

Sentiment Analysis and Polarity Detection of Tweets through R Programming on Bengaluru Traffic, India

Annie Syrien¹, Dr.M.Hanumanthappa², Dr.S.Kavitha³

^{1,2}Department of Computer Applications, Bangalore University, Bangalore

³Department of Computer Applications, Dayananda Sagar College, Bangalore

Abstract:

This paper aims to retrieve data from twitter and pre-process the data into comprehensive particulars, and analyze the sentiments of tweets via providing scores to the tweets obtained on Bengaluru traffic with the help of opinion lexicon English, which has both negative and positive words, where these words are compared with collected data from the twitter. Simple naïve bayes algorithm is used to carry out experiment in r studio. Twitter users sentiments on the Bengaluru traffic is presented as the result through the R studio plots.

Keywords — Social media, R programming, Naïve bayes, sentiment analysis and polarity detection.

I. INTRODUCTION

Social media is an internet based communication tool that empowers people to share information. To understand better the term social media, social indicates associating with people and spending time in order to develop their relationships whereas media indicates tool for communication such as internet, TV, radio, newspaper etc. here our focus is internet. So social media can be stated as, an electronic platform for socializing people. Some example of social media sites are Facebook, Twitter, Youtube, LinkedIn, Digg etc. Initially people involved in social media to associate with their friends and lost friends, gradually they improved to the status of updating and consuming any information on social media, these led to vast generation of user data which could be further processed for future development [1].

In social media millions of people share their ideas, views, reviews, ratings on various events, topics, organizations, business products where these information can be further analyzed in order to extract the useful information where it can be used to draw the useful conclusions and suggestions for future implementations and developments. In the recent past opinion mining and sentiment analysis gained a vast scope in the field of web data mining. The fast growth of online user generated contents in

blogs, wikis and web forums lead to the era of processing to derive important information. Due to the huge data production from the online social networking users. Expansion of internet and extensive use of internet for information retrieval [2] has also contributed to the changes. The user contribution is the chief value driver in most of the Web 2.0 social networking applications [3].

Social media mining is acquiring useful insights from the user messages on the user media sites. Social media mining is the trending research topics where many researchers focus, due to the advancement of number of users in social media sites, the sentiments of the users are easily predicted via their likes, share, and clicks. Determining the sentiments such as positive and negative plays a vital role in determining the demand for a product, celebrity or reviews. The eWOM (electronic word of mouth) statements conveyed on the web are more prevalent in service industry and business to empower customer to share their point of view [4].

According to [5] Sentiment analysis is the branch of study involved in analyzing people's sentiments, opinions, appraisals, evaluations, attitudes and emotions towards services, products, organizations, issues, individuals, and topics. In other words sentiment analysis and opinion mining is an exercise of analyzing peoples' opinions or sentiments from the textual data [6]. On the other hand business wanted to know about the public

opinions and consumers wanted to know about the products and services of the existing users opinions [7]. 'What other people think' has been an major piece of information for most of the domains during the decision-making process [8]. Sentiment analysis extents the innumerous research challenges and promises useful insights to the interested individuals [9]. Sentiment analysis is a growing area of NLP [10]. According to Daiyan [11] the elevate in opinion mining research is due to the advancement of machine learning and NLP and information retrieval, realization of intellectual challenges. Sentiment analysis and opinion mining can be effectuated in three different levels: aspect level, sentence level and document level [12]. The approach in opinion mining has three main stages, such Opinion Retrieval, Opinion Classification and Opinion Summarization [13]. Some of the applications of sentiment analysis are finding out the reviews such as movies, products., and context sensitive information detection, spam detection, knowing consumer attitudes and trends [14].

PROBLEM STATEMENT

Twitter is one of the popular microblogging service where users generate their status messages known as tweets [15]. The microblogs messages are short, simple and interactive [16]. By default twitter messages are public and come in large size with high rate [17]. Social networks sites and blogs are valuable resources for sentiments mining [18]. Many firms are interested in twitter data for acquiring the people's opinion [19]. Data extraction is one of the challenges faced by the researchers and organizations despite the software tools availability. Web based sentiment analysis focus on single tweet analysis only posing restrictions on mass sentiment analysis [20]. Sentiment analysis is difficult with the use of informal language usages such as slangs and other shortcuts in the words [20]. Interrogative sentences, sarcastic, sentiment without sentiment words, conditional sentences and spam reviews also raised challenges regarding the sentiment analysis [21].

OBJECTIVE

The objective of this paper is to analyze the twitterer sentiments on bengaluru traffic through R programming language and plot the chart to represent the data comprehended. Data from the twitter is extracted through R Studio via Twitter app over the network, and then the data is processed and cleaned for the analysis. Finally sentiment analysis is performed on the obtained data.

RELATED WORK

Vibhor Singh et al [22] presented the research work in Opinion Mining and Analysis of Movie Reviews. They focused on sentiment analysis on texts via manually defined dictionary for negative, positive and intensifier words. Data was collected from three different sources such as twitter, third party website and their website. Data was stored in text file and compared via dictionary based algorithm with predefined semantic orientation and sentiments were obtained for a positive and negative for a particular movie and rating were generated based on trained data sets via three machine learning algorithms such as decision tree, k nearest neighbors, and naïve bayes. According to their research naïve bayes gives the best accurate results and tabulated as 54.10% and other methods such as decision tree 44.26% and k nearest neighbors as 50.28%. Accuracy calculated with the below measure.

Accuracy: Close to the standard value = $(TP+TN)/(TP+TN+FP+FN)$

Where as

TP stands for True positive weighted that is a rating which is predicted as positive which is really positive.

TN stands for True negative weighted that is a rating which is predicted as negative which is really negative.

FP stands for False positive weighted that is a rating which is predicted as positive which is really negative.

FN stands for False Negative weighted that is a rating which is predicted as negative which is really positive.

Sanket Patil et al [23] worked on Data preprocessing, Sentiment Analysis & NER on Twitter Data, in their paper they focused on data preprocessing such as stop word removal, special characters and slang words and processed data is analyzed through sentiment analysis model via Name entity recognition(NER). The name entity recognition separates the data in appropriate entity category, such as location, name, and organization. Finally the tweets are classified as positive, negative and neutral.

Peiman Barnaghi et al [24] contributed significant work on Opinion mining and sentiment polarity on twitter and correlation between events and sentiment, their focal point was to collect the data from twitter streaming API, preprocess the data and apply two types of methods to find the sentiments of the tweets. One method was manually label tagging as positive and negative, and extracting the features to unigram, bigram, term frequency – inverse document, then applying the sentiment polarity classifier through which building a trained model and obtaining the sentiments. The other method was tweets selection via keyword search then building a trained model and obtaining sentiments. The trained model was based on Bayesian logistic regression classification method. FIFA world cup 2014 was their topic of interest.

METHODOLOGY

The data from twitter is collected through the twitter APP online via statistical package R Studio which is effective, fast and powerful tool for data, and text processing. The collected data is preprocessed. Preprocessing of data involves detaching noise such as user mentions, URLs, punctuations, hashtags, controls and special characters [25]. Then simple naïve bayes algorithm is applied in order to compute the sentiment of the tweets with the help of opinion lexicon English obtained from github. Each tweet will be compared with the opinion lexicon English both positive and negative words, the positive words are assigned with +1, +2 and so on sign and negative sentiments are assigned

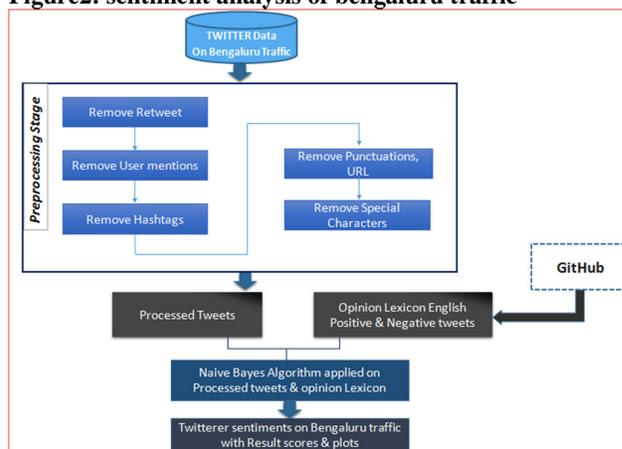
with -1,-2 and so on. The neutral words are assigned with 0 value. Neutral words are those tweets that does not imply anything on the bengaluru traffic sentiments but irrelevant.

The below figure 1 shows the illustration of positive, negative and neutral tweets

Figure1: positive, negative and neutral tweets

	text	score
6	Impossible has been done hrs late but I yam readddd...	-1
7	As draws to a close lets salute the Bangalore City Pol...	1
8	guess my last visit to Bangalore was yrs back not in I...	0
9	Bangalore is a bit ahead Traffic jam in air too <U+653...	-2
10	Yeah this happens all the time Though considering B...	-1
11	Fucked up traffic from Bangalore to MysoreIt might ta...	-1
12	Wah Bangalore Jugaad traffic jams and quiet nights A ...	0
13	She had a thought Which makes sense cause it will al...	1
14	Bangalore to Hosur traffic near checklist is well know...	0
15	I just mean thinking of Bangalore my head fills with th...	1
16	Ipl fever mg road full traffic my headblast in signals B...	-2
17	Here is the rule applicable across the country This w...	0

Figure2: sentiment analysis of bengaluru traffic



The figure 2 shows the flow of various steps involved in this research paper

NAÏVE BAYES

Naïve bayes is simple probabilistic classifier which has strong conditional independence assumption for optimal classifying classes with highly dependent features. Positive words, negative words and tweets from the twitter classes are calculated using probability based on the bayes theorem. In spite of this method being simple probabilistic classifier it has produced remarkable results.

The figure 3 shows the naïve bayes classifier in R studio

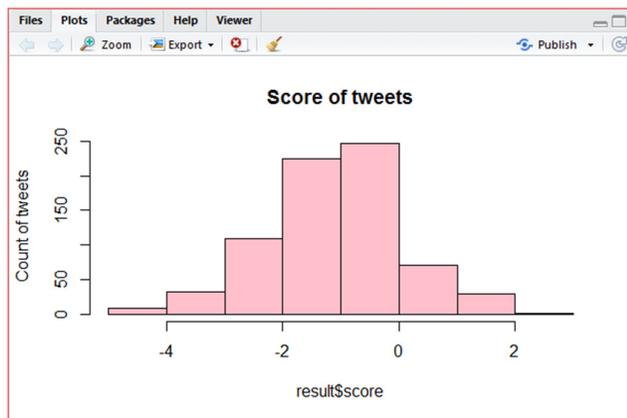
Figure 3: Naïve bayes classifier code

```

146 opinion.pos = scan('positive.txt', what = 'accolade', comment.char = ';')
147
148
149 opinion.neg = scan('negative.txt', what = 'character', comment.char = ';')
150
151 head(opinion.pos)
152 head(opinion.neg)
153 positvewords = c(opinion.pos,'good', 'clear', 'no traffic')
154 negativewords = c(opinion.neg,'traffic', 'waiting', 'bad', 'roads', 'time')
155
156
157 getsentimentscore = function(textPreprocessing, positvewords,
158                             negativewords, .progress='none')
159 {
160   require(plyr)
161   require(stringr)
162   scores = lapply(TextPreprocessing,
163                  function(TextPreprocessing, positvewords, negativewords) {
164                     # Let first remove the digit, punctuation character and contro
165                     TextPreprocessing = gsub('[[:cntrl:]]', '', gsub('[[:punct:]]',
166                     # Then lets convert all to lower sentence case:
167                     #sentence = tolower(sentence)
168                     # Now lets split each sentence by the space delimiter
169                     words = unlist(str_split(TextPreprocessing, "\\s+"))
170                     # get the boolean match of each words with the positive & negat
171                     pos.matches = !is.na(match(words, positvewords))
172                     neg.matches = !is.na(match(words, negativewords))
173                     # Now get the score as total positive sentiment minus the tota
174                     score = sum(pos.matches) - sum(neg.matches)
175                     return(score)
176                   }, positvewords, negativewords, .progress=.progress )
177   # Return a data frame with respective sentence and the score
178   return(data.frame(text=TextPreprocessing, score=scores))
179 }
180 }
181 }
182 }
183 }
184 }
185 }
186 }
187 }
188 }
189 }
190 }
191 }
192 }
193 }
194 }
195 }
196 }
197 }
198 }
199 }
200 }
201 }
202 }
203 }
204 }
205 }
206 }
207 }
208 }
209 }
210 }
211 }
212 }
213 }
214 }
215 }
216 }
217 }
218 }
219 }
220 }
221 }
222 }
223 }
224 }
225 }
226 }
227 }
228 }
229 }
230 }
231 }
232 }
233 }
234 }
235 }
236 }
237 }
238 }
239 }
240 }
241 }
242 }
243 }
244 }
245 }
246 }
247 }
248 }
249 }
250 }
251 }
252 }
253 }
254 }
255 }
256 }
257 }
258 }
259 }
260 }
261 }
262 }
263 }
264 }
265 }
266 }
267 }
268 }
269 }
270 }
271 }
272 }
273 }
274 }
275 }
276 }
277 }
278 }
279 }
280 }
281 }
282 }
283 }
284 }
285 }
286 }
287 }
288 }
289 }
290 }
291 }
292 }
293 }
294 }
295 }
296 }
297 }
298 }
299 }
300 }
301 }
302 }
303 }
304 }
305 }
306 }
307 }
308 }
309 }
310 }
311 }
312 }
313 }
314 }
315 }
316 }
317 }
318 }
319 }
320 }
321 }
322 }
323 }
324 }
325 }
326 }
327 }
328 }
329 }
330 }
331 }
332 }
333 }
334 }
335 }
336 }
337 }
338 }
339 }
340 }
341 }
342 }
343 }
344 }
345 }
346 }
347 }
348 }
349 }
350 }
351 }
352 }
353 }
354 }
355 }
356 }
357 }
358 }
359 }
360 }
361 }
362 }
363 }
364 }
365 }
366 }
367 }
368 }
369 }
370 }
371 }
372 }
373 }
374 }
375 }
376 }
377 }
378 }
379 }
380 }
381 }
382 }
383 }
384 }
385 }
386 }
387 }
388 }
389 }
390 }
391 }
392 }
393 }
394 }
395 }
396 }
397 }
398 }
399 }
400 }
401 }
402 }
403 }
404 }
405 }
406 }
407 }
408 }
409 }
410 }
411 }
412 }
413 }
414 }
415 }
416 }
417 }
418 }
419 }
420 }
421 }
422 }
423 }
424 }
425 }
426 }
427 }
428 }
429 }
430 }
431 }
432 }
433 }
434 }
435 }
436 }
437 }
438 }
439 }
440 }
441 }
442 }
443 }
444 }
445 }
446 }
447 }
448 }
449 }
450 }
451 }
452 }
453 }
454 }
455 }
456 }
457 }
458 }
459 }
460 }
461 }
462 }
463 }
464 }
465 }
466 }
467 }
468 }
469 }
470 }
471 }
472 }
473 }
474 }
475 }
476 }
477 }
478 }
479 }
480 }
481 }
482 }
483 }
484 }
485 }
486 }
487 }
488 }
489 }
490 }
491 }
492 }
493 }
494 }
495 }
496 }
497 }
498 }
499 }
500 }
501 }
502 }
503 }
504 }
505 }
506 }
507 }
508 }
509 }
510 }
511 }
512 }
513 }
514 }
515 }
516 }
517 }
518 }
519 }
520 }
521 }
522 }
523 }
524 }
525 }
526 }
527 }
528 }
529 }
530 }
531 }
532 }
533 }
534 }
535 }
536 }
537 }
538 }
539 }
540 }
541 }
542 }
543 }
544 }
545 }
546 }
547 }
548 }
549 }
550 }
551 }
552 }
553 }
554 }
555 }
556 }
557 }
558 }
559 }
560 }
561 }
562 }
563 }
564 }
565 }
566 }
567 }
568 }
569 }
570 }
571 }
572 }
573 }
574 }
575 }
576 }
577 }
578 }
579 }
580 }
581 }
582 }
583 }
584 }
585 }
586 }
587 }
588 }
589 }
590 }
591 }
592 }
593 }
594 }
595 }
596 }
597 }
598 }
599 }
600 }
601 }
602 }
603 }
604 }
605 }
606 }
607 }
608 }
609 }
610 }
611 }
612 }
613 }
614 }
615 }
616 }
617 }
618 }
619 }
620 }
621 }
622 }
623 }
624 }
625 }
626 }
627 }
628 }
629 }
630 }
631 }
632 }
633 }
634 }
635 }
636 }
637 }
638 }
639 }
640 }
641 }
642 }
643 }
644 }
645 }
646 }
647 }
648 }
649 }
650 }
651 }
652 }
653 }
654 }
655 }
656 }
657 }
658 }
659 }
660 }
661 }
662 }
663 }
664 }
665 }
666 }
667 }
668 }
669 }
670 }
671 }
672 }
673 }
674 }
675 }
676 }
677 }
678 }
679 }
680 }
681 }
682 }
683 }
684 }
685 }
686 }
687 }
688 }
689 }
690 }
691 }
692 }
693 }
694 }
695 }
696 }
697 }
698 }
699 }
700 }
701 }
702 }
703 }
704 }
705 }
706 }
707 }
708 }
709 }
710 }
711 }
712 }
713 }
714 }
715 }
716 }
717 }
718 }
719 }
720 }
721 }
722 }
723 }
724 }
725 }
726 }
727 }
728 }
729 }
730 }
731 }
732 }
733 }
734 }
735 }
736 }
737 }
738 }
739 }
740 }
741 }
742 }
743 }
744 }
745 }
746 }
747 }
748 }
749 }
750 }
751 }
752 }
753 }
754 }
755 }
756 }
757 }
758 }
759 }
760 }
761 }
762 }
763 }
764 }
765 }
766 }
767 }
768 }
769 }
770 }
771 }
772 }
773 }
774 }
775 }
776 }
777 }
778 }
779 }
780 }
781 }
782 }
783 }
784 }
785 }
786 }
787 }
788 }
789 }
790 }
791 }
792 }
793 }
794 }
795 }
796 }
797 }
798 }
799 }
800 }
801 }
802 }
803 }
804 }
805 }
806 }
807 }
808 }
809 }
810 }
811 }
812 }
813 }
814 }
815 }
816 }
817 }
818 }
819 }
820 }
821 }
822 }
823 }
824 }
825 }
826 }
827 }
828 }
829 }
830 }
831 }
832 }
833 }
834 }
835 }
836 }
837 }
838 }
839 }
840 }
841 }
842 }
843 }
844 }
845 }
846 }
847 }
848 }
849 }
850 }
851 }
852 }
853 }
854 }
855 }
856 }
857 }
858 }
859 }
860 }
861 }
862 }
863 }
864 }
865 }
866 }
867 }
868 }
869 }
870 }
871 }
872 }
873 }
874 }
875 }
876 }
877 }
878 }
879 }
880 }
881 }
882 }
883 }
884 }
885 }
886 }
887 }
888 }
889 }
890 }
891 }
892 }
893 }
894 }
895 }
896 }
897 }
898 }
899 }
900 }
901 }
902 }
903 }
904 }
905 }
906 }
907 }
908 }
909 }
910 }
911 }
912 }
913 }
914 }
915 }
916 }
917 }
918 }
919 }
920 }
921 }
922 }
923 }
924 }
925 }
926 }
927 }
928 }
929 }
930 }
931 }
932 }
933 }
934 }
935 }
936 }
937 }
938 }
939 }
940 }
941 }
942 }
943 }
944 }
945 }
946 }
947 }
948 }
949 }
950 }
951 }
952 }
953 }
954 }
955 }
956 }
957 }
958 }
959 }
960 }
961 }
962 }
963 }
964 }
965 }
966 }
967 }
968 }
969 }
970 }
971 }
972 }
973 }
974 }
975 }
976 }
977 }
978 }
979 }
980 }
981 }
982 }
983 }
984 }
985 }
986 }
987 }
988 }
989 }
990 }
991 }
992 }
993 }
994 }
995 }
996 }
997 }
998 }
999 }
1000 }

```

Figure 4: Scores of the tweets



CONCLUSION AND RECOMMENDATION

R Studio is open source and free integrated development environment. The article provides the sentiments of twitterer through simple naïve bayes algorithm in R programming on bengaluru city traffic. In the figure 4 it is evident that many of the twitterers are unhappy about the bengaluru traffic as the scores in the negative values are more. The zero score are indicates they are normal tweets without expressing any sentiments. It is found that R for twitter data analysis serves as effective tool. The future scope for this study would be comparing the different machine learning algorithms for the twitter data analysis.

REFERENCES

- [1] M. Hanumanthappa, S Regina Lourdhu Suganthi Annie Syrien, "Survey on Mining Opinions and Sentiments in Social Media," *International Journal of Engineering Research & Technology (IJERT)*, vol. 4, no. 27, p. 39, 2016.
- [2] Yazan Alaya AL-Khassawneh, Naomie Salim, and Mutasem Jarrah, "Improving Triangle-Graph Based Text Summarization using Hybrid Similarity Function," *Indian Journal of Science and Technology*, vol. 10, 2017.
- [3] K. C. Ravi Kumar, D. Teja Santosh, and B. Vishnu Vardhan, "Extracting Opinion Targets from Product Reviews using Comprehensive Feature Extraction Model in Opinion Mining," *Indian Journal of Science and Technology*, vol. 8, 2017.
- [4] Kumar Ravi and Vadlamani Ravi, "A survey on opinion mining and sentiment analysis: tasks, approaches and applications," *Knowledge-Based Systems*, vol. 89, pp. 14-46, 2015.
- [5] Bing Liu, "Sentiment analysis and opinion mining," *Synthesis lectures on human language technologies*, vol. 5, pp. 1-167, 2012.
- [6] Kai Yang et al., "An effective hybrid model for opinion mining and sentiment analysis," in *Big Data and Smart Computing (BigComp), 2017 IEEE International Conference on*, 2017, pp. 465-466.
- [7] Bing Liu and Lei Zhang, "A survey of opinion mining and sentiment analysis," in *Mining text data.*: Springer, 2012, pp. 415-463.
- [8] Bo Pang, Lillian Lee, and others, "Opinion mining and

- sentiment analysis," *Foundations and Trends*
- [9] Bing Liu, *Sentiment analysis: Mining opinions, sentiments, and emotions.*: Cambridge University Press, 2015.
- [10] Efthymios Kouloumpis, Theresa Wilson, and Johanna D. Moore, "Twitter sentiment analysis: The good the bad and the omg!," *Icwsn*, vol. 11, p. 164, 2011.
- [11] Md Daiyan, Dr S. K. Tiwari, Manish Kumar, and M. Aftab Alam, "A Literature Review on Opinion Mining and Sentiment Analysis," *International Journal of Emerging Technology and Advanced Engineering*, vol. 5, 2015.
- [12] Imene Guellil and Kamel Boukhalfa, "Social big data mining: a survey focused on opinion mining and sentiments analysis," in *Programming and Systems (ISPS), 2015 12th International Symposium on*, 2015, pp. 1-10.
- [13] K. L. Santhosh Kumar, Jayanti Desai, and Jharna Majumdar, "Opinion mining and sentiment analysis on online customer review," in *Computational Intelligence and Computing Research (ICCIC), 2016 IEEE International Conference on*, 2016, pp. 1-4.
- [14] Subhabrata Mukherjee and Pushpak Bhattacharyya, "Sentiment analysis: A literature survey," *arXiv preprint arXiv:1304.4520*, 2013.
- [15] Alec Go, Lei Huang, and Richa Bhayani, "Twitter sentiment analysis," *Entropy*, vol. 17, p. 252, 2009.
- [16] Lu Lin, Jianxin Li, Richong Zhang, Weiren Yu, and Chenggen Sun, "Opinion Mining and Sentiment Analysis in Social Networks: A Retweeting Structure-Aware Approach," in *Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing*, 2014, pp. 890-895.
- [17] Sakkachin Wongcharoen and Twittie Senivongse, "Twitter analysis of road traffic congestion severity estimation," in *Computer Science and Software Engineering (JCSSE), 2016 13th International Joint Conference on*, 2016, pp. 1-6.
- [18] P. Ambika and M. R. Binu Rajan, "Survey on diverse facets and research issues in social media mining," in *Research Advances in Integrated Navigation Systems (RAINS), International Conference on*, 2016, pp. 1-6.
- [19] Mondher Bouazizi and Tomoaki Ohtsuki, "Opinion mining in Twitter: How to make use of sarcasm to enhance sentiment analysis," in *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*, 2015, pp. 1594-1597.
- [20] Chayanit Nadam, Shuib Basri Aliza Sarlan, "Twitter Sentiment Analysis," *IEEE*, p. 212, November 2014.
- [21] Vidisha M. Pradhan, Jay Vala, and Prem Balani, "A survey on Sentiment Analysis Algorithms for opinion mining," *International Journal of Computer Applications*, vol. 133, 2016.
- [22] Vibhor Singh, Priyansh Saxena, Siddharth Singh, and S. Rajendran, "Opinion Mining and Analysis of Movie Reviews," *Indian Journal of Science and Technology*, vol. 10, 2017.
- [23] Varsha Wangikar, K. Jayamalini Sanket Patil, "Data Preprocessing, Sentiment Analysis & NER On Twitter Data," *IOSR Journal of Computer Engineering*, pp. 73-79, 2017.
- [24] Peiman Barnaghi, Parsa Ghaffari, and John G. Breslin, "Opinion mining and sentiment polarity on twitter and correlation between events and sentiment," in *Big Data Computing Service and Applications (BigDataService), 2016 IEEE Second International Conference on*, 2016, pp. 52-57.
- [25] M. Hanumanthappa, Annie Syrien, and N. M. Jyothi, "Twitter Data Retrieval and Preprocessing through R Programming in Social Media: An Overview of Bengaluru Traffic, India," *Indian Journal of Science and Technology*, vol. 10, 2017.