository if they have not been stored before. In the case of this research paper, the extracted friends' lists just consist of the top friends who we assumed the user may have a strong affiliation with. The data in the repository can be used for data mining purposes in the future to find patterns.

5.PROPOSED SYSTEM

Proposed Approach

In this section, we describe the details of our approach to tackle the problem of purchase intention detection. We will begin by describing our data collection and annotation process. Then we will describe our approach for data preprocessing and transforming the data to train text analytical models.

Data collection and annotation

As there are no annotated Twitter tweets corpora available publicly for detection of purchase intent, we had to create our own. This was done using a web crawler developed by JohnBakerFish which crawled the website to collect the data. We had collected over 100,000 tweets but since they were not annotated, we had to cut down to just 3200 tweets which were randomly selected out of the dataset and we manually annotated them.

We used just 3200 tweets out of such a large dataset as we were limited by time. We defined definition of Purchase Intention as object that is having action word like (buy, want, desire) associated with it. Each tweet was read by 3 people and final class was decided by maximum voting.

Data preprocessing

- **Data preprocessing techniques:**

Next, we preprocessed the tweets using these techniques:

1. **LOWERCASE:** So, we started our groundwork by converting our text into lower case, to get case uniformity.
2. **REMOVE PUNC:** Then we passed that lower case text to punctuations and special characters removal function. Text may contain unwanted special characters, spaces, tabs and etcetera which has no significant use in text classification.
3. **STOPWORDS REMOVAL:** Text also contains useless words which are routine part of the sentence and grammar but do not contribute to the meaning of the sentence. Likes of “the”, “a”, “an”, “in” and etcetera are the words mentioned above. So, we do not need these words, and it is better to remove these.

4. **COMMON WORD REMOVAL:** Then there also lots of repetitive words which from their recurrence do not contribute to the meaning in the sentence. This can also be the result of mistake as the data we are analyzing is an informal data where formal sentence norms are not taken into consideration.
5. **RARE WORDS REMOVAL:** We also removed some rare words like names, brand words (not iphone x), left out html tags etc. These are unique words which do not contribute much to interpretation in the model.
6. **SPELLING CORRECTION:** Social media data is full of spelling mistakes. And it is our job to get rid of these mistakes and give our model the correct word as an input.
7. **STEMMING:** Then we stemmed the words to their root. Stemming works like by cutting the end or beginning of the word, considering the common prefixes or suffixes that can be found in that word. For our purpose, we used Porters Stemmer, which is available with NLTK.
8. **LEMMATIZATION:** Then we also performed lemmatization on our text. This analysis is performed in morphological order. A word is traced back to its lemma, and lemma is returned as the output.

After preprocessing the tweets, we are left with about 1300 tweets for training data and remaining for testing.

➢ Document Vector

Next, we made 3 types of document vectors:

1. **TF:** First is the term frequency document vector. We have stored text and its labeled class in data frame. And we have constructed a new data frame with columns as the words and document count as the rows. So, individual frequency of words in a document count is recorded.
2. **IDF:** It is a weighting method to retrieve information from the document. Term frequency and inverse document frequency scores calculated and then product of TF*IDF is called TF-IDF. IDF is important in finding how relevant a word is. Normally words like ‘is’, ‘the’, ‘and’ etc. have greater TF. So IDF calculated a weight to tell how important least occurring words are.
3. **TF-IDF with textblob library:** With the help of textblob library we calculated sentiments of individual word and then multiplied the sentiment score with TF and TF-IDF of that word.

➢ Algorithms

Once the corpus was ready, we then used different text analytical models to test which one gave the best results. We used the following models:

1. **Support Vector Machine (SVM):**

Simply put, SVM is a supervised machine learning algorithm which does complex transformation on the data. And then it tries to separate data on classes we have defined on our data.

2. **Naive Bayes:**

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

3. **Logistic Regression:**

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

4. **Decision Tree:**

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

5. **Neural Network:**

It is deep learning machine algorithm, which is arranged in a layer of neuron. There is an input layer, output layer and hidden layers of neurons. Neuron network is adaptive as neurons in these layers learn from their initial input and subsequent runs.

6.PROBLEM STATEMENT

- Problem

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.

- Complexity

The complexity of our approach is that we have to calculate how to measure the purchase intention from a tweet. Exploring the different type of text analytical methods and choosing the best one for our task will be quite challenging. Measuring the results of our machine learning model and then deciding the best one will involve a lot of factors which we will have to calculate.

- Motivation

We want to develop a machine learning model which can predict the numerical value for the consumer intention for a tweet. By doing this we can prove that we social media such as Twitter is also an important tool which marketers can use when deciding to target a customer. We believe that our work can be valuable to applications focusing on exploiting purchase intentions from social media.

- Challenges

The first challenge we faced was that we were not able to find any public dataset regarding purchase intention. We had to scrap the data from Twitter using a web scraper. Secondly, since we ourselves gathered the data we had to manually annotate the tweets. Again, this process was extremely time consuming as we had to go through each tweet and decide the purchase intention. Thirdly, we had limited annotated data because of the lengthy process of manual annotation and time constraint.

7.SYSTEM ARCHITECTURE

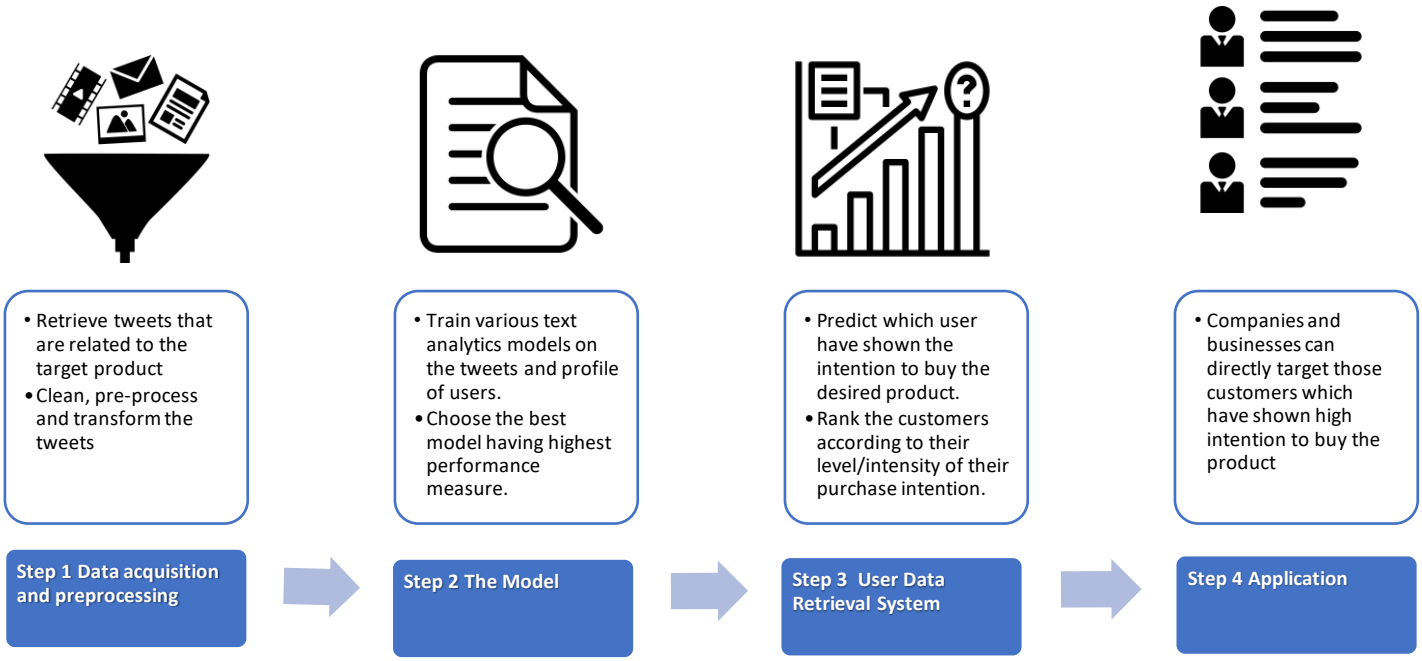

Fig 1: Steps involved in User Interest Based Social Media Data Retrieval System

PROJECT DESCRIPTION:

➢ Background and Motivation:

- Currently we have many recommendation systems available which recommend different products to the user, most of which are not efficient. No such effective model for businesses to identify potential customers.

- We want 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.

➢ Project Goal:

- We aim to analyze the tweets related to a product and identify the purchase intention in it. In this way we can rank the tweets which have high purchase intention and report the name of the person who tweeted as potential customer of product.

- We will make a model by gathering tweets from users who have already expressed intention to buy the product and see their tweet history and if possible, their web search history as well. Using this model, we will input potential customers who have tweeted about the product but have not bought it. And based on the training data the model will estimate a prediction/likelihood of whether the customer will buy it or not.

➢ Project Requirements:

- We require data from twitter to analyze purchase intention. For now, we are gathering the data by scraping the tweets of a product through a scraper. We will need a mechanism to save the scrapped tweets in storage for further processing and performing analysis and for that we have decided to use mongo db. We will need to annotate data retrieved from scrapper. Once model is trained, we will need to develop a website so that users can easily access our application through the graphical interface of the website. We will show on the website the purchase intention rank of tweets on level of 1 to 5 for the desired product of the user.

1.Functional Requirements

- Website : Website will be required so that users can easily access our application.

- Dashboard : 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.

- Get Relevant Users : The user will press the Get Relevant Users button through which he will be taken to the next screen which will show the list of users and tweets who have shown intention to buy the product.
- Find Purchase Intention of The Users : The user will then press The Find Purchase Intention Of The Users button through which he will be taken to the final screen which will show the results in a pie chart and list format for all the users who should be targeted and will also show the result in level form meaning that the users will be categorized according the level of intention they have expressed to buy the product.

2.Constraints

- We are scrapping the data not getting data through twitter Search API. This is sometimes not reliable and robust, so we will need to either get the data from some company which sells such data or will need access to the twitter search API.

- Another constraint is that we have a lot of tweets which we will use to train our model and we need annotated data so that we can evaluate our model but since annotating each and every tweet is time-consuming as well as there is no way to verify the annotated tweets, we will have to use methods of majority voting and averaging.

3.Objectives

- We aim to analyze the tweets related to a product and identify the purchase intention in it.

- We want to rank the tweets according to the level of purchase intention and show a list of potential customers to the user which he can use to directly target the customer.

➢ Validation and Acceptance Tests:

We will test the model accuracy by confusion matrix (Accuracy, Precision, Recall, F-Measure). This will give us a percentage of accuracy achieved by our model.
For our application development, we are going to run unit testing in which individual units of source code, would be tested to determine if they are fit to use. Then, we are

going to run integration testing where the individual source codes would be merged and tested as group.

Algorithms using :

Once the corpus was ready, we then used different text analytical models to test which one gave the best results. We used the following models:

1.Support Vector Machine (SVM):

Simply put, SVM is a supervised machine learning algorithm which does complex transformation on the data. And then it tries to separate data on classes we have defined on our data.

2.Naive Bayes:

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between every pair of features given the value of the class variable.

3.Logistic Regression:

Logistic regression is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables.

4.Decision Tree:

Decision Trees are a non-parametric supervised learning method used for classification and regression. The goal is to create a model that predicts the value of a target variable by learning simple decision rules inferred from the data features.

5.Neural Network:

It is deep learning machine algorithm, which is arranged in a layer of neuron. There is an input layer, output layer and hidden layers of neurons. Neuron network is adaptive as neurons in these layers learn from their initial input and subsequent runs.

Program Details:

➢ User interface

- 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.
- Secondly, the user will press the UPLOAD ANNOTATED DATASET button through which he will be asked to upload the dataset he wants to test for his product and find the relevant customers from that dataset. The output will be in a form of a pie chart and a table showing the different level of customers and the scores assigned to each customer.
- Thirdly, when the user presses the ANALYSIS button, he will be taken to the screen which will show the detailed analysis of our dataset that we have built containing word clouds, positive vs negative tweets and the most used words for that product.

➢ Errors

- Initially we encountered some errors in running the python environment and setting up a separate virtual environment for our project. However, we were soon able to overcome this by reading up on the documentation and following tutorials.
- We also encountered some errors running the Django framework. Setting up a sperate virtual environment and installing the missing dependencies later fixed this problem.

- We also faced some issues in integrating the python code with the website because it was our first time configuring a back-end server to run code scripts and sending the output to the html page and getting it to display correctly.

➢ Trails and tests

- We have tested the model accuracy by confusion matrix (Accuracy, Precision, Recall, F-Measure). Further we have also considered the True Negative Rate, The True Positive Rate and the shape of the ROC curve for more insights. This will give us a percentage of accuracy achieved by our model.

- For our application development, we opted to use unit testing in which individual units of source code, were tested to determine if they are fit to use. Then, we ran integration testing where the individual source was merged and tested as group.

- We also tested the usability of our website by carrying out the tasks and functions of the website in different scenarios and checking if they successfully completed.

Evaluation:

To evaluate our models, we using the following techniques:

1. Confusion Matrix
2. Accuracy
3. Precision
4. Recall
5. F-Measure
6. True Negative Rate

Further we have also considered The True Positive Rate and the shape of the ROC curve for more insights.

8.HARDWARE & SOFTWARE REQUIREMENT

➢ HARDWARE REQUIREMENTS

- Processor : > i3
- Ram : 4GB.
- Hard Disk : 500 GB.
- Input device : Standard Keyboard and Mouse.
- Compact Disk : 650 Mb.
- Output device : High Resolution Monitor.

➢ SOFTWARE REQUIREMENTS

- Operating system : macOS, Windows XP/7 or higher version.
- Coding language : Python 3.6.
- Framework : Python Django Framework.
- Storsge : Mongo DB. (if needed)
- Libraries : scikit-learn library for Python.
- Internet browser : Google Chrome.
- Code Editor :A Python code editor.

9.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:

i 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.

ii 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.

Looking at the other researches that are done in the similar field, our project also stands apart since we have implemented 5 different models and after evaluating them, we choose the best one customized to the product data.

We were not able to get more than 80% accuracy because of the two problems highlighted above. To achieve even 80% accuracy with an imbalance class data and such a small dataset is a victory.

After showing the website to a few potential clients we have received positive remarks about our product and people seem interest in this new approach to customer identification and targeted marketing.

REFERENCES

References:

1. Data Retrieval from Online Social Network Profiles for Social Engineering Applications, Sophia Alim, Ruquya Abdul-Rahman, Daniel Neagu and Mick Ridley *Department of Computing, University of Bradford, BD7 1DP {S.Alim, R.S.H Abdul-Rahman, D.Neagu, M.J.Ridley}@bradford.ac.uk.*
2. Social networks and information retrieval, how are they converging? A survey, a taxonomy and an analysis of social information retrieval approaches and platforms, Information Systems56(2016)1–18.