PROJECT REPORT ON

Efficient Celebrity Profiling in Twitter Social Network

“A dissertation submitted in partial fulfilment of the requirements of Bachelor of Technology Degree in Computer Science and Engineering of the Maulana Abul Kalam Azad University of Technology for the year 2016-2020”

Submitted by

Nafisa Anjum	14800116085
Anarpit Dey	14800116125
Anirban Patra	14800116120
Sayantan Das	14800116054

Under the guidance of

Kumar Gourav Das

Assistant Professor

Dept. of Computer Science & Engineering

Future Institute of Engineering and Management

Department of Computer Science and Engineering

Future Institute of Engineering & Management

(Affiliated to Maulana Abul Kalam Azad University of Technology, West Bengal)

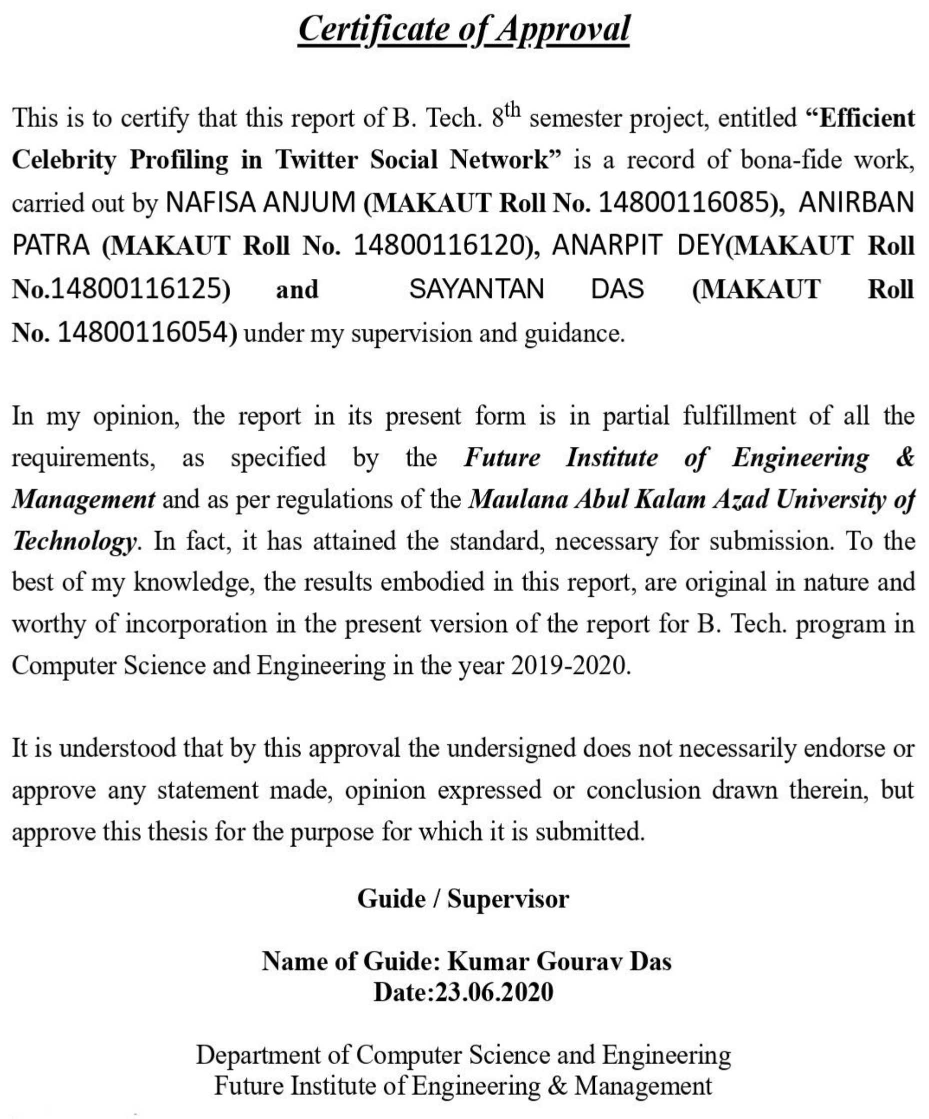
Kolkata-700150, WB

Gmail **Nafisa Anjum <nafisaanjum13@gmail.com>**

certificate of approval

1 message

Gourav Das <kumargouravdas18@gmail.com> Thu, Jun 25, 2020 at 10:39 AM
To: Nafisa Anjum <nafisaanjum13@gmail.com>



Certificate of Approval

This is to certify that this report of B. Tech. 8th semester project, entitled **"Efficient Celebrity Profiling in Twitter Social Network"** is a record of bona-fide work, carried out by NAFISA ANJUM **(MAKAUT Roll No. 14800116085),** ANIRBAN PATRA **(MAKAUT Roll No. 14800116120),** ANARPIT DEY**(MAKAUT Roll No.14800116125) and** SAYANTAN DAS **(MAKAUT Roll No. 14800116054)** under my supervision and guidance.

In my opinion, the report in its present form is in partial fulfillment of all the requirements, as specified by the ***Future Institute of Engineering & Management*** and as per regulations of the ***Maulana Abul Kalam Azad University of Technology***. In fact, it has attained the standard, necessary for submission. To the best of my knowledge, the results embodied in this report, are original in nature and worthy of incorporation in the present version of the report for B. Tech. program in Computer Science and Engineering in the year 2019-2020.

It is understood that by this approval the undersigned does not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve this thesis for the purpose for which it is submitted.

Guide / Supervisor

Name of Guide: Kumar Gourav Das
Date:23.06.2020

Department of Computer Science and Engineering
Future Institute of Engineering & Management

ACKNOWLEDGEMENT

We would like to express our gratitude to our Prof. Kumar Gourav Das for taking us under his wings. It has been an honour to work with him in this seminar project he has taught us both consciously and unconsciously how good pattern recognition is done. The joy and enthusiasm he has for his research was contagious and motivational for us, even during tough times in the pursuit of completing this project.

We are immensely grateful to all the faculty members technical/laboratory staff and attendance of FIEM CSE department who has contributed exceptionally to our personal and professional time at FIEM.

We would like to thank Prof. Tapas Roy head of department computer science and engineering, FIEM, for allowing us to participate in this project we are especially grateful to him for his time and patience despite the innumerable queries we had for him during the course of the project

Our heartfelt gratitude goes to *Dr Aloke Kumar Ghosh, principal, FIEM* for all his extensive support to us during our time in FIEM.

Lastly, we are grateful to our classmates for all their love encouragement and camaraderie and for inspiring us with their competitiveness thank you.

Nafisa Anjum (sign)_______________

University Roll no 14800116085 Registration no- 161480110052

Anarpit Dey (sign)_______________

University Roll no 14800116125 Registration no- 161480110017

Anirban Patra (sign)_______________

University Roll no 14800116120 Registration no- 161480110012

Sayantan Das (sign)_______________

University Roll no 14800116054 Registration no- 161480110083

PROJECT ABSTRACT

The topic of our final year project is “Efficient Celebrity Profiling in Twitter Social Network”. The project is based on Natural Language Processing (NLP). Our objective is to user classification and predict unknown celebrity’s most tweets belong to which class. We basically do this project upon python. We’ll use topic modelling to classify (multi class classification) the category and for this we’ll use **Data Dictionary.**

- Firstly, The labelled data set of Twitter users along with six bipolar domains (Business, Sports, Entertainment, Education, Technology &Politics)was obtained in a semi-automatic manner wherein we first manually gather Twitter account handles for a specific class over the Internet and then use the Tweepy python library for extracting the Twitter feed of the user. We first focused on Twitter handles that were available in the Internet and then filled the remaining instance slots with users across the globe so as to balance the data-set. For every Twitter User we collect up to last 600 tweets made by the user.
- Secondly, we have to do the pre-processing where hashtags, user mentions and emojis are removed using python Regular expression package.
- Then stop words from the NLTK toolkit will be removed from the set of words.
- After this step we’ll find out the most frequent word of each domain to create a data dictionary. By which we predict how we can find our objective.
- For each document we create a dictionary reporting how many words and how many times those words appear.
- Then We will collect an unknown celebrity’s tweets and pre-process it and by comparing with the data dictionary we will conclude the category of that person.

.

CONTENTS

CHAPTER 1

INTODUCTION

With the success of social networks, the alliance of social information has become imperative. We study strategies to build profiles of the Twitter user. Twitter is today the most leading micro-blogging service accessible on the Web. People publish short messages about their everyday actions on Twitter. Twitter is an American online news and social networking assistance on user post and communicates with messages known as “tweets”, Tweets were formerly restricted to 140 characters containing only text or hyperlinks, primarily, it’s a micro-blogging site that provides millions of users to communicate, fasten connection and more.

The skyrocketing demand for social networking sites has created vast resources of user-generated content. As of January 2019, witter users in the leading market. As of October 2018, Twitter reports a monthly usage of 500 million active users with more than 800 million tweets swapped per day. Twitter’s Streaming API systems provide a secure and programmatic path to the vast amount of data generated in the social network. This has made Twitter an active hub for user personality and profiles related research. Some of the studies that have been carried out include classifying user’s demographic information, foretelling brand-related events from user’s tweets and tweet topic identification.

Studies have also been carried in finding out finely-tuned features like divining the type of Twitter account publishing an event (individual, news organization or other).

In this context, the problem of automatically identifying user interests and user profiling has gained significant attention.

User profiling is essential in several areas including marketing, forensic science, and security. For example, from a marketing perspective, it is always useful to know details about user of text in blogs and reviews, so that relevant recommendations can be provided to users. The linguistic profile of an author of abusive message would be helpful from a forensic linguistics view point. User profiling has gained significant attention. Twitter's Streaming API methods provide easy and programmatic access to the vast amount of data generated in the social network. This has made Twitter an active hub for user personality and profile related research. Some of the studies that have been carried out include identifying user's demographic information, predicting brand-related events from user's tweets and tweet topic identification. Researches have also been carried in finding out finely-tuned features like predicting the type of Twitter account reporting an event (individual, news organization or other).

1.1 OBJECTIVE

Knowing the profile of an author could be of key importance. For instance, from a forensic linguistics perspective being able to know what is the linguistic profile of a suspected text message (language used by a certain type of people) and identify characteristics (language as evidence) just by analysing the text would certainly help considering suspects. Similarly, from a marketing viewpoint, companies may be interested in knowing, on the basis of the analysis of blogs and online product reviews, what types of people like or dislike their products. The training corpus provided for this task contains English text of Indian to avoid code mixing.

A person's syntactic construct or lexical uses can give cues to his authorship, but to describe and quantify such characteristics despite several prior research on authorship attribution and some author profiling.

1.2 MOTIVATION

Being able to infer an author’s gender, age, native language, dialects, or personality opens a world of possibilities—among others in marketing, where companies may analyse online reviews to improve targeted advertising, or in forensics, where the profile of authors could be used as valuable additional evidence in criminal investigations, and in security, where knowing the demographics of social media users (age and gender), as well as cultural and social context such as native language and dialects, may help to identify potential terrorists.

CHAPTER 2

Why we choose this topic?

- Celebrity Author profiling from text has been an interesting topic recently because of the increase in the availability of texts. This is mostly because of the internet where text is one of the forms of communication. This could be present in blogs, website, customer review, and even twitter posts.

- Author anonymity has been present mostly in the web, using profiling can be useful, especially in aspects such as marketing, advertising as well as security, forensics, plagiarism detection and terrorism prevention. Profiling mainly uses such text to determine certain aspects of the author such as age, gender and certain personality traits.

- The author profiling task is a yet unsolved problem due to its difficulty. For instance, exploration of more features such as stylistic feature. It has been studied by many researches and while some great progress and good results, it still has many unexplored areas and room for improvement**.**

CHAPTER 3

SOFTWARE REQUIREMENT SPECIFICATION (SRS)

1. Anaconda

    1.1 Python 3.6

2.Microsoft Office suite

    1.1 Microsoft Excel

    1.2 Microsoft Power Point

3.Twitter data download API (Tweepy package)

4. Libraries

1.1 Pandas: Library for data manipulation and analysis. In particular, it offers data structures and operations for manipulating numerical tables and time series.

1.2 NLTK: The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English

1.3 Regular Expression: The Python module re provides full support for Perl-like regular expressions in Python.

1.4 Matplotlib: Matplotlib is a comprehensive library for creating static, animated, and interactive visualizations in Python.

1.5 Genism: It is an open-source library for unsupervised topic modeling and natural language processing, using modern statistical machine learning

CHAPTER 4

THE DOMAIN ON WHICH OUR PROJECT IS BASED ON

NATURAL LANGUAGE PROCESSING (NLP):

Natural language processing (NLP) is the ability of a computer program to understand human language as it is spoken. NLP is a component of artificial intelligence (AI).The development of NLP applications is challenging because computers traditionally require humans to "speak" to them in a programming language that is precise, unambiguous and highly structured, or through a limited number of clearly enunciated voice commands. Human speech, however, is not always precise -- it is often ambiguous and the linguistic structure can depend on many complex variables, including slang, regional dialects and social context.

NLP can be used to interpret free text and make it analysable. There is a tremendous amount of information stored in free text files, like patients' medical records, for example. Prior to deep learning -based NLP models, this information was inaccessible to computer-assisted analysis and could not be analysed in any kind of systematic way. But NLP allows analysts to sift through massive troves of free text to find relevant information in the files.

WORKS DONE ON THIS DOMAIN

1. Sentiment Analysis
2. Twitter trending
3. Author Profiling
4. User classification

CHAPTER 5

5.1 SYSTEM OVERVIEW

Our method is composed of six steps:

Pre-processing:

- **Data Collection**

  We've collected the tweets of six individual domains of Indian user mostly in English. We've collected the data using twitter API.

- **Pre-Processing**

  We've pre-processed the tweets by removing hashtags, user mentions, Emojis, extra space between words and punctuations.

- **Tokenization**

  It is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

- **Stop Word Removal**:

  The Stop words are removed using NLTK tool kit.

- **Text Analysis**:

  Patterns within written text are not the same across all authors or languages. This allows linguists to study the language of origin or potential authorship of texts where these characteristics are not directly known such as the Federalist Papers of the American Revolution. We started by calculating the number of occurrences of all words found in the corpus ranking them in order of their appearances. We calculate the CF (the class frequency) for each class of attributes in order to measure the frequency of occurrence of each class of attributes in each document of the corpus.

- **Data Dictionary:**

We have made a data dictionary of each class depending upon the CF, where the most frequent word of each class is there.

5.1.1: LIMITATIONS OF EXISTING SYSTEM

1. **Redundant data**: In this section of work we may face that the data we collect for usage maybe redundant in nature. For this reason, we may not predict our objective because some of the texts contain only emoji or some of the text contain regional language that may not have any relevant meaning.

2. **Code mixing**: In this field of work we face code mixing, that means for any particular region any particular language may be used in various type. So, for avoiding this thing we mostly download Indian personalities in English document.

3. **Classifier**: If we have some problem of collecting redundant data or if the text contains mixed language then the classifier may not be able to classify the text. For that reason, we may not predict the unknown person, or in the other language may be our accuracy not good as well.

5.2.1 Data Collection:

We've collected the tweets of six individual domains of Indian Celebrity mostly in English. We've collected the data using twitter API.

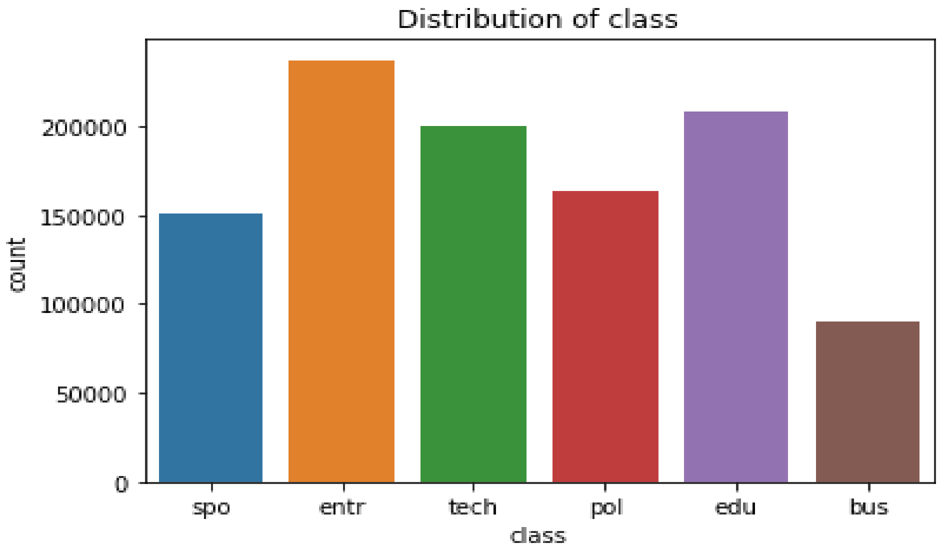
The Twitter API allows you to access the features of Twitter without having to go through the website interface. This can be useful for doing things like posting tweets or sending directed messages in an automated way with scripts.

Table 1: Distribution of Twitter User Profiles collected

Interest Class	User Instances
politics	100
Entertainment	102
Education	100
Business	95
Technology	93
Sports	100
Total User	590

Graphical Distribution of tweets

Image I: Distribution of Tweet of User Profiles collected



Tokenization is the process of tokenizing or splitting a string, text into a list of tokens. One can think of token as parts like a word is a token in a sentence, and a sentence is a token in a paragraph.

Stop Word Removal:

Stop Words: A stop word is a commonly used word (such as "the", "a" , "an", "in") that a search engine has been programmed to ignore both when indexing entries for searching and when retrieving them as the result of a search query. The Stop words are removed using NLTK tool kit.

Text Analysis:

Patterns within written text are not the same across all authors or languages. This allows linguists to study the language of origin or potential authorship of texts where these characteristics are not directly known such as the Federalist Papers of the American Revolution. We started by calculating the number of occurrences of all words found in the corpus ranking them in order of their appearances. We calculate the CF (the class frequency) for each class of attributes in order to measure the frequency of occurrence of each class of attributes in each document of the corpus

. **Feature Set Generation:**

The most common approach in the literature distinguish two main types of attributes that can be used to detect the celebrity's profile is the stylistic and content-based ones. We manually grouped the term belonging to the same class attribute. We identified six classes, that is:

1: Sports 2: Entertainment 3: Technology 4: Business

5: Education 6: Politics

5.2.1: OBJECTIVE OF THE PROPOSED SYSTEM

The objective was to predict a celebrity user’s tweet classification, on different given classes. We basically focused on tweets i.e. a social media application. In this section we describe the construction of the corpus and discuss particular properties, limitations and the job we will have to do moreover, the evolution measures are described.

We have merged the training and test sets from the twitter API. Firstly, we look for public link in profiles that share a twitter account. We verified whether the twitter account exists whether it is written in one of the languages we are interested in whether it is updated by only one person and whether this person is easily identifiable. We describe the organizational twitter account if we were not sure that the account was updated by the person identified in the linked in id.

5.2.2: USAGE OF THE PROPOSED SYSTEM

User profiling is essential in several areas including marketing, forensic science, and security. For example, from a marketing perspective, it is always useful to know details about user of text in blogs and reviews, so that relevant recommendations can be provided to users. The linguistic profile of an author of abusive message would be helpful from a forensic linguistics view point. User profiling has gained significant attention.

CHAPTER 6

6.1 Tools used in development

Anaconda: anaconda is a free and open source distribution of the Python and R programming languages for large scale data processing predictive analytics and scientific computing that aims to simplify package management and deployment.

Spyder IDE: Spyder is an open source cross-platform integrated development environment (IDE) for scientific programming in the Python language. Spyder integrates with a number of prominent packages in the scientific Python stack, including NumPy, SciPy, Matplotlib, pandas, I Python, SymPy and Cython, as well as other open source software.

Jupyter Notebook: Jupyter Notebook App is a server-client application that allows editing and running notebook documents via a web browser. The Jupyter Notebook App can be executed on a local desktop requiring no internet access (as described in this document) or can be installed on a remote server and accessed through the internet.

In addition to displaying/editing/running notebook documents, the Jupyter Notebook App has a “Dashboard” (Notebook Dashboard), a “control panel” showing local files and allowing to open notebook documents or shutting down their kernels.

Data Collection: Twitter API a python rapid for performing API requests such as searching for users and downloading to its tweets this library handles all of the queries for you and provides it to you in a simple Python interface.

DEVELOPMENT ENVIROMENT

Windows 10: It is a personal computer operating system developed and released by Microsoft as apart of Windows NT family of operating system,

HARDWARE USED:

1. CPU 1.9GHz Intel core i3-4030U (3MB cache)

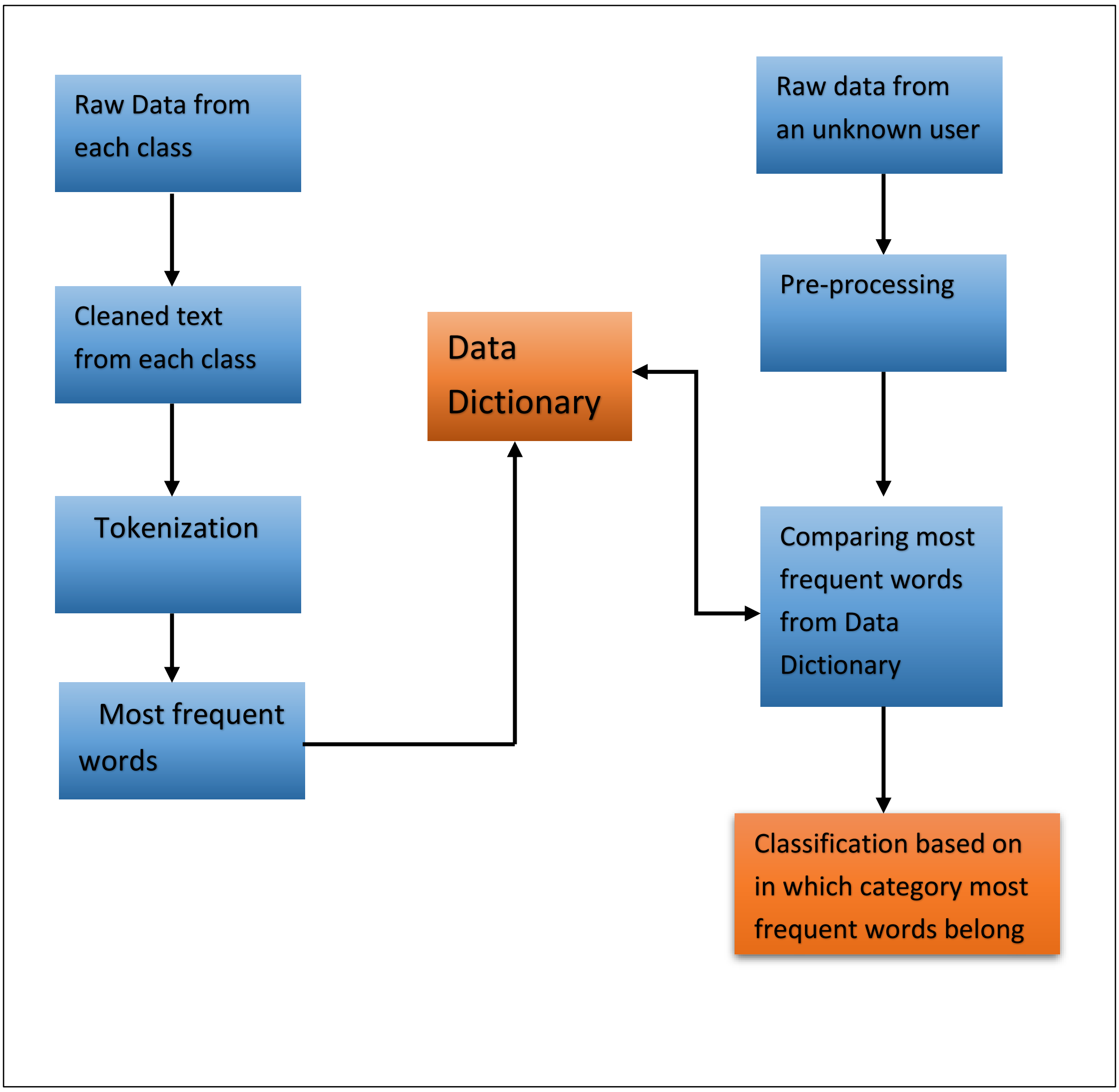
2. Graphics Intel HD Graphics 4400

3. RAM 8GB DDRL RAM

4. Storage 500GB (5400rpm with 16 GB SSD cache)

CHAPTER 7

DATAFLOW DIAGRAM


Raw Data from each class
Cleaned text from each class
Tokenization
Most frequent words
Data Dictionary
Raw data from an unknown user
Pre-processing
Comparing most frequent words from Data Dictionary
Classification based on in which category most frequent words belong

CHAPTER 8

DATA DICTIONARY:

Here is some example of most frequent words of each class:

Business: “Rise”, “market”, “Marketing” etc.

Education: “Books”, “Learning”, “Grade” etc.

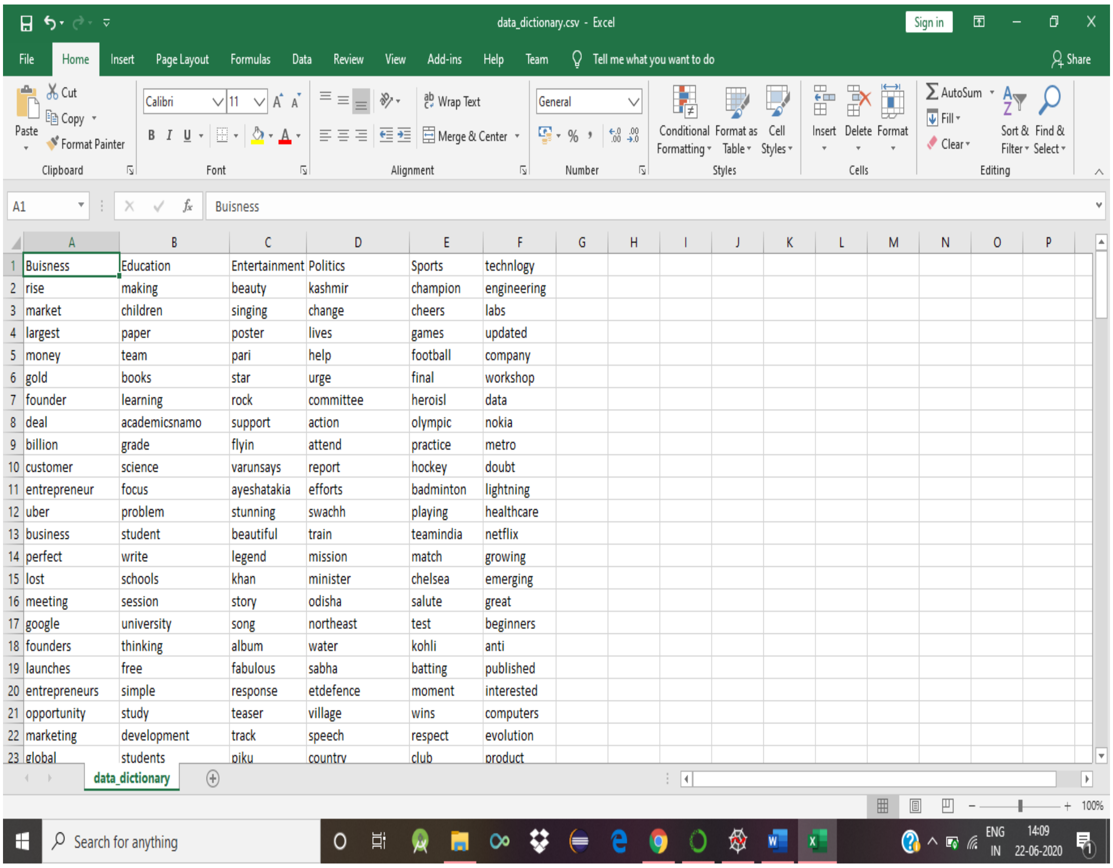
Entertainment: “Album”, “Teaser”, “Movie” etc.

Politics: “Minister”, “leadership”, “election” etc.

Sports: “Olympic”, “Practice”, “Hockey” etc.

Technology: “Engineering”, “Data”, “Apps” etc.

Snapshot of the CSV file:

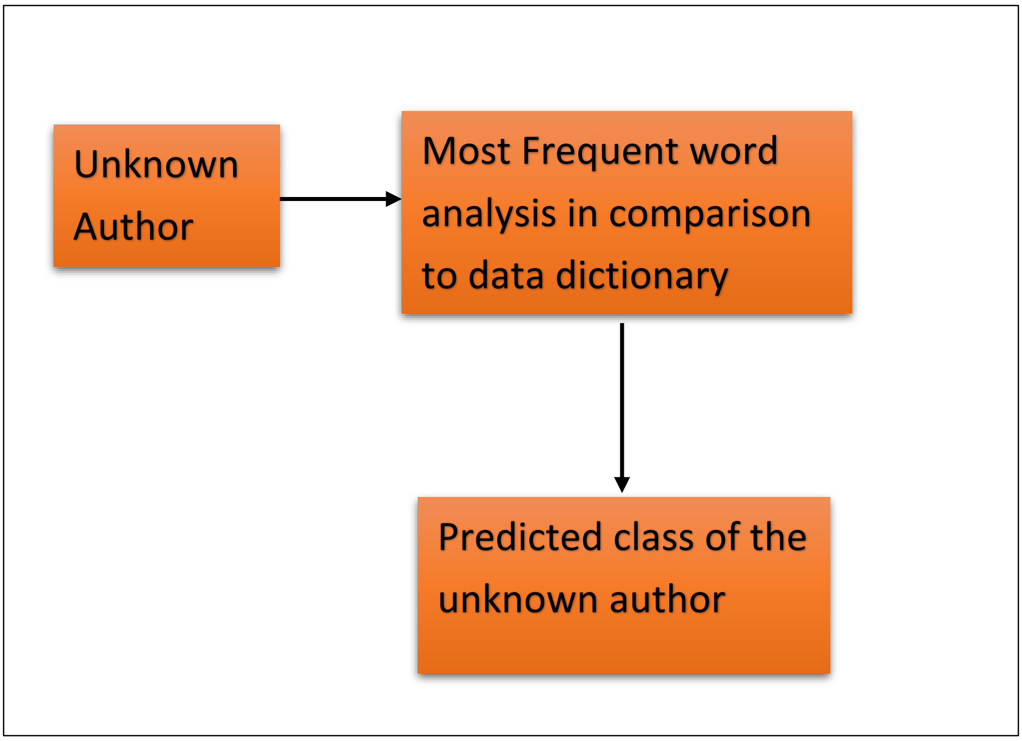


A1 — Buisness

	A	B	C	D	E	F
1	Buisness	Education	Entertainment	Politics	Sports	technlogy
2	rise	making	beauty	kashmir	champion	engineering
3	market	children	singing	change	cheers	labs
4	largest	paper	poster	lives	games	updated
5	money	team	pari	help	football	company
6	gold	books	star	urge	final	workshop
7	founder	learning	rock	committee	heroisl	data
8	deal	academicsnamo	support	action	olympic	nokia
9	billion	grade	flyin	attend	practice	metro
10	customer	science	varunsays	report	hockey	doubt
11	entrepreneur	focus	ayeshatakia	efforts	badminton	lightning
12	uber	problem	stunning	swachh	playing	healthcare
13	business	student	beautiful	train	teamindia	netflix
14	perfect	write	legend	mission	match	growing
15	lost	schools	khan	minister	chelsea	emerging
16	meeting	session	story	odisha	salute	great
17	google	university	song	northeast	test	beginners
18	founders	thinking	album	water	kohli	anti
19	launches	free	fabulous	sabha	batting	published
20	entrepreneurs	simple	response	etdefence	moment	interested
21	opportunity	study	teaser	village	wins	computers
22	marketing	development	track	speech	respect	evolution
23	global	students	piku	country	club	product

CHAPTER 9

User classification Diagram: -


Unknown Author
Most Frequent word analysis in comparison to data dictionary
Predicted class of the unknown author

We will do this project upon multi-level user classification of celebrities. To classify tweets using various machine learning techniques, a proper set of features known most frequent word as is to be required to extract from the tweets. For extracting tweets tokenization approach will be used. The frequency of each word is to be used as data. As there may be a large number of words extracted from in different tweets, using all data will cause to increase the overload and dimension. Thus, we will first identify the common words and remove them from the dataset and create a data dictionary from it. Next, we'll take an unknown user's tweet and compare from the data dictionary and the most frequent word from each class is obtained. The highest number of words from the class will conclude which category the author/celebrity belongs.

For instance, we are taking Sachin Tendulkar's tweet and by analysing the tweets the following result is obtained:

BUSINESS: meeting

No of words Found: 1

EDUCATION: making, children, team, kids

No of words Found: 4

ENTERTAINMENT: support, beautiful, kind, fans, movie, watching, film, enjoyed

No of words Found: 8

POLITICS: country, health, life, soon, nation

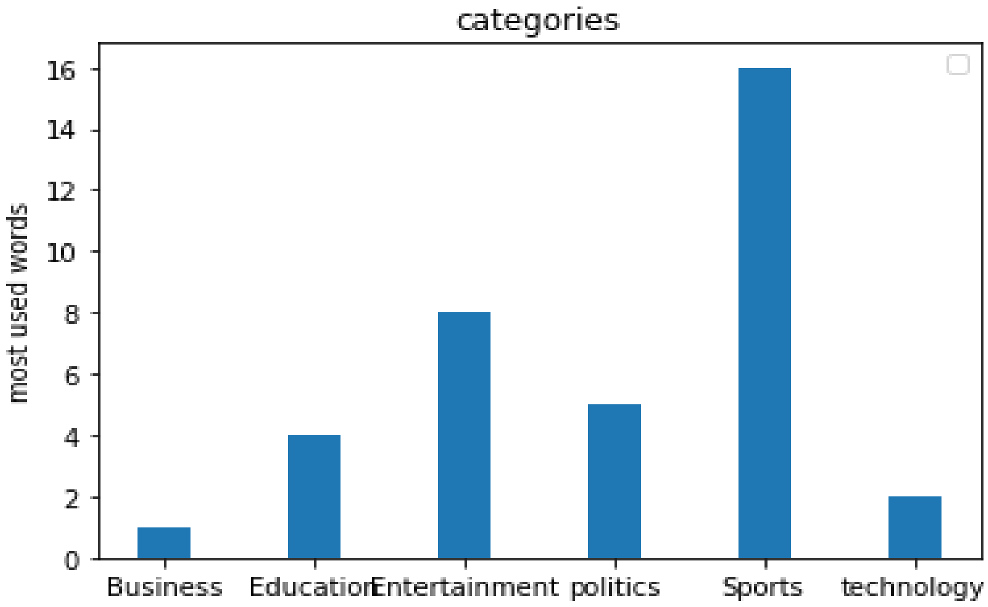
No of words Found: 5

SPORTS: playing, match, test, batting, moment, cricket, season, game, luck, performance, fantastic, played, play, series, sports, proud

No of words Found: 16

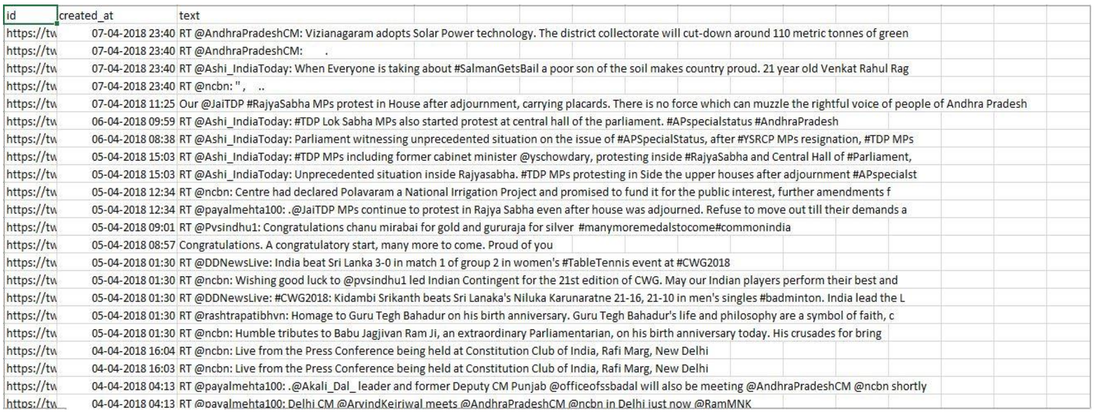
TECHNOLOGY: great, watch

No of words Found: 2



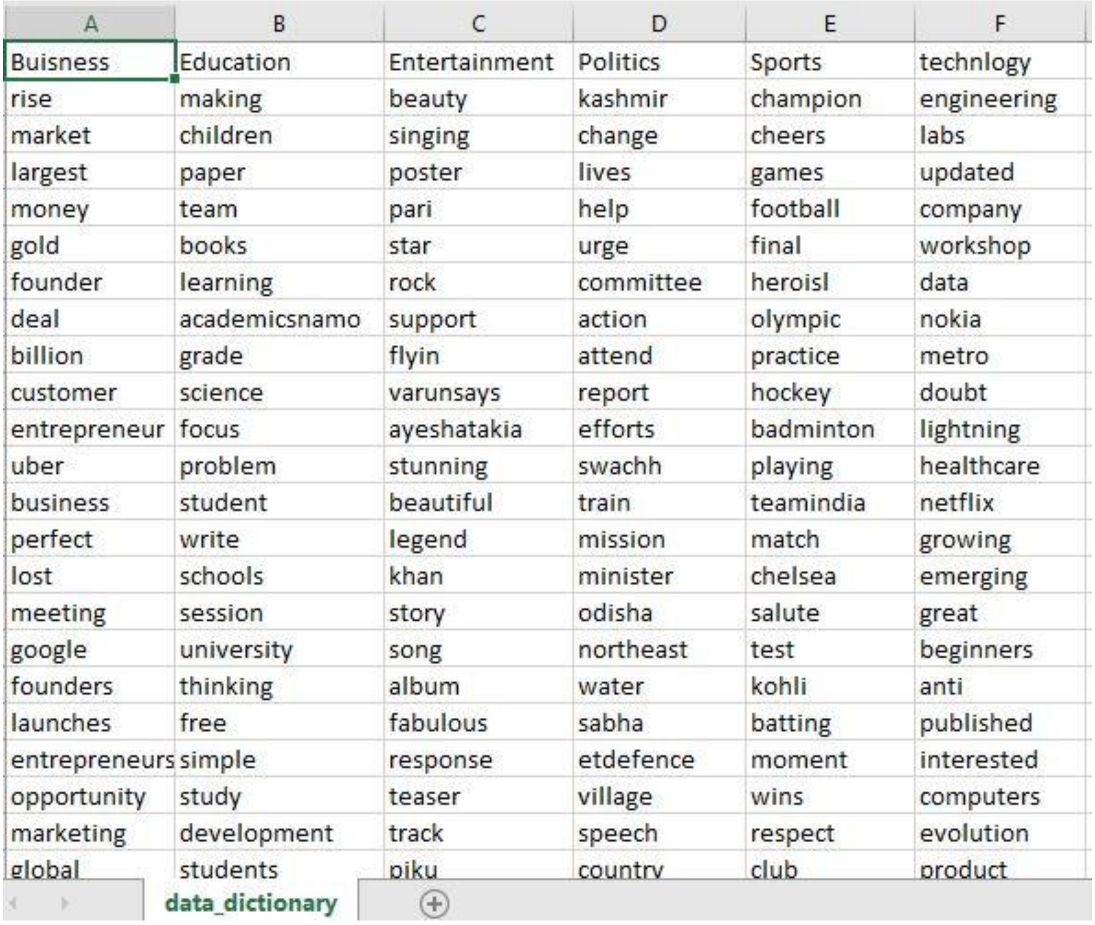
CHAPTER 10

Csv file of an unknown user:



id	created_at	text
https://tw	07-04-2018 23:40	RT @AndhraPradeshCM: Vizianagaram adopts Solar Power technology. The district collectorate will cut-down around 110 metric tonnes of green
https://tw	07-04-2018 23:40	RT @AndhraPradeshCM: .
https://tw	07-04-2018 23:40	RT @Ashi_IndiaToday: When Everyone is taking about #SalmanGetsBail a poor son of the soil makes country proud. 21 year old Venkat Rahul Rag
https://tw	07-04-2018 23:40	RT @ncbn: " , ..
https://tw	07-04-2018 11:25	Our @JaiTDP #RajyaSabha MPs protest in House after adjournment, carrying placards. There is no force which can muzzle the rightful voice of people of Andhra Pradesh
https://tw	06-04-2018 09:59	RT @Ashi_IndiaToday: #TDP Lok Sabha MPs also started protest at central hall of the parliament. #APspecialstatus #AndhraPradesh
https://tw	06-04-2018 08:38	RT @Ashi_IndiaToday: Parliament witnessing unprecedented situation on the issue of #APSpecialStatus, after #YSRCP MPs resignation, #TDP MPs
https://tw	05-04-2018 15:03	RT @Ashi_IndiaToday: #TDP MPs including former cabinet minister @yschowdary, protesting inside #RajyaSabha and Central Hall of #Parliament,
https://tw	05-04-2018 15:03	RT @Ashi_IndiaToday: Unprecedented situation inside Rajyasabha. #TDP MPs protesting in Side the upper houses after adjournment #APspecialst
https://tw	05-04-2018 12:34	RT @ncbn: Centre had declared Polavaram a National Irrigation Project and promised to fund it for the public interest, further amendments f
https://tw	05-04-2018 12:34	RT @payalmehta100: .@JaiTDP MPs continue to protest in Rajya Sabha even after house was adjourned. Refuse to move out till their demands a
https://tw	05-04-2018 09:01	RT @Pvsindhu1: Congratulations chanu mirabai for gold and gururaja for silver #manymoremedalstocome#commonindia
https://tw	05-04-2018 08:57	Congratulations. A congratulatory start, many more to come. Proud of you
https://tw	05-04-2018 01:30	RT @DDNewsLive: India beat Sri Lanka 3-0 in match 1 of group 2 in women's #TableTennis event at #CWG2018
https://tw	05-04-2018 01:30	RT @ncbn: Wishing good luck to @pvsindhu1 led Indian Contingent for the 21st edition of CWG. May our Indian players perform their best and
https://tw	05-04-2018 01:30	RT @DDNewsLive: #CWG2018: Kidambi Srikanth beats Sri Lanaka's Niluka Karunaratne 21-16, 21-10 in men's singles #badminton. India lead the L
https://tw	05-04-2018 01:30	RT @rashtrapatibhvn: Homage to Guru Tegh Bahadur on his birth anniversary. Guru Tegh Bahadur's life and philosophy are a symbol of faith, c
https://tw	05-04-2018 01:30	RT @ncbn: Humble tributes to Babu Jagjivan Ram Ji, an extraordinary Parliamentarian, on his birth anniversary today. His crusades for bring
https://tw	04-04-2018 16:04	RT @ncbn: Live from the Press Conference being held at Constitution Club of India, Rafi Marg, New Delhi
https://tw	04-04-2018 16:03	RT @ncbn: Live from the Press Conference being held at Constitution Club of India, Rafi Marg, New Delhi
https://tw	04-04-2018 04:13	RT @payalmehta100: .@Akali_Dal_ leader and former Deputy CM Punjab @officeofssbadal will also be meeting @AndhraPradeshCM @ncbn shortly
https://tw	04-04-2018 04:13	RT @payalmehta100: Delhi CM @ArvindKejriwal meets @AndhraPradeshCM @ncbn in Delhi just now @RamMNK

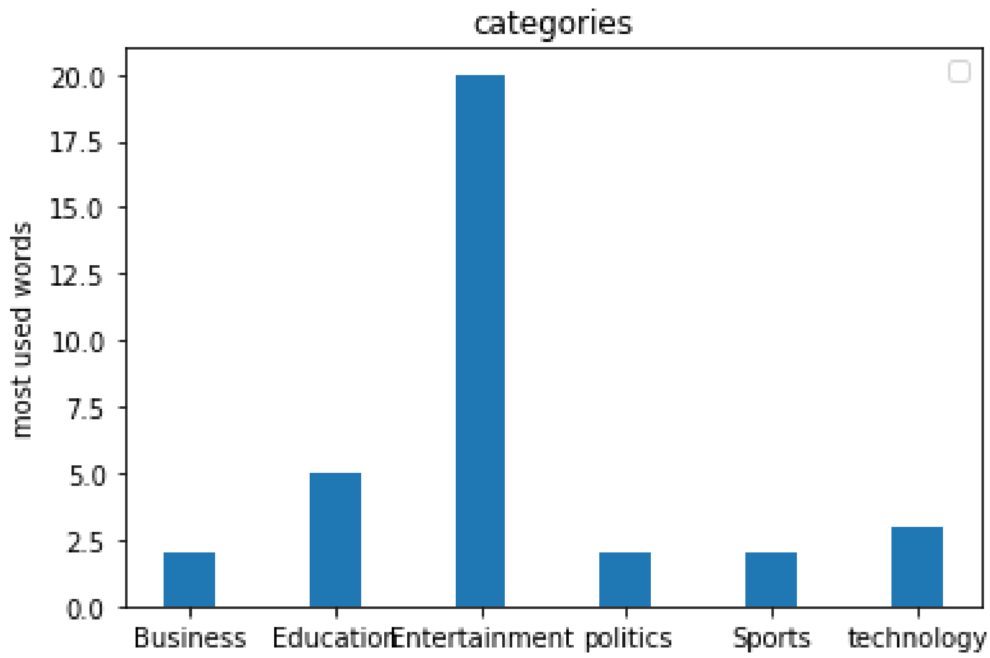
Data Dictionary:



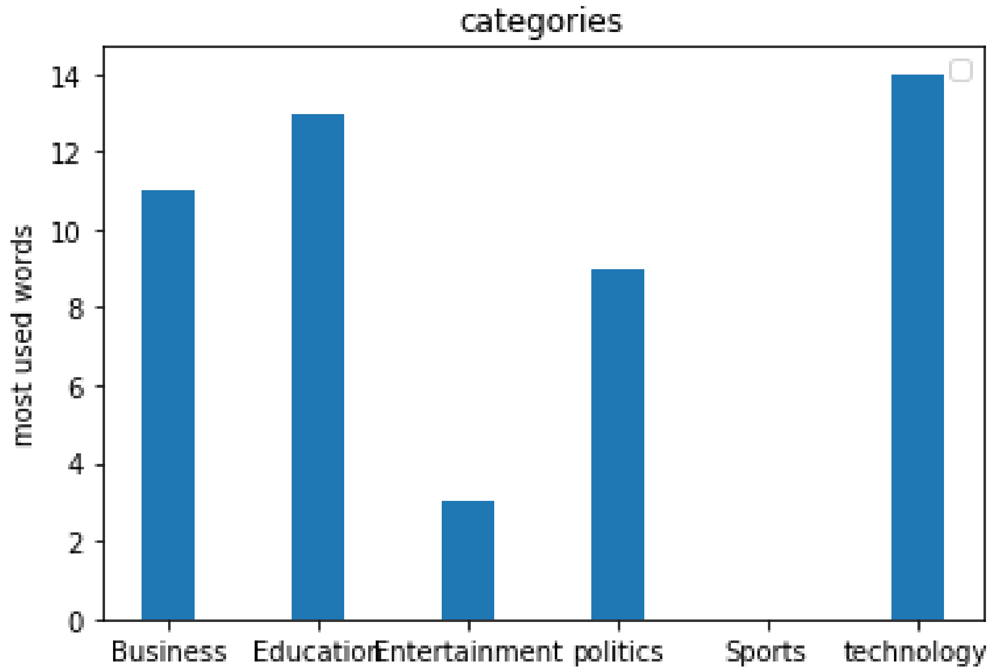
A	B	C	D	E	F
Buisness	Education	Entertainment	Politics	Sports	technlogy
rise	making	beauty	kashmir	champion	engineering
market	children	singing	change	cheers	labs
largest	paper	poster	lives	games	updated
money	team	pari	help	football	company
gold	books	star	urge	final	workshop
founder	learning	rock	committee	heroisl	data
deal	academicsnamo	support	action	olympic	nokia
billion	grade	flyin	attend	practice	metro
customer	science	varunsays	report	hockey	doubt
entrepreneur	focus	ayeshatakia	efforts	badminton	lightning
uber	problem	stunning	swachh	playing	healthcare
business	student	beautiful	train	teamindia	netflix
perfect	write	legend	mission	match	growing
lost	schools	khan	minister	chelsea	emerging
meeting	session	story	odisha	salute	great
google	university	song	northeast	test	beginners
founders	thinking	album	water	kohli	anti
launches	free	fabulous	sabha	batting	published
entrepreneurs	simple	response	etdefence	moment	interested
opportunity	study	teaser	village	wins	computers
marketing	development	track	speech	respect	evolution
global	students	piku	country	club	product

Output of various users:

Deepika Padukone:



Abhijit Bhaduri:

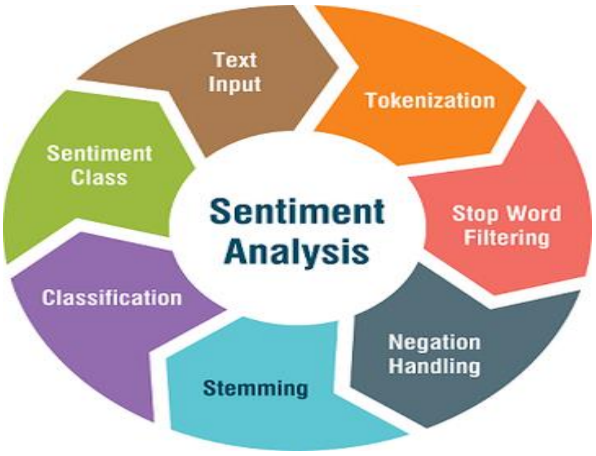


CHAPTER 11

The Works Done on This Domain

- Sentiment analysis can be defined as a process that automates mining attitudes, opinions, views and emotions from text, speech, tweets and database source through Natural Language Processing (NLP). Sentiment analysis involves classifying opinions in text into categories like “positive” or “negative”. It’s also referred as subjectivity analysis, opinion mining, and appraisal extraction.

- Sentiment analysis aims to determine the attitude of a speaker, writer, or other subject with respect to some topic or the overall contextual polarity or emotional reaction to a document, interaction, or event. The attitude may be a judgment or evaluation affective state (that is to say, the emotional state of the author or speaker), or the intended emotional communication (that is to say, the emotional effect intended by the author or interlocutor).



CHAPTER 12

Conclusion and Future Work: -

Twitter’s Streaming API methods provide easy and programmatic access to the vast amount of data generated in the social network. This has made Twitter an active hub for user personality and profile related research. Some of the studies that have been carried out include identifying user’s demographic information, predicting brand-related events from user’s tweets and tweet topic identification.

User profiling is essential in several areas including marketing, forensic science, and security. For example, from a marketing perspective, it is always useful to know details about user of text in blogs and reviews, so that relevant recommendations can be provided to users. The linguistic profile of an author of abusive message would be helpful from a forensic linguistics view point. User profiling has gained significant attention.

CHAPTER 13

REFERENCES: -

1. Twitter. https://about.twitter.com/company
2. Twitter Streaming APIs. https://dev.twitter.com/streaming/overview
3. Weka 3: Data Mining Software in Java. http://www.cs.waikato.ac.nz/ml/weka/
4. India to have third-largest Twitter population by 2014: eMarketer (2014). http:// indianexpress.com/article/india/politics/india-to-have-third-largest-twitter-population-by-2014-emarketer
5. GitHub - Twitter User Categorization (2015). https://github.com/AKSHAYH/ twitterusercategorization
6. An, J., Cha, M., Gummadi, P.K., Crowcroft, J.: Media landscape in twitter: a world of new conventions and political diversity. In: ICWSM (2011)
7. Bifet, A., Frank, E.: Sentiment knowledge discovery in twitter streaming data. In: Pfahringer, B., Holmes, G., Hoffmann, A. (eds.) DS 2010. LNCS, vol. 6332, pp. 1–15. Springer, Heidelberg (2010)
8. Bollen, J., Mao, H., Zeng, X.: Twitter mood predicts the stock market. Journal of Com-putational Science **2**(1), 1–8 (2011)
9. Chakrabarti, D., Punera, K.: Event summarization using tweets. In: ICWSM 2011, pp. 66–73 (2011)
10. De Choudhury, M., Diakopoulos, N., Naaman, M.: Unfolding the event landscape on twitter: classification and exploration of user categories. In: Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 241–244. ACM (2012)
11. Kouloumpis, E., Wilson, T., Moore, J.: Twitter sentiment analysis: the good the bad and the omg! In: ICWSM 2011, pp. 538–541 (2011)
12. McCord, M., Chuah, M.: Spam detection on twitter using traditional classifiers. In: Calero, J.M.A., Yang, L.T., Mármol, F.G., García Villalba, L.J., Li, A.X., Wang, Y. (eds.) ATC 2011. LNCS, vol. 6906, pp. 175–186. Springer, Heidelberg (2011)