**Twitter User Interest Analysis**

**RAJENDRAKUMAR VM(2012103579)**

**PIRASAD T(2012103568)**

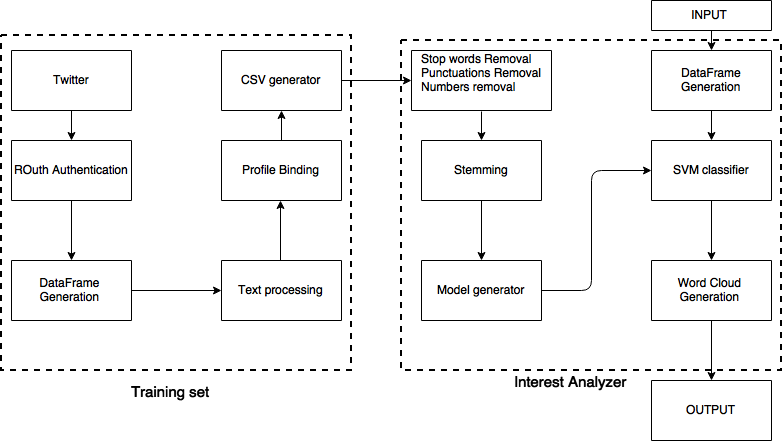
**Abstract :**

This project utilizes text mining tools to tokenize text and train classifier.By applying classification models on the testing data,we are able to evaluate the performance of the model generated before. After that, twitter users’ timeline can be retrieved to analyze their interests.. The goal of our work is to automatically classify incoming tweets into different categories so that users are not overwhelmed by the raw data. This is particularly useful because it is directly showing users the topics that the twitter user concerns and filter the tweets of wanted topics, especially when Twitter is accessed via handheld devices like smart phones.Thus it is important and necessary to classify these topics into general categories with high accuracy for better information retrieval.

**Problem Statement :**

Twitter, the popular online social networking service that enables users to send short 140-character messages and currently has more than 284 million monthly active users and 500 million tweets sent per day contains too much raw data which are not sorted or categorized. The social networking service provides a platform for people to interact with each other by following, commenting and re-tweeting. Hence, it is ideal for Twitter users to know the interest of others before following them.However, it is impractical to look through all his tweets if he is an active user.Thus we categorize the tweets of a particular user into various domains such as Food,Technology,Business,Sports Movies,News,Arts and Fashion.

**Block Diagram :**

****

**Methodology :**

We use Twitter API for extracting tweets of specific user using OAuth authorization. The API returns a JSON object as its response.We have to train the data set which generally involves the text pre processing and profile binding Interest analyzer process the tweets using SVM classifier and from which the word cloud can be generated We can get various user information such as tweets, friends list and followers list.A user’s Tweets can be retrieved using both the REST and the Streaming API.Any application which needs to access the tweets of a user must first create a sample application for which secret token is provided, which can be further used for accessing the tweets.To summarise the info about the list of tweets in data frame Data frame consists of favorite count,time created,retweet count From this we get the text field only along with the URL.

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis.It is used to train the data set and map them to a particular category in which the tweets fall upon.Stop word removal – in this process we remove the question words i.e which,what,when,where etc.. And also the,is,at,on Number removal is used to remove the numbers from tweet Stemming-it is the process done to get root words.The obtained model of svm is trained against user preprocessed timeline dataframe and based on the normality obtained the results are generated.

**Existing Work**:

The existing work makes use of brute force methodology by obtaining the user dataframes and the process of stemming is done.The resultant stemmed words are searched in a search engine to identify the profile the user is interested and based on further normalization the user interest is identified.

**Proposed Work** :

We have worked up on the assumption that we will be able to find out the user interest from their tweets based on the verified profiles of twitter under each category.The users in the twitter blogosphere tends to retweet the ideas and thus resonate in tandem with the verified profiles.We train our model with reliable sources of categories under consideration and then analyse with the user profile under examination.

**Evaluation Parameters :**

This Project identifies the users topic of interest based on tweets in real time.We propose a probability based approach by building a SVM model of verified twitter handles against that of the user.The following are identified as the evaluation for the project

* Accuracy
* Random sample of tweets
* Real time variation of tweets

**Accuracy:**

The user handles are identified and classified in major domains such as Food,art.business,technology,movies,sports.The accuracy depends on the frequency of words in the domain.Accuracy is hampered by inter dependability across major domains.

|  |  |
| --- | --- |
| Interest | Mean probabilistic Distribution |
| Technology | 0.8 |
| Movies | 0.46 |
| Fashion | 0.55 |
| Sport | 0.48 |
| Food | 0.52 |
| Music | 0.27 |
| Arts | 0.18 |
| Business | 0.22 |
| Politics | 0.34 |

**Random sample of tweets:**

The sample on which the SVM model is built is based on a [pseudorandom number generator](https://en.wikipedia.org/wiki/Pseudorandom_number_generator) **Mersenne Twister algorithm** that has been set during every run.Thus the SVM model generated is highly randomized and not localized.

for i = 0 to 226

temp = first bit of a(i) followed by last 31 bits of a(i+1) ;

a(i) = temp shifted right one bit xor

X'9908B0DF' if temp is odd xor

a(i+397) ;

next i

for i = 227 to 622

temp = first bit of a(i) followed by last 31 bits of a(i+1) ;

a(i) = temp shifted right one bit xor

X'9908B0DF' if temp is odd xor

a(i-227) ;

next i

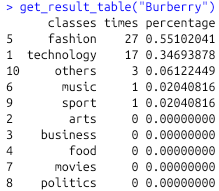
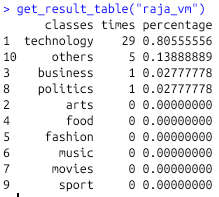
temp = first bit of a(623) followed by last 31 bits of a(0) ;

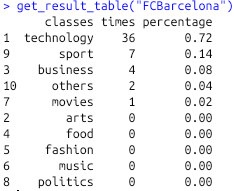
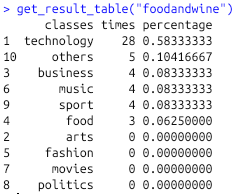
a(i) = temp shifted right one bit xor

X'9908B0DF' if temp is odd xor

a(396) ;

**Mean distribution for Handles:**





**Conclusion :**

For this project, we spot the problem of raw data and proposed a solution to it by building a model from our training data obtained through Twitter and practically using it to suggest the possible interests of the Twitter user. This model allows people to have a clearer understanding on other Twitter users and ease their information search journey in Twitter, which further enhance the Twitter user experience. Thoughthe model we built based on SVM have reasonable and acceptable accuracy percentage, it also has a few limitations that dent its efficiency. Hence, the improvements on our data collection process, model selection and model modification are favored to increase the efficiency of our solution. In conclusion, our model can practically solve our problem and be further used to build few applications to enhance Twitter user experience.