



PREDICTING ELECTION RESULTS WITH SOCIAL MEDIA AND NEWS DATA

Under Guidance of-

**Ms. Archana Purwar
Dr. Shikha Mehta**

Undertaken By-

PRATIGYA AGARWAL	14103180
GAURAV SINGH	14103159
AMAN	14103168

CERTIFICATE

This is to certify that the work titled “ **ELECTION PREDICTION**” submitted by “**PRATIGYA AGARWAL,GAURAV SINGH, AMAN** ” in partial fulfilment for the award of degree of B.TECH IN COMPUTER SCIENCE of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor:

Name of Supervisor: Ms. Archana Purwar

Dr. Shikha Mehta

Date: 4-May-2017

ACKNOWLEDGEMENT

A lot of efforts have been taken in this project. This project consumed huge amount of work, research and dedication. However, it would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

We are highly indebted to Ms. Archana Purwar , Dr. Shikha Mehta for his guidance and constant supervision as well as for providing necessary information regarding the project & also for his support in completing the project.

My thanks also go to my colleagues in developing the project and people who have willingly helped me out to the best their abilities.

Objective

This project aims to predict the result of elections with the use of data collected from twitter and news website.

Abstract

Micro-blogging provider Twitter has become very popular communication tools for Internet and Mobile users. People write about their life, share opinions on a variety of topics and discuss current political issues.

This huge amount of raw data can be used for industrial or studies purpose by organizing according to our requirement and processing.

Data is in the form of tweets which are opinions of people on different topics which lie in political category .

We present the results of machine learning algorithms for classifying the sentiment of Twitter messages using a novel feature vector. Our training data consists of publicly available tweets obtained through automated means. We show that machine learning algorithms (Naive Bayes, Maximum Entropy, and SVM) can achieve competitive accuracy when trained using feature vector and the publicly available dataset. Our project also describes the preprocessing steps of the dataset needed in order to achieve high accuracy. The main contribution is the novel feature vector of weighted unigrams used to train the machine learning classifiers.

DIVISION OF WORK AMONGST THE GROUP MEMBERS

The project has been built with equal contribution from all the group members and a lot of time and effort has been put into collaborating common slots to build the project collectively as a group. Each and every member has contributed in the implementation of every feature in this project.

On a broad basis,

Pratigya Agarwal -14103180

- **Data collection of twitter and news data using scraping.**
- **Data cleaning and pre-processing and algorithms (Naive Bayes,Support VectorMachine,Maximum entropy and Modified polarity lexicon method) implementation.**

Gaurav Singh -14103159

- **Front end website designing.**
- **Resultant pie charts and data dictionary of positive, negative words from the tweet data.**
- **Preparation of training dataset for sentiment analysis.**

Aman -14103168

- **Data Dictionary preparation of positive and negative tweets.**
- **Preparation of training dataset for sentiment analysis and data loading in hive table.**

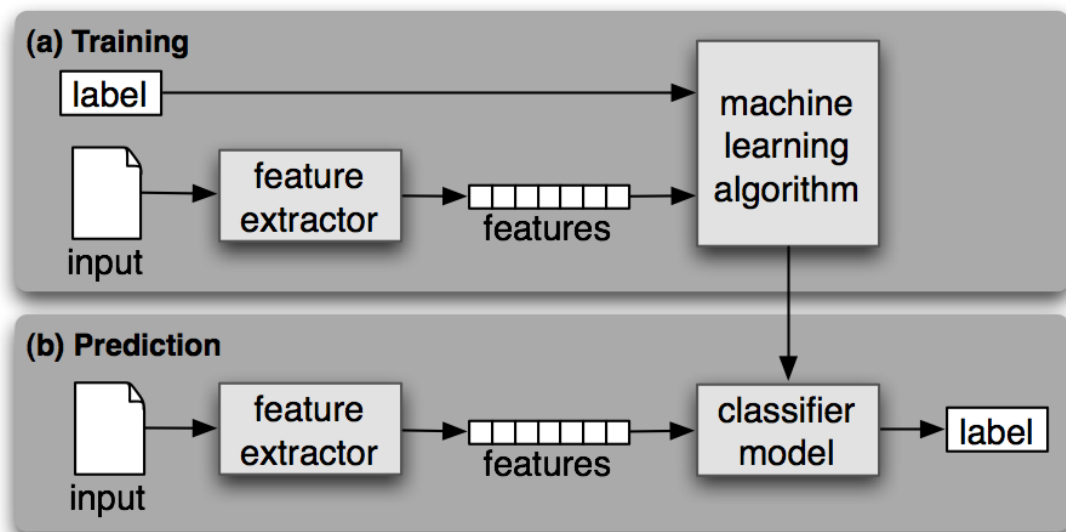
BACKGROUND STUDY

- **In order to build a sentiment analyzer, first we need to equip ourselves with the right tools and methods. Machine learning is one such tool where people have developed various methods to classify. Classifiers may or may not need training data. In particular, we will deal with the following machine learning classifiers, namely, Naive Bayes Classifier, Maximum Entropy Classifier and Support Vector Machines. All of these classifiers require training data and hence these methods fall under the category of supervised classification.**
- **After reading numerous research papers related to predictive analytics and text mining, we found out that**
- **A large amount of data generated today is in unstructured form.**
- **To extract useful information from it, we have to perform text mining techniques.**
- **In text mining, we have to clean the data and save it in a particular format so that information can be obtained from it.**
- **We learn about various classification algorithms like naive bayes, svm and maximum entropy classification.**
- **Also we learnt about various data collection technologies like selenium ,beautiful soup for web scraping.**

- **We learnt to use and work on anaconda framework using pandas and sklearn libraries.**
- **We learn to connect hive with python and running map reduce jobs from python script.**
- **We learn how to use pandas data frame technology to visualize the results obtained after implementing algorithms.**
- **We learn to apply modified polarity lexicon method to predict the results and percentage of probability of winning of a party.**

FINDINGS

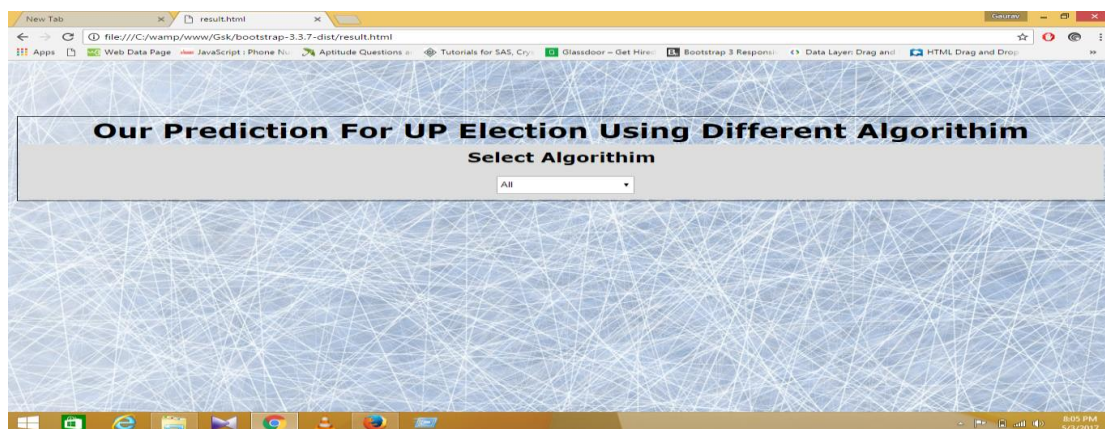
Since social media is usually formed and constructed by daily and continuous communication between participants, we have decided to investigate its potential in predicting real-world outcome. That is, using the information posted by social media participants to detect their sentiment or opinion about the different 2017 UP political parties. The sentiment used to predict the outcome of the election.



DESIGNING

Website to display final result of our prediction.

FRONTEND OF THE WEBSITE



W Naive Bayes classifier - V X Facebook X const.html X Gaurav

file:///E:/minor%20project/Gsk/bootstrap-3.3.7-dist/const.html

Apps Web Data Page JavaScript: Phone No. Aptitude Questions Tutorials for SAS, Cry Glassdoor - Get Hire Android Tutorial Getting Started | And Learn Python The Ha CodeMirror

List Of Constituency

Uttar Pradesh Result Status	
Total no of Constituencies:403	
Name of Constituency	Constituency No.
Lucknow Cantt.	175
Agra Cantt.	87
Agra North	89
Akbarpur	281
Aliganj	289
Alapur	279
Azamgarh	247
Ayodhya	275
Atroulia	73
Babaganj	245
Baheru	233
Bairia	222
badaun	275
Badapur	364
Baheri	118
Bahraich	286
Bairia	363
Lucknow Cantt.	175
Lucknow Cantt.	175
Lucknow Cantt.	175
Lucknow Cantt.	175

Windows Taskbar: 1:21 AM 3/25/2017

EXIT POLLS RESULTS

W Naive Bayes classifier - V X Facebook X exitpoll.html X Gaurav

file:///E:/minor%20project/Gsk/bootstrap-3.3.7-dist/exitpoll.html

Apps Web Data Page JavaScript: Phone No. Aptitude Questions Tutorials for SAS, Cry Glassdoor - Get Hire Android Tutorial Getting Started | And Learn Python The Ha CodeMirror

Uttar Pradesh 2017 Election Results|Exit Polls|Opinion Poll Results-SP,BSP,BJP,Congress

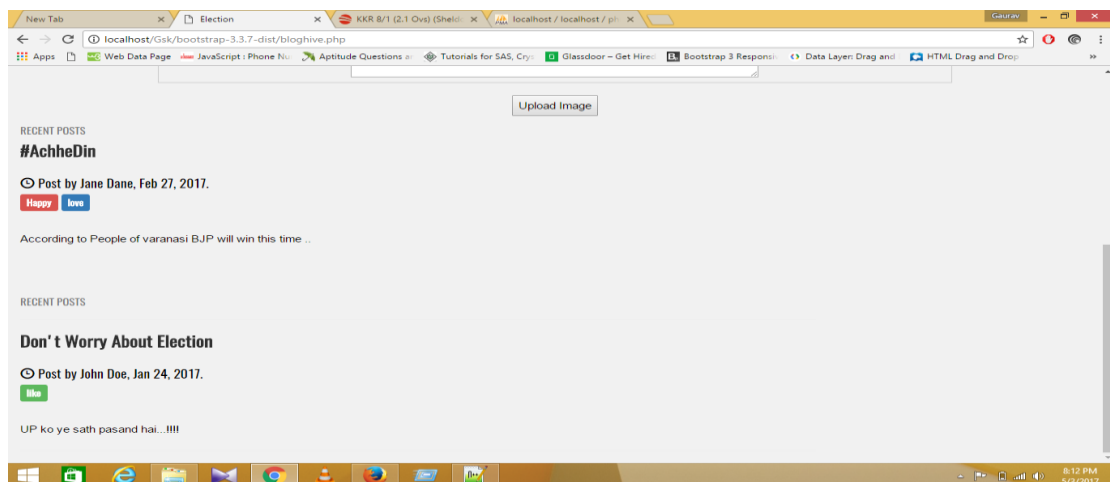
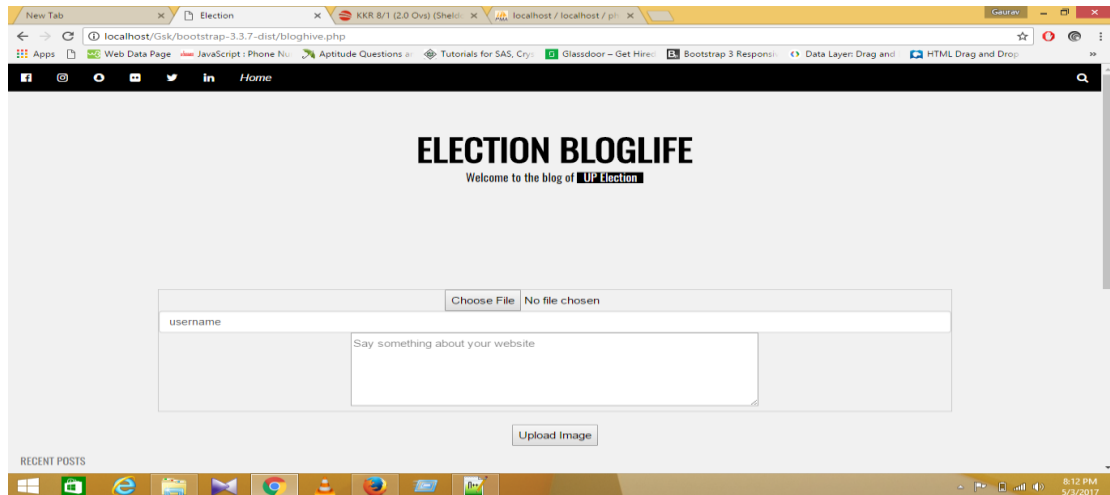
Pollsters	BJP	SP+Cong	BSP	Others
TIMES NOW VMR	190-210	110-130	57-74	19
India News MRC	185	120	90	8
ABP-CSDS	164-176	156-169	60-72	2-6
India TV-CVoter	155-167	135-147	81-93	8-20
News 24 Chankya	285	88	27	3
Axix-My-India	185	173	41	4

Activate Windows
Go to PC settings to activate Windows.

Windows Taskbar: 1:22 AM 3/25/2017

BLOG

- Blog is provided where users can share their opinions to others.



Step by step implementation

Implementation Details

We will be using Python (3.x) along with the Natural Language Toolkit (nltk) and libsvm libraries to implement the classifiers.

- **Data collection-**

Data is collected by web scraping. A technique used to collect information from web pages. We have collected data from 1st October until 8th march for our analysis.

Website: Twitter

News Website: Indian express

Language used: Python

Selenium Web driver for scrolling through the web page


Libraries used:

- of twitter
- **Beautiful soup** for collecting data from multiple pages in next link.

- **Format for data storage:**

Text file storage for tweets and news website data.

CODE:-

A screenshot of a code editor with multiple tabs. The active tab is 'data_clean1.py'. The code is a Python script using Selenium to scrape tweets from a search results page. It imports 'time' and 'Keys' from 'selenium.webdriver.common'. It sets up a Chrome browser with a specific driver path. The script navigates to a Twitter search URL, scrolls down 1800 times, finds tweet text elements using a CSS selector, and writes the text to a file named 'foo.txt'.

```
import time
from selenium import webdriver
from selenium.webdriver.common.keys import Keys

browser=webdriver.Chrome("C:\\Selenium\\chromedriver.exe")
base_url= u'https://twitter.com/search?q='
query =u'%23Upelections%20until%3A2017-03-10&f=realtime'
url=base_url+query

browser.get(url)
time.sleep(20)

body=browser.find_element_by_tag_name('body')

for _ in range(1800):
    body.send_keys(Keys.PAGE_DOWN)
    time.sleep(0.3)

tweets =browser.find_elements_by_css_selector('.TweetTextSize.js-tweet-text.tweet-text')
fo = open("foo.txt", "w",encoding='utf8')
for tweet in tweets :
    print(tweet.text)
    # fo.write(tweet.text + "\n"+"\n"+'$')
fo.close()

# Close opened file
```

1.Data Cleaning:-

One of the first steps in working with text data is to pre-process it. It is an essential step before the data is ready for analysis. Majority of available text data is highly unstructured and noisy in nature – to achieve better insights or to build better algorithms, it is necessary to play with clean data. For example, social media data is highly unstructured – it is an informal communication – typos, bad grammar, usage of slang, presence of unwanted content like stop-words , special characters etc. are the usual suspects.

Preprocess tweets

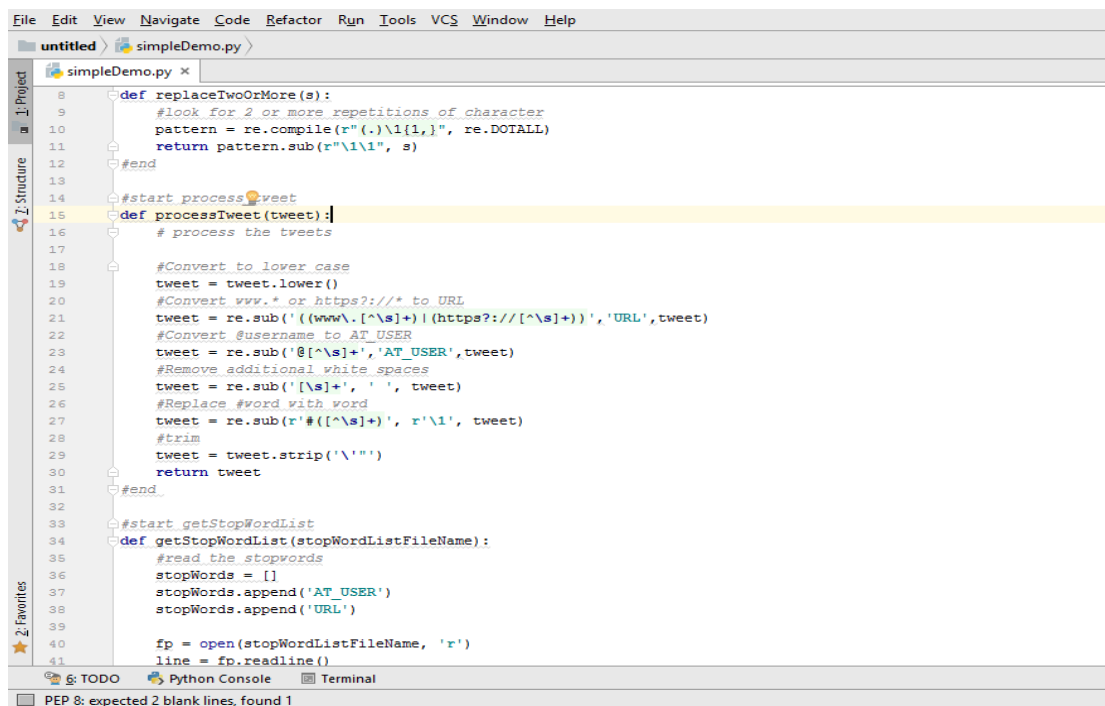
- Lower Case - Convert the tweets to lower case.
- URLs - I don't intend to follow the short urls and determine the content of the site, so we can eliminate all of these URLs via regular expression matching or replace with generic word URL.

- @username - we can eliminate "@username" via regex matching or replace it with generic word AT_USER.
- #hashtag - hash tags can give us some useful information, so it is useful to replace them with the exact same word without the hash. E.g. #BJP replaced with 'BJP'.
- Punctuations and additional white spaces - remove punctuation at the start and ending of the tweets. It is also helpful to replace multiple whitespaces with a single whitespace.

Language used: Python

Library used: Regex

Code:-



```

File Edit View Navigate Code Refactor Run Tools VCS Window Help
untitled simpleDemo.py
simpleDemo.py x
8 def replaceTwoOrMore(s):
9     #look for 2 or more repetitions of character
10    pattern = re.compile(r"(\1){1,}", re.DOTALL)
11    return pattern.sub(r"\1", s)
12    #end
13
14    #start process tweet
15    def processTweet(tweet):
16        # process the tweets
17
18        #Convert to lower case
19        tweet = tweet.lower()
20        #Convert www.* or https?/* to URL
21        tweet = re.sub('((www\.[^\s]*)|(https?:\/\/[^\s]*))', 'URL', tweet)
22        #Convert @username to AT_USER
23        tweet = re.sub('@[^\s]*+', 'AT_USER', tweet)
24        #Remove additional white spaces
25        tweet = re.sub('[\s]+', ' ', tweet)
26        #Replace #word with word
27        tweet = re.sub(r'#([^\s]+)', r'\1', tweet)
28        #trim
29        tweet = tweet.strip('\'\"')
30        return tweet
31    #end
32
33    #start getStopWordList
34    def getStopWordList(stopWordListFileName):
35        #read the stopwords
36        stopWords = []
37        stopWords.append('AT_USER')
38        stopWords.append('URL')
39
40        fp = open(stopWordListFileName, 'r')
41        line = fp.readline()
42
43    #end
44
45    #start main
46    #read the tweets
47    #process the tweets
48    #get the stop word list
49    #remove the stop words
50    #print the cleaned tweets
51
52    #end
53
54    #start main
55    #read the tweets
56    #process the tweets
57    #get the stop word list
58    #remove the stop words
59    #print the cleaned tweets
60
61    #end
62
63    #start main
64    #read the tweets
65    #process the tweets
66    #get the stop word list
67    #remove the stop words
68    #print the cleaned tweets
69
70    #end
71
72    #start main
73    #read the tweets
74    #process the tweets
75    #get the stop word list
76    #remove the stop words
77    #print the cleaned tweets
78
79    #end
80
81    #start main
82    #read the tweets
83    #process the tweets
84    #get the stop word list
85    #remove the stop words
86    #print the cleaned tweets
87
88    #end
89
90    #start main
91    #read the tweets
92    #process the tweets
93    #get the stop word list
94    #remove the stop words
95    #print the cleaned tweets
96
97    #end
98
99    #start main
100   #read the tweets
101   #process the tweets
102   #get the stop word list
103   #remove the stop words
104   #print the cleaned tweets
105
106   #end
107
108   #start main
109   #read the tweets
110   #process the tweets
111   #get the stop word list
112   #remove the stop words
113   #print the cleaned tweets
114
115   #end
116
117   #start main
118   #read the tweets
119   #process the tweets
120   #get the stop word list
121   #remove the stop words
122   #print the cleaned tweets
123
124   #end
125
126   #start main
127   #read the tweets
128   #process the tweets
129   #get the stop word list
130   #remove the stop words
131   #print the cleaned tweets
132
133   #end
134
135   #start main
136   #read the tweets
137   #process the tweets
138   #get the stop word list
139   #remove the stop words
140   #print the cleaned tweets
141
142   #end
143
144   #start main
145   #read the tweets
146   #process the tweets
147   #get the stop word list
148   #remove the stop words
149   #print the cleaned tweets
150
151   #end
152
153   #start main
154   #read the tweets
155   #process the tweets
156   #get the stop word list
157   #remove the stop words
158   #print the cleaned tweets
159
160   #end
161
162   #start main
163   #read the tweets
164   #process the tweets
165   #get the stop word list
166   #remove the stop words
167   #print the cleaned tweets
168
169   #end
170
171   #start main
172   #read the tweets
173   #process the tweets
174   #get the stop word list
175   #remove the stop words
176   #print the cleaned tweets
177
178   #end
179
180   #start main
181   #read the tweets
182   #process the tweets
183   #get the stop word list
184   #remove the stop words
185   #print the cleaned tweets
186
187   #end
188
189   #start main
190   #read the tweets
191   #process the tweets
192   #get the stop word list
193   #remove the stop words
194   #print the cleaned tweets
195
196   #end
197
198   #start main
199   #read the tweets
200   #process the tweets
201   #get the stop word list
202   #remove the stop words
203   #print the cleaned tweets
204
205   #end
206
207   #start main
208   #read the tweets
209   #process the tweets
210   #get the stop word list
211   #remove the stop words
212   #print the cleaned tweets
213
214   #end
215
216   #start main
217   #read the tweets
218   #process the tweets
219   #get the stop word list
220   #remove the stop words
221   #print the cleaned tweets
222
223   #end
224
225   #start main
226   #read the tweets
227   #process the tweets
228   #get the stop word list
229   #remove the stop words
230   #print the cleaned tweets
231
232   #end
233
234   #start main
235   #read the tweets
236   #process the tweets
237   #get the stop word list
238   #remove the stop words
239   #print the cleaned tweets
240
241   #end
242
243   #start main
244   #read the tweets
245   #process the tweets
246   #get the stop word list
247   #remove the stop words
248   #print the cleaned tweets
249
250   #end
251
252   #start main
253   #read the tweets
254   #process the tweets
255   #get the stop word list
256   #remove the stop words
257   #print the cleaned tweets
258
259   #end
260
261   #start main
262   #read the tweets
263   #process the tweets
264   #get the stop word list
265   #remove the stop words
266   #print the cleaned tweets
267
268   #end
269
270   #start main
271   #read the tweets
272   #process the tweets
273   #get the stop word list
274   #remove the stop words
275   #print the cleaned tweets
276
277   #end
278
279   #start main
280   #read the tweets
281   #process the tweets
282   #get the stop word list
283   #remove the stop words
284   #print the cleaned tweets
285
286   #end
287
288   #start main
289   #read the tweets
290   #process the tweets
291   #get the stop word list
292   #remove the stop words
293   #print the cleaned tweets
294
295   #end
296
297   #start main
298   #read the tweets
299   #process the tweets
300   #get the stop word list
301   #remove the stop words
302   #print the cleaned tweets
303
304   #end
305
306   #start main
307   #read the tweets
308   #process the tweets
309   #get the stop word list
310   #remove the stop words
311   #print the cleaned tweets
312
313   #end
314
315   #start main
316   #read the tweets
317   #process the tweets
318   #get the stop word list
319   #remove the stop words
320   #print the cleaned tweets
321
322   #end
323
324   #start main
325   #read the tweets
326   #process the tweets
327   #get the stop word list
328   #remove the stop words
329   #print the cleaned tweets
330
331   #end
332
333   #start main
334   #read the tweets
335   #process the tweets
336   #get the stop word list
337   #remove the stop words
338   #print the cleaned tweets
339
340   #end
341
342   #start main
343   #read the tweets
344   #process the tweets
345   #get the stop word list
346   #remove the stop words
347   #print the cleaned tweets
348
349   #end
350
351   #start main
352   #read the tweets
353   #process the tweets
354   #get the stop word list
355   #remove the stop words
356   #print the cleaned tweets
357
358   #end
359
360   #start main
361   #read the tweets
362   #process the tweets
363   #get the stop word list
364   #remove the stop words
365   #print the cleaned tweets
366
367   #end
368
369   #start main
370   #read the tweets
371   #process the tweets
372   #get the stop word list
373   #remove the stop words
374   #print the cleaned tweets
375
376   #end
377
378   #start main
379   #read the tweets
380   #process the tweets
381   #get the stop word list
382   #remove the stop words
383   #print the cleaned tweets
384
385   #end
386
387   #start main
388   #read the tweets
389   #process the tweets
390   #get the stop word list
391   #remove the stop words
392   #print the cleaned tweets
393
394   #end
395
396   #start main
397   #read the tweets
398   #process the tweets
399   #get the stop word list
400   #remove the stop words
401   #print the cleaned tweets
402
403   #end
404
405   #start main
406   #read the tweets
407   #process the tweets
408   #get the stop word list
409   #remove the stop words
410   #print the cleaned tweets
411
412   #end
413
414   #start main
415   #read the tweets
416   #process the tweets
417   #get the stop word list
418   #remove the stop words
419   #print the cleaned tweets
420
421   #end
422
423   #start main
424   #read the tweets
425   #process the tweets
426   #get the stop word list
427   #remove the stop words
428   #print the cleaned tweets
429
430   #end
431
432   #start main
433   #read the tweets
434   #process the tweets
435   #get the stop word list
436   #remove the stop words
437   #print the cleaned tweets
438
439   #end
440
441   #start main
442   #read the tweets
443   #process the tweets
444   #get the stop word list
445   #remove the stop words
446   #print the cleaned tweets
447
448   #end
449
450   #start main
451   #read the tweets
452   #process the tweets
453   #get the stop word list
454   #remove the stop words
455   #print the cleaned tweets
456
457   #end
458
459   #start main
460   #read the tweets
461   #process the tweets
462   #get the stop word list
463   #remove the stop words
464   #print the cleaned tweets
465
466   #end
467
468   #start main
469   #read the tweets
470   #process the tweets
471   #get the stop word list
472   #remove the stop words
473   #print the cleaned tweets
474
475   #end
476
477   #start main
478   #read the tweets
479   #process the tweets
480   #get the stop word list
481   #remove the stop words
482   #print the cleaned tweets
483
484   #end
485
486   #start main
487   #read the tweets
488   #process the tweets
489   #get the stop word list
490   #remove the stop words
491   #print the cleaned tweets
492
493   #end
494
495   #start main
496   #read the tweets
497   #process the tweets
498   #get the stop word list
499   #remove the stop words
500   #print the cleaned tweets
501
502   #end
503
504   #start main
505   #read the tweets
506   #process the tweets
507   #get the stop word list
508   #remove the stop words
509   #print the cleaned tweets
510
511   #end
512
513   #start main
514   #read the tweets
515   #process the tweets
516   #get the stop word list
517   #remove the stop words
518   #print the cleaned tweets
519
520   #end
521
522   #start main
523   #read the tweets
524   #process the tweets
525   #get the stop word list
526   #remove the stop words
527   #print the cleaned tweets
528
529   #end
530
531   #start main
532   #read the tweets
533   #process the tweets
534   #get the stop word list
535   #remove the stop words
536   #print the cleaned tweets
537
538   #end
539
540   #start main
541   #read the tweets
542   #process the tweets
543   #get the stop word list
544   #remove the stop words
545   #print the cleaned tweets
546
547   #end
548
549   #start main
550   #read the tweets
551   #process the tweets
552   #get the stop word list
553   #remove the stop words
554   #print the cleaned tweets
555
556   #end
557
558   #start main
559   #read the tweets
560   #process the tweets
561   #get the stop word list
562   #remove the stop words
563   #print the cleaned tweets
564
565   #end
566
567   #start main
568   #read the tweets
569   #process the tweets
570   #get the stop word list
571   #remove the stop words
572   #print the cleaned tweets
573
574   #end
575
576   #start main
577   #read the tweets
578   #process the tweets
579   #get the stop word list
580   #remove the stop words
581   #print the cleaned tweets
582
583   #end
584
585   #start main
586   #read the tweets
587   #process the tweets
588   #get the stop word list
589   #remove the stop words
590   #print the cleaned tweets
591
592   #end
593
594   #start main
595   #read the tweets
596   #process the tweets
597   #get the stop word list
598   #remove the stop words
599   #print the cleaned tweets
600
601   #end
602
603   #start main
604   #read the tweets
605   #process the tweets
606   #get the stop word list
607   #remove the stop words
608   #print the cleaned tweets
609
610   #end
611
612   #start main
613   #read the tweets
614   #process the tweets
615   #get the stop word list
616   #remove the stop words
617   #print the cleaned tweets
618
619   #end
620
621   #start main
622   #read the tweets
623   #process the tweets
624   #get the stop word list
625   #remove the stop words
626   #print the cleaned tweets
627
628   #end
629
630   #start main
631   #read the tweets
632   #process the tweets
633   #get the stop word list
634   #remove the stop words
635   #print the cleaned tweets
636
637   #end
638
639   #start main
640   #read the tweets
641   #process the tweets
642   #get the stop word list
643   #remove the stop words
644   #print the cleaned tweets
645
646   #end
647
648   #start main
649   #read the tweets
650   #process the tweets
651   #get the stop word list
652   #remove the stop words
653   #print the cleaned tweets
654
655   #end
656
657   #start main
658   #read the tweets
659   #process the tweets
660   #get the stop word list
661   #remove the stop words
662   #print the cleaned tweets
663
664   #end
665
666   #start main
667   #read the tweets
668   #process the tweets
669   #get the stop word list
670   #remove the stop words
671   #print the cleaned tweets
672
673   #end
674
675   #start main
676   #read the tweets
677   #process the tweets
678   #get the stop word list
679   #remove the stop words
680   #print the cleaned tweets
681
682   #end
683
684   #start main
685   #read the tweets
686   #process the tweets
687   #get the stop word list
688   #remove the stop words
689   #print the cleaned tweets
690
691   #end
692
693   #start main
694   #read the tweets
695   #process the tweets
696   #get the stop word list
697   #remove the stop words
698   #print the cleaned tweets
699
700   #end
701
702   #start main
703   #read the tweets
704   #process the tweets
705   #get the stop word list
706   #remove the stop words
707   #print the cleaned tweets
708
709   #end
710
711   #start main
712   #read the tweets
713   #process the tweets
714   #get the stop word list
715   #remove the stop words
716   #print the cleaned tweets
717
718   #end
719
720   #start main
721   #read the tweets
722   #process the tweets
723   #get the stop word list
724   #remove the stop words
725   #print the cleaned tweets
726
727   #end
728
729   #start main
730   #read the tweets
731   #process the tweets
732   #get the stop word list
733   #remove the stop words
734   #print the cleaned tweets
735
736   #end
737
738   #start main
739   #read the tweets
740   #process the tweets
741   #get the stop word list
742   #remove the stop words
743   #print the cleaned tweets
744
745   #end
746
747   #start main
748   #read the tweets
749   #process the tweets
750   #get the stop word list
751   #remove the stop words
752   #print the cleaned tweets
753
754   #end
755
756   #start main
757   #read the tweets
758   #process the tweets
759   #get the stop word list
760   #remove the stop words
761   #print the cleaned tweets
762
763   #end
764
765   #start main
766   #read the tweets
767   #process the tweets
768   #get the stop word list
769   #remove the stop words
770   #print the cleaned tweets
771
772   #end
773
774   #start main
775   #read the tweets
776   #process the tweets
777   #get the stop word list
778   #remove the stop words
779   #print the cleaned tweets
780
781   #end
782
783   #start main
784   #read the tweets
785   #process the tweets
786   #get the stop word list
787   #remove the stop words
788   #print the cleaned tweets
789
790   #end
791
792   #start main
793   #read the tweets
794   #process the tweets
795   #get the stop word list
796   #remove the stop words
797   #print the cleaned tweets
798
799   #end
800
801   #start main
802   #read the tweets
803   #process the tweets
804   #get the stop word list
805   #remove the stop words
806   #print the cleaned tweets
807
808   #end
809
810   #start main
811   #read the tweets
812   #process the tweets
813   #get the stop word list
814   #remove the stop words
815   #print the cleaned tweets
816
817   #end
818
819   #start main
820   #read the tweets
821   #process the tweets
822   #get the stop word list
823   #remove the stop words
824   #print the cleaned tweets
825
826   #end
827
828   #start main
829   #read the tweets
830   #process the tweets
831   #get the stop word list
832   #remove the stop words
833   #print the cleaned tweets
834
835   #end
836
837   #start main
838   #read the tweets
839   #process the tweets
840   #get the stop word list
841   #remove the stop words
842   #print the cleaned tweets
843
844   #end
845
846   #start main
847   #read the tweets
848   #process the tweets
849   #get the stop word list
850   #remove the stop words
851   #print the cleaned tweets
852
853   #end
854
855   #start main
856   #read the tweets
857   #process the tweets
858   #get the stop word list
859   #remove the stop words
860   #print the cleaned tweets
861
862   #end
863
864   #start main
865   #read the tweets
866   #process the tweets
867   #get the stop word list
868   #remove the stop words
869   #print the cleaned tweets
870
871   #end
872
873   #start main
874   #read the tweets
875   #process the tweets
876   #get the stop word list
877   #remove the stop words
878   #print the cleaned tweets
879
880   #end
881
882   #start main
883   #read the tweets
884   #process the tweets
885   #get the stop word list
886   #remove the stop words
887   #print the cleaned tweets
888
889   #end
890
891   #start main
892   #read the tweets
893   #process the tweets
894   #get the stop word list
895   #remove the stop words
896   #print the cleaned tweets
897
898   #end
899
900   #start main
901   #read the tweets
902   #process the tweets
903   #get the stop word list
904   #remove the stop words
905   #print the cleaned tweets
906
907   #end
908
909   #start main
910   #read the tweets
911   #process the tweets
912   #get the stop word list
913   #remove the stop words
914   #print the cleaned tweets
915
916   #end
917
918   #start main
919   #read the tweets
920   #process the tweets
921   #get the stop word list
922   #remove the stop words
923   #print the cleaned tweets
924
925   #end
926
927   #start main
928   #read the tweets
929   #process the tweets
930   #get the stop word list
931   #remove the stop words
932   #print the cleaned tweets
933
934   #end
935
936   #start main
937   #read the tweets
938   #process the tweets
939   #get the stop word list
940   #remove the stop words
941   #print the cleaned tweets
942
943   #end
944
945   #start main
946   #read the tweets
947   #process the tweets
948   #get the stop word list
949   #remove the stop words
950   #print the cleaned tweets
951
952   #end
953
954   #start main
955   #read the tweets
956   #process the tweets
957   #get the stop word list
958   #remove the stop words
959   #print the cleaned tweets
960
961   #end
962
963   #start main
964   #read the tweets
965   #process the tweets
966   #get the stop word list
967   #remove the stop words
968   #print the cleaned tweets
969
970   #end
971
972   #start main
973   #read the tweets
974   #process the tweets
975   #get the stop word list
976   #remove the stop words
977   #print the cleaned tweets
978
979   #end
980
981   #start main
982   #read the tweets
983   #process the tweets
984   #get the stop word list
985   #remove the stop words
986   #print the cleaned tweets
987
988   #end
989
990   #start main
991   #read the tweets
992   #process the tweets
993   #get the stop word list
994   #remove the stop words
995   #print the cleaned tweets
996
997   #end
998
999   #start main
1000  #read the tweets
1001  #process the tweets
1002  #get the stop word list
1003  #remove the stop words
1004  #print the cleaned tweets
1005
1006  #end
1007
1008  #start main
1009  #read the tweets
1010  #process the tweets
1011  #get the stop word list
1012  #remove the stop words
1013  #print the cleaned tweets
1014
1015  #end
1016
1017  #start main
1018  #read the tweets
1019  #process the tweets
1020  #get the stop word list
1021  #remove the stop words
1022  #print the cleaned tweets
1023
1024  #end
1025
1026  #start main
1027  #read the tweets
1028  #process the tweets
1029  #get the stop word list
1030  #remove the stop words
1031  #print the cleaned tweets
1032
1033  #end
1034
1035  #start main
1036  #read the tweets
1037  #process the tweets
1038  #get the stop word list
1039  #remove the stop words
1040  #print the cleaned tweets
1041
1042  #end
1043
1044  #start main
1045  #read the tweets
1046  #process the tweets
1047  #get the stop word list
1048  #remove the stop words
1049  #print the cleaned tweets
1050
1051  #end
1052
1053  #start main
1054  #read the tweets
1055  #process the tweets
1056  #get the stop word list
1057  #remove the stop words
1058  #print the cleaned tweets
1059
1060  #end
1061
1062  #start main
1063  #read the tweets
1064  #process the tweets
1065  #get the stop word list
1066  #remove the stop words
1067  #print the cleaned tweets
1068
1069  #end
1070
1071  #start main
1072  #read the tweets
1073  #process the tweets
1074  #get the stop word list
1075  #remove the stop words
1076  #print the cleaned tweets
1077
1078  #end
1079
1080  #start main
1081  #read the tweets
1082  #process the tweets
1083  #get the stop word list
1084  #remove the stop words
1085  #print the cleaned tweets
1086
1087  #end
1088
1089  #start main
1090  #read the tweets
1091  #process the tweets
1092  #get the stop word list
1093  #remove the stop words
1094  #print the cleaned tweets
1095
1096  #end
1097
1098  #start main
1099  #read the tweets
1100  #process the tweets
1101  #get the stop word list
1102  #remove the stop words
1103  #print the cleaned tweets
1104
1105  #end
1106
1107  #start main
1108  #read the tweets
1109  #process the tweets
1110  #get the stop word list
1111  #remove the stop words
1112  #print the cleaned tweets
1113
1114  #end
1115
1116  #start main
1117  #read the tweets
1118  #process the tweets
1119  #get the stop word list
1120  #remove the stop words
1121  #print the cleaned tweets
1122
1123  #end
1124
1125  #start main
1126  #read the tweets
1127  #process the tweets
1128  #get the stop word list
1129  #remove the stop words
1130  #print the cleaned tweets
1131
1132  #end
1133
1134  #start main
1135  #read the tweets
1136  #process the tweets
1137  #get the stop word list
1138  #remove the stop words
1139  #print the cleaned tweets
1140
1141  #end
1142
1143  #start main
1144  #read the tweets
1145  #process the tweets
1146  #get the stop word list
1147  #remove the stop words
1148  #print the cleaned tweets
1149
1150  #end
1151
1152  #start main
1153  #read the tweets
1154  #process the tweets
1155  #get the stop word list
1156  #remove the stop words
1157  #print the cleaned tweets
1158
1159  #end
1160
1161  #start main
1162  #read the tweets
1163  #process the tweets
1164  #get the stop word list
1165  #remove the stop words
1166  #print the cleaned tweets
1167
1168  #end
1169
1170  #start main
1171  #read the tweets
1172  #process the tweets
1173  #get the stop word list
1174  #remove the stop words
1175  #print the cleaned tweets
1176
1177  #end
1178
1179  #start main
1180  #read the tweets
1181  #process the tweets
1182  #get the stop word list
1183  #remove the stop words
1184  #print the cleaned tweets
1185
1186  #end
1187
1188  #start main
1189  #read the tweets
1190  #process the tweets
1191  #get the stop word list
1192  #remove the stop words
1193  #print the cleaned tweets
1194
1195  #end
1196
1197  #start main
1198  #read the tweets
1199  #process the tweets
1200  #get the stop word list
1201  #remove the stop words
1202  #print the cleaned tweets
1203
1204  #end
1205
1206  #start main
1207  #read the tweets
1208  #process the tweets
1209  #get the stop word list
1210  #remove the stop words
1211  #print the cleaned tweets
1212
1213  #end
1214
1215  #start main
1216  #read the tweets
1217  #process the tweets
1218  #get the stop word list
1219  #remove the stop words
1220  #print the cleaned tweets
1221
1222  #end
1223
1224  #start main
1225  #read the tweets
1226  #process the tweets
1227  #get the stop word list
1228  #remove the stop words
1229  #print the cleaned tweets
1230
1231  #end
1232
1233  #start main
1234  #read the tweets
1235  #process the tweets
1236  #get the stop word list
1237  #remove the stop words
1238  #print the cleaned tweets
1239
1240  #end
1241
1242  #start main
1243  #read the tweets
1244  #process the tweets
1245  #get the stop word list
1246  #remove the stop words
1247  #print the cleaned tweets
1248
1249  #end
1250
1251  #start main
1252  #read the tweets
1253  #process the tweets
1254  #get the stop word list
1255  #remove the stop words
1256  #print the cleaned tweets
1257
1258  #end
1259
1260  #start main
1261  #read the tweets
1262  #process the tweets
1263  #get the stop word list
1264  #remove the stop words
1265  #print the cleaned tweets
1266
1267  #end
1268
1269  #start main
1270  #read the tweets
1271  #process the tweets
1272  #get the stop word list
1273  #remove the stop words
1274  #print the cleaned tweets
1275
1276  #end
1277
1278  #start main
1279  #read the tweets
1280  #process the tweets
1281  #get the stop word list
1282  #remove the stop words
1283  #print the cleaned tweets
1284
1285  #end
1286
1287  #start main
1288  #read the tweets
1289  #process the tweets
1290  #get the stop word list
1291  #remove the stop words
1292  #print the cleaned tweets
1293
1294  #end
1295
1296  #start main
1297  #read the tweets
1298  #process the tweets
1299  #get the stop word list
1300  #remove the stop words
1301  #print the cleaned tweets
1302
1303  #end
1304
1305  #start main
1306  #read the tweets
1307  #process the tweets
1308  #get the stop word list
1309  #remove the stop words
1310  #print the cleaned tweets
1311
1312  #end
1313
1314  #start main
1315  #read the tweets
1316  #process the tweets
1317  #get the stop word list
1318  #remove the stop words
1319  #print the cleaned tweets
1320
1321  #end
1322
1323  #start main
1324  #read the tweets
1325  #process the tweets
1326  #get the stop word list
1327  #remove the stop words
1328  #print the cleaned tweets
1329
1330  #end
1331
1332  #start main
1333  #read the tweets
1334  #process the tweets
1335  #get the stop word list
1336  #remove the stop words
1337  #print the cleaned tweets
1338
1339  #end
1340
1341  #start main
1342  #read the tweets
1343  #process the tweets
1344  #get the stop word list
1345  #remove the stop words
1346  #print the cleaned tweets
1347
1348  #end
1349
1350  #start main
1351  #read the tweets
1352  #process the tweets
1353  #get the stop word list
1354  #remove the stop words
1355  #print the cleaned tweets
1356
1357  #end
1358
1359  #start main
1360  #read the tweets
1361  #process the tweets
1362  #get the stop word list
1363  #remove the stop words
1364  #print the cleaned tweets
1365
1366  #end
1367
1368  #start main
1369  #read the tweets
1370  #process the tweets
1371  #get the stop word list
1372  #remove the stop words
1373  #print the cleaned tweets
1374
1375  #end
1376
1377  #start main
1378  #read the tweets
1379  #process the tweets
1380  #get the stop word list
1381  #remove the stop words
1382  #print the cleaned tweets
1383
1384  #end
1385
1386  #start main
1387  #read the tweets
1388  #process the tweets
1389  #get the stop word list
1390  #remove the stop words
1391  #print the cleaned tweets
1392
1393  #end
1394
1395  #start main
1396  #read the tweets
1397  #process the tweets
1398  #get the stop word list
1399  #remove the stop words
1400  #print the cleaned tweets
1401
1402  #end
1403
1404  #start main
1405  #read the tweets
1406  #process the tweets
1407  #get the stop word list
1408  #remove the stop words
1409  #print the cleaned tweets
1410
1411  #end
1412
1413  #start main
1414  #read the tweets
1415  #process the tweets
1416  #get the stop word list
1417  #remove the stop words
1418  #print the cleaned tweets
1419
1420  #end
1421
1422  #start main
1423  #read the tweets
1424  #process the tweets
1425  #get the stop word list
1426  #remove the stop words
1427  #print the cleaned tweets
1428
1429  #end
1430
1431  #start main
1432  #read the tweets
1433  #process the tweets
1434  #get the stop word list
1435  #remove the stop words
1436  #print the cleaned tweets
1437
1438  #end
1439
1440  #start main
1441  #read the tweets
1442  #process the tweets
1443  #
```

3.Data loading into Hadoop file storage system:-

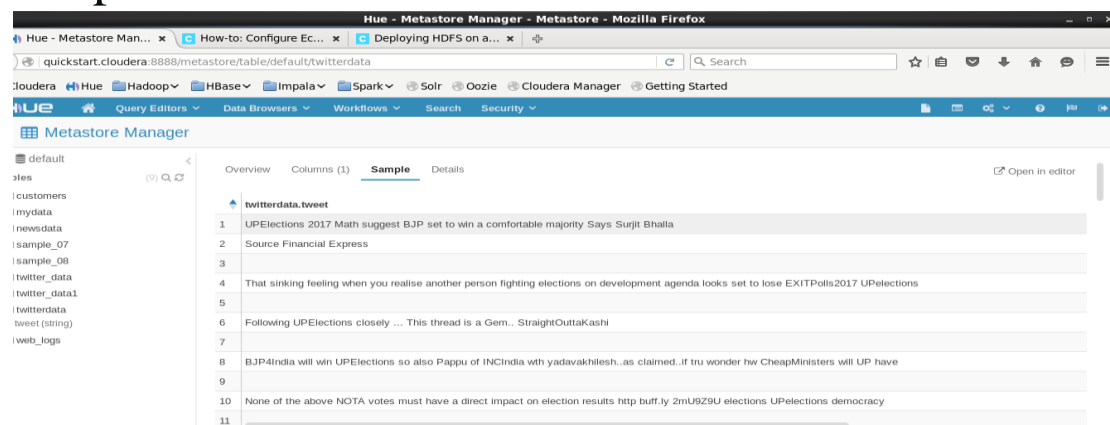
Using put command load text data files into file storage.

Command:-hdfs dfs -put textfilename.txt

4.Data loading into hive table:-

For loading data from hdfs into hive, we have used external table named twitterdata and newsdata1 using load data statement.

Snapshot:-



The screenshot shows the Hue Metastore Manager interface. On the left, a sidebar lists various databases and tables, including 'default', 'xles', 'customers', 'mydata', 'newsdata', 'sample_07', 'sample_08', 'twitter_data', 'twitter_data1', 'twitterdata', 'tweet (string)', and 'web_logs'. The main panel displays the 'Sample' tab for the 'twitterdata.tweet' table. It shows 11 rows of data, each containing a tweet text snippet. The interface includes a top navigation bar with links to 'Query Editors', 'Data Browsers', 'Workflows', 'Search', and 'Security'. The bottom of the main panel has a 'Details' tab and an 'Open in editor' link.

	twitterdata.tweet
1	UPElections 2017 Math suggest BJP set to win a comfortable majority Says Surjit Bhalla
2	Source Financial Express
3	
4	That sinking feeling when you realise another person fighting elections on development agenda looks set to lose EXITPolls2017 UPElections
5	
6	Following UPElections closely ... This thread is a Gem.. StraightOuttaKashi
7	
8	BJP4India will win UPElections so also Pappu of INCIndia with yadavakhlesh..as claimed..if tru wonder hw CheapMinisters will UP have
9	
10	None of the above NOTA votes must have a direct impact on election results http buff.ly 2mU9Z9U elections UPElections democracy
11	

5.Data dictionary of positive and negative words :-

We have prepared 4 separate text files containing positive and negative words . For opinion mining, data dictionary is needed.

[Reference:-Kabir Ismail Umar Data mining for social media analysis using twitter.]

6.Feature Vector

Feature vector is the most important concept in implementing a classifier. A good feature vector directly determines how successful your classifier will be. The feature vector is used to build a model which the classifier learns from the training data and further can be used to classify previously unseen data.

Similarly, in tweets, we can use the presence/absence of words that appear in tweet as features. In the training data, consisting of positive, negative and neutral tweets, we can split each tweet into words and add each word to the feature vector. Some of the words might not have any say in indicating the sentiment of a tweet and hence we can filter them out. Adding individual (single) words to the feature vector is referred to as 'unigrams' approach.

7.Filtering tweet words (for feature vector):

- Stop words - a, is, the, with etc. The full list of stop words can be found at **Stop Word List**. These words don't indicate any sentiment and can be removed.
- Repeating letters - if you look at the tweets, sometimes people repeat letters to stress the emotion. E.g. hunggrryyy, huuuuuuungry for 'hungry'. We can look for 2 or more repetitive letters in words and replace them by 2 of the same.
- Punctuation - we can remove punctuation such as comma, single/double quote, question marks at the start and end of each word. E.g. beautiful!!!!!! replaced with beautiful

- Words must start with an alphabet - For simplicity sake, we can remove all those words which don't start with an alphabet. E.g. 15th, 5.34am

```
simpleDemo.py x
34 def getStopWordList(stopWordListFileName):
35     #read the stopwords
36     stopWords = []
37     stopWords.append('AT_USER')
38     stopWords.append('URL')
39
40     fp = open(stopWordListFileName, 'r')
41     line = fp.readline()
42     while line:
43         word = line.strip()
44         stopWords.append(word)
45         line = fp.readline()
46     fp.close()
47     return stopWords
48 #end
49
50 #start getFeatureVector
51 def getFeatureVector(tweet, stopWords):
52     featureVector = []
53     words = tweet.split()
54     for w in words:
55         #replace two or more with two occurrences
56         w = replaceTwoOrMore(w)
57         #strip punctuation
58         w = w.strip('\\"?,.')
59         #check if it consists of only words
60         val = re.search(r"^[a-zA-Z][a-zA-Z0-9]*[a-zA-Z]+[a-zA-Z0-9]*$", w)
61         #ignore if it is a stopWord
62         if(w in stopWords or val is None):
63             continue
64         else:
65             featureVector.append(w.lower())
66     return featureVector
67 #end
```

8.Feature Extraction:

The following code, extracts the tweets and label from the csv file and processes it as above and obtains a feature vector and stores it in a variable called "tweets".

```
#Read the tweets one by one and process it
inpTweets = csv.reader(open('data/215.csv', 'r'), delimiter=',')
stopWords = getStopWordList('data/feature_list/stopwords.txt')
count = 0;
featureList = []
tweets = []
for row in inpTweets:
    sentiment = row[0]
    tweet = row[1]
    processedTweet = processTweet(tweet)
    featureVector = getFeatureVector(processedTweet, stopWords)
    featureList.extend(featureVector)
    tweets.append((featureVector, sentiment));
#end_loop
```

9.Extract Features Method

```
#start extract_features
def extract_features(tweet):
    tweet_words = set(tweet)
    features = {}
    for word in featureList:
        features['contains(%s)' % word] = (word in tweet_words)
    return features
#end
```

```
#Read the tweets one by one and process it
```

10. Naive Bayes Classifier:

For classifying the data/tweets into positive, negative and neutral. Naive Bayes is a simple model which works well on text categorization. We use a multinomial Naive Bayes model. Class c^* is assigned to tweet d , where

$$c^* = \operatorname{argmax}_c P_{NB}(c|d)$$

$$P_{NB}(c|d) := \frac{(P(c) \sum_{i=1}^m P(f_i|c)^{n_i(d)})}{P(d)}$$

In this formula, f represents a feature and $n_i(d)$ represents the count of feature f_i found in tweet d . There are a total of m features. Parameters $P(c)$ and $P(f/c)$ are obtained through maximum likelihood estimates, and add-1 smoothing is utilized for unseen features. I used the Python based Natural Language Toolkit library to train and classify using the Naive Bayes method.

Code:-

```
# Generate the training set
training_set = nltk.classify.util.apply_features(extract_features, tweets)

# Train the Naive Bayes classifier
NBClassifier = nltk.NaiveBayesClassifier.train(training_set)

# Test the classifier
list_of_parties=['BJP','Congress','SP','BSP']
iTweet= csv.reader(open('data/b.csv', 'r'),delimiter=',')
#myfile = open('data/naive_result.csv', 'w')
#wr = csv.writer(myfile, quoting=csv.QUOTE_ALL)
#testTweet = 'Disappointing day. Attended a car boot sale to raise some fund.
fp = open('final_naive_bayes.txt', 'w')
```

[Reference:-The Predictive Power of Social Media: On the Predictability of U.S.

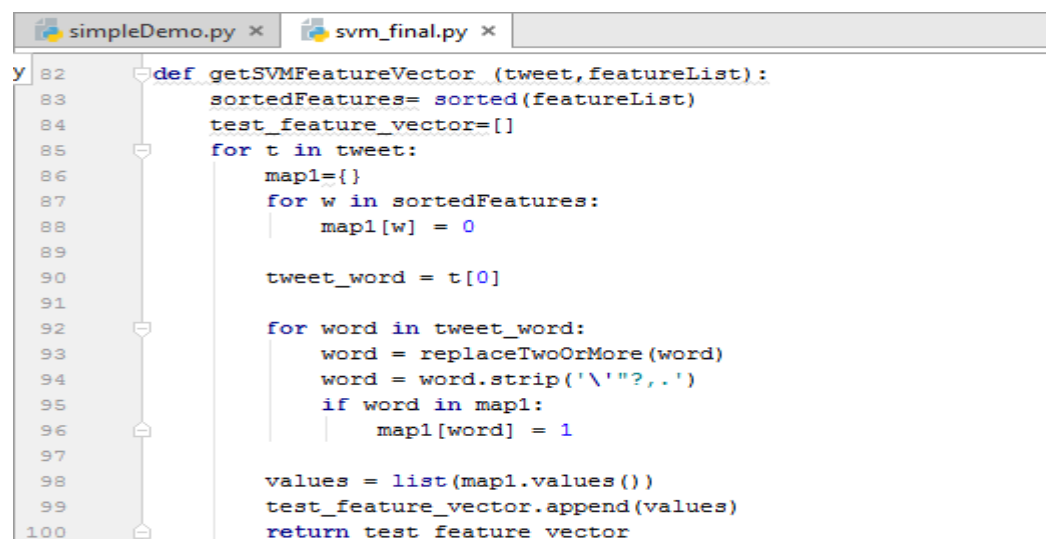
Presidential Elections using Twitter by Kazem Jahanbakhsh ,Yumi Moon]

11.Support Vector Machine:

Support Vector Machines (SVM) is pretty much the standard classifier which is used for any general purpose classification. I will use the libsvm library (written in C++ and has a python handle)

Support Vector Machines is another popular classification technique . I have used libsvm library with a linear kernel. My input data are two sets of vectors of size m. Each entry in the vector corresponds to the presence a feature. In the unigram feature extractor, each feature is a single word found in a tweet. If the feature is present, the value is 1, but if the feature is absent, then the value is 0. I use feature presence, as opposed to a count, so that I do not have to scale the input data, which speeds up overall processing .

Code:-



```
simpleDemo.py x svm_final.py x
y 82 def getSVMFeatureVector (tweet,featureList):
83     sortedFeatures= sorted(featureList)
84     test_feature_vector=[]
85     for t in tweet:
86         map1={}
87         for w in sortedFeatures:
88             map1[w] = 0
89
90         tweet_word = t[0]
91
92         for word in tweet_word:
93             word = replaceTwoOrMore(word)
94             word = word.strip('\\"?,.')
95             if word in map1:
96                 map1[word] = 1
97
98         values = list(map1.values())
99         test_feature_vector.append(values)
100     return test_feature_vector
```

12. Maximum Entropy Classifier:

We use the 'General Iterative Scaling' algorithm and stick to 10 iterations.

The idea behind Maximum Entropy models is that one should prefer the most uniform models that satisfy a given constraint . MaxEnt models are feature-based models. In a two class scenario, it is the same as using logistic regression to find a distribution over the classes. MaxEnt makes no independence assumptions for its features, unlike Naive Bayes. The model is represented by the following:

$$P_{ME}(c|d, \lambda) = \frac{\exp[\sum_i \lambda_i f_i(c, d)]}{\sum_{c'} \exp[\sum_i \lambda_i f_i(c', d)]}$$

In this formula, c is the class, d is the tweet, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the λ_i 's so as to maximize the conditional probability.

We use the Python NLTK library to train and classify using the Maximum Entropy method. For training the weights we use conjugate gradient ascent.

Theoretically, Max Entropy performs better than Naive Bayes because it handles feature overlap better. However, in practice, Naive Bayes can still perform well on a variety of problems..

```

featureList = list(set(featureList))

# Generate the training set
training_set = nltk.classify.util.apply_features(extract_features, tweets)

)# Train the Naive Bayes classifier
)#NBClassifier = nltk.NaiveBayesClassifier.train(training_set)
MaxEntClassifier = nltk.classify.maxent.MaxentClassifier.train(training_set, max_iter = 10) #, '(
#testTveet = 'just had some bloodwork done. My arm hurts'

list_of_parties=['BJP', 'Congress', 'SP', 'BSP']
iTweet= csv.reader(open('data/b.csv', 'r'),delimiter=',')
)#myfile = open('data/maxent_result.csv', 'w')
)#vr = csv.writer(myfile, quoting=csv.QUOTE_ALL)
fp = open('final_maxent.txt', 'w')

```

14.Results for three algorithms stored on hive:

We stored result data from three algorithms in 3 hive tables.

These tables are then used to retrieve no of positive and negative instances for a articular party.

15.Code for retrieving count from hive table:

We connected python with hive using pyhive package and wrote HiveQL queries in the python script.

6 map reduce jobs were run using this script.

16.Polarity lexicon model modified by Gayo-Avello et al.:- Used for prediction of the results

$$p(c_1) = \frac{pos(c_1) + neg(c_2)}{pos(c_1) + neg(c_1) + pos(c_2) + neg(c_2)}$$

c_1 =party 1

c_2 =party 2

Pos(c_1) positive words for party 1

Pos(c_2) positive words for party 2

Neg(c_1) negative words for party 1

Neg(c_2) negative words for party 2

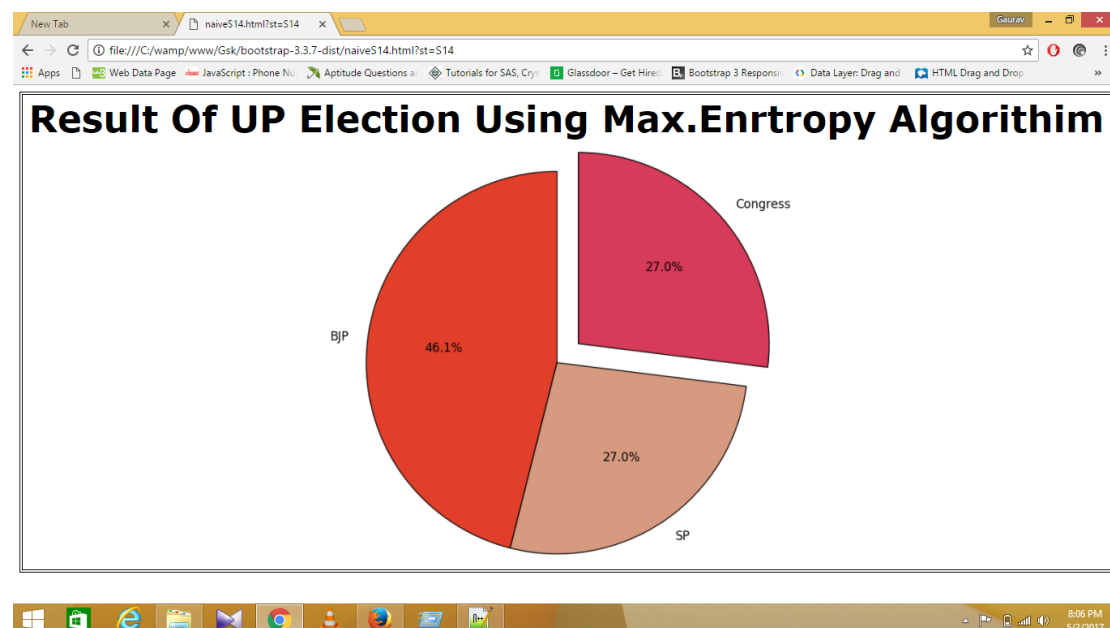
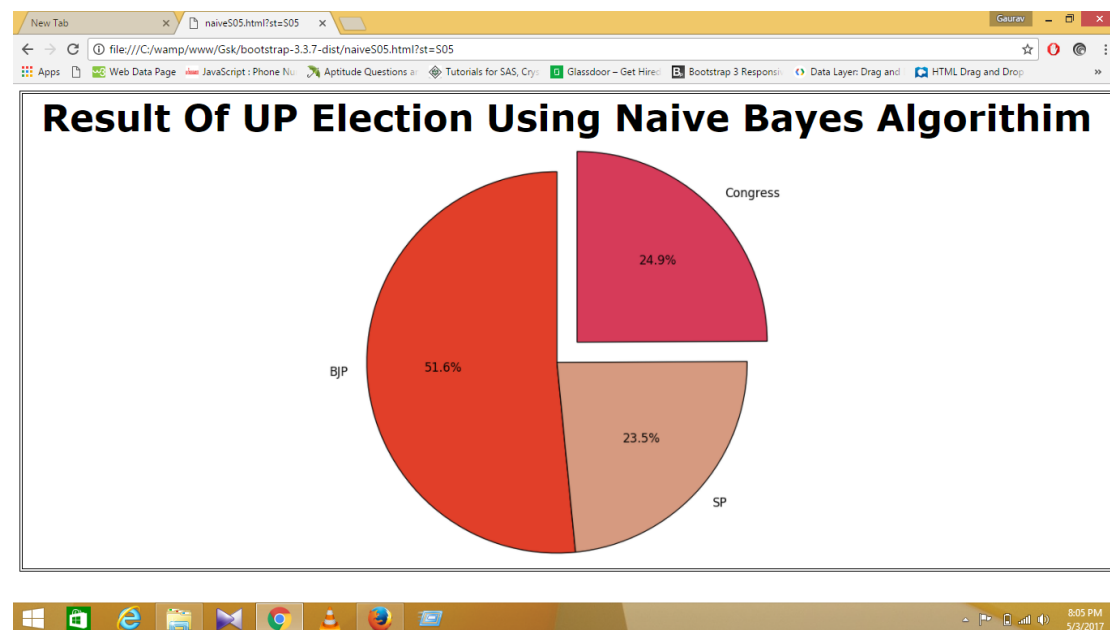
However, the equation does not use the neutral tweets as they don't express a candidate preference.

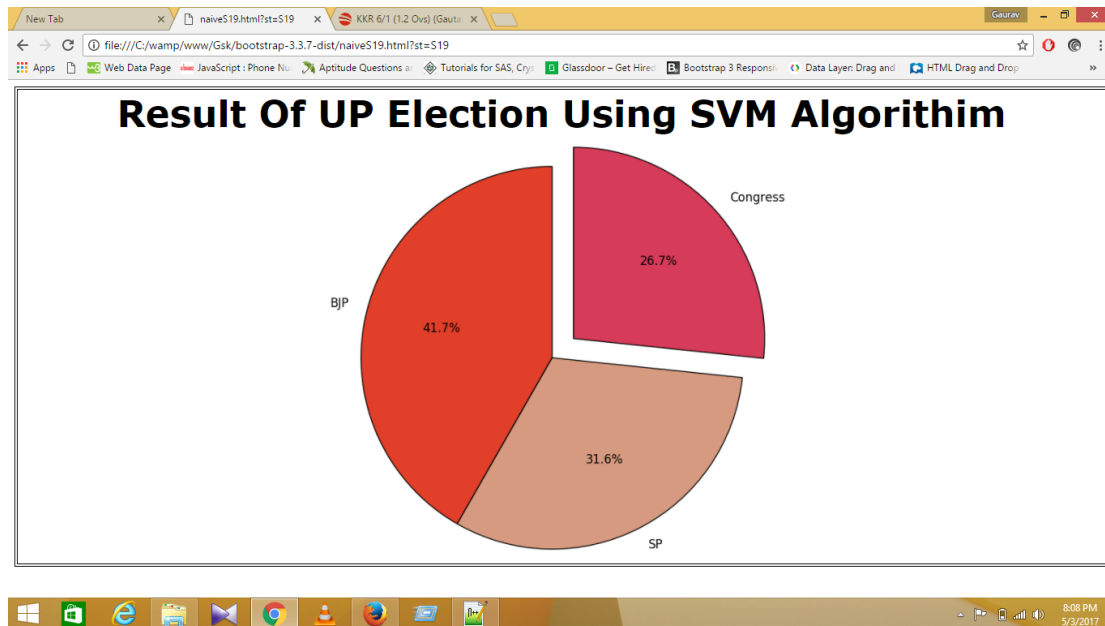
[Reference:-Kabir Ismail Umar Data mining for social media analysis using twitter.]

17. Result visualization:

For the visualization of final result, we used pandas library of python. In that we created data frame from the data obtained after prediction. We used functions for creation of pie chart.

Final pie charts are displayed on website.



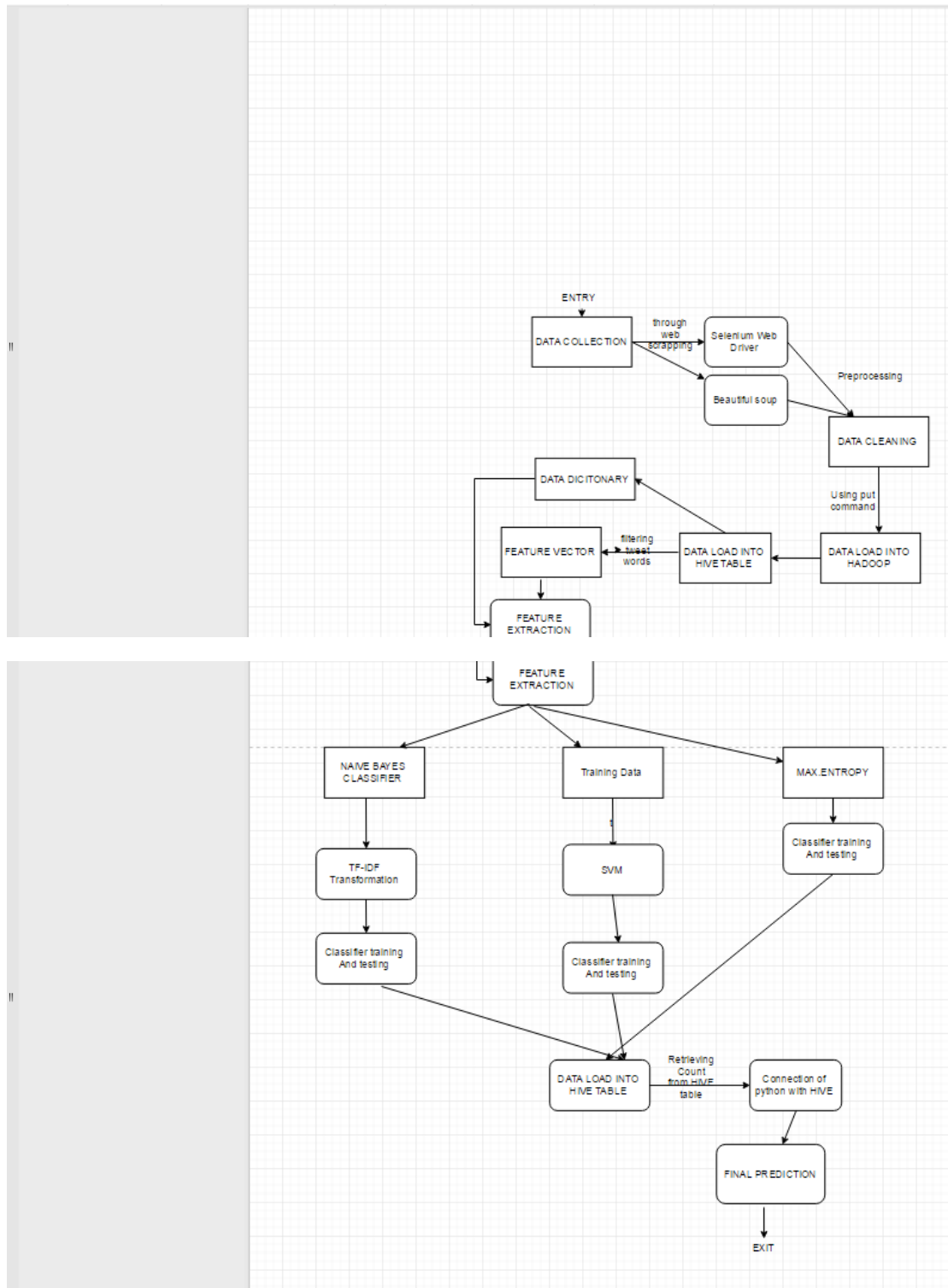


Website designing:-

An online portal to display our prediction in the form of pie charts . In addition to it, various constituencies and political parties list can be found on the website.

[Source:- election commission website.]

FLOWCHART



Language, database systems, packages used

We have used python for implementing all the algorithms and also for scraping twitter and data collection.

Database System used:-

We have stored the data in hive table. In addition to it party and sentiment for the tweets is stored in 3 different tables for different algorithms.

Packages Used:-

- Nltk for naive bayes and maximum entropy.
- Libsvm for svm.
- Pandas for visualization.
- Pyhive for connecting python script with hive table and performing queries.
- Selenium and BeautifulSoup for scraping of twitter data.
- In addition to it we have used anaconda framework for all the machine learning packages on windows as well as cloudera.

Testing

Test case	Input tweet	Naive Bayes Output	SVM Output	Maximum Entropy Output
1	“The only belief of BJP is Vikas Vikas Vikas people of UP believe that only BJP can make this possible..”	bjp positive	bjp positive	bjp positive
2	“UPElections 2017 Math suggest BJP set to win a comfortable majority Says Surjit Bhalla”	bjp positive	bjp positive	bjp positive
3	UPElections Strongly condemn road show. Urge the EC to take action against BJP for violation of model code of conduct.	bjp negative	bjp negative	bjp negative
4	UPElections So how did rape accused Prajapati get SP ticket writes Rajesh Singh	sp negative	sp negative	sp positive
5	Maharashtra civic poll result indicates clear majority of BJP4UP in UPElections thanks to people of Maha. for showing their faith BJP	bjp positive	bjp positive	bjp positive

FUTURE WORK

Machine learning techniques perform well for classifying sentiment in tweets. I believe the accuracy of the system could be still improved. Below is a list of ideas I think could help the classification:-

Semantics The algorithms classify the overall sentiment of a tweet. The polarity of a tweet may depend on the perspective you are interpreting the tweet from. For example, in the tweet “BJP beats SP :)”, the sentiment is positive for BJP and negative for SP. In this case, semantics may help. Using a semantic role label may indicate which noun is mainly associated with the verb and the classification would take place accordingly. This may allow “BJP beats SP :)” to be classified differently from “SP beats BJP :)”.

Bigger Dataset The training dataset in the order of millions will cover a better range of twitter words and hence better unigram feature vector resulting in an overall improved model. This would vastly improve upon the existing classifier results.

Internationalization Currently, we focus only on English tweets but Twitter has a huge international audience. It should be possible to use our approach to classify sentiment in other languages with a language specific positive/negative keyword list.

REFERENCES:-

[1] Kabir Ismail Umar & Fatima Chiroma .Data Mining for Social Media Analysis using Twitter to predict the 2016 US Presidential Election.1 September, 2016.

[2] Harald Schoen, University of Bamberg (Germany), Daniel Gayo-Avello, University of Oviedo (Spain), Panagiotis Takis Metaxas, Wellesley College and Harvard University.The Power of Prediction with Social Media.

[3] Donald E. Brown and Ahmed Abbasi, University of Virginia. Predictive analytics.

[4] Patrick Hummel (Google) ,David Rothschild (Microsoft Research) PHummel@alumni.gsb.stanford.edu. Fundamental Models for Forecasting Election

[5] David W. Nickerson University of Notre Dame ,Todd Rogers Harvard Kennedy School .Political Campaigns and Big Data Faculty Research Working Paper Series

[6] Lei Shi Opera Solutions,Neeraj Agarwal, Ankur Agrawal,Predicting US Primary Elections with Twitter.

[7] Ronald MacDonald¹ and Xuxin Mao² .18 January 2016. Forecasting the 2015 General Election with Internet Big Data: An Application of the TRUST Framework

[8] Rohan Sampath, Yue Teng.Classification and Regression Approaches to Predicting United States Senate Elections

[9] Data Mining in US presidential election campaign.
<https://www.promptcloud.com/blog/data-mining-in-presidential-election-campaign/>

[10] Robert Tibshirani. The lasso method for variable selection .In Statistics in Medicine, pages 385–395, 1997.

[11] D. O. Computer, C. wei Hsu, C. chung Chang, and C. jen Lin. A practical guide to support vector classification chih-wei hsu, chih-chung chang, and chih-jen lin. Technical report, 2003.

[12] N. Cristianini and J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge University Press, March 2000.

[13] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Micro-blogging as online word of mouth branding. In CHI EA '09: Proceedings of the 27th international conference extended abstracts on Human factors in computing systems, pages 3859–3864, New York, NY, USA, 2009. ACM.

[14] T. Joachims. Making large-scale support vector machine learning practical. In B. Scholkopf, C. J. C. Burges, and A. J. Smola, editors, Advances in kernel methods: support vector learning, pages 169–184. MIT Press, Cambridge, MA, USA, 1999.

[15] C. D. Manning and H. Schutze. Foundations of statistical natural language processing. MIT Press, 1999.

[16] G. Mishne. Experiments with mood classification in blog posts. In 1st Workshop on Stylistic Analysis Of Text For Information Access, 2005.

[17] K. Nigam, J. Lafferty, and A. McCallum. Using maximum entropy for text classification. In IJCAI-99 Workshop on Machine Learning for Information Filtering, pages 61–67, 1999.

[18] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[19] B. Pang, L. Lee, and S. Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 79–86, 2002.

