

Sentimental Analysis of Reddit Social Media Website

Aarti, Prabhdeep kaur, Geetika Chatley

School of Computer Science and Engineering

Lovely Professional University, Phagwara

aarti.1208@gmail.com, prabhdeep.25127@lpu.co.in, geetika.25107@lpu.co.in

Abstract: Reddit is social news aggregation and discussion website with large communities of people across different categories. This paper focus on mining/extracting data efficiently from popular subreddits and analyzing the dataset. Analysis of sentiment is the statistical analysis of the opinions, feelings, emotions and attitude of people expressed in written language. It is one of the most active research areas in the processing of natural languages and text mining in recent years. The value of sentiment analysis correlates with social media growth such as posts, discussion forums, articles, micro-blogs, Twitter, and online platforms. In almost every company and social field, sentiment analysis systems are implemented because feelings are fundamental to nearly every human activity and are essential influencers of our behaviours. In this research, reddit website is analysed using R and results are evaluated using different parameters.

Keywords: Reddit, Opinion mining, text mining, social-field

1. Introduction:

Reddit is a media, social networking and news website where registered community members may upload material, such as text messages or links to websites. Material entries are grouped into topics of interest named subreddits. Members can vote for the submissions in the form of up or down to decide their place on the site's page based on the net positive votes that the entry received called Karma. In fact, users may post on these submissions and vote on these posts to decide their place on Karma's website page as well[1][2]. Over the years, the website has become massively popular, attracting throughout 7.5 billion overall views of the page and 170 million unique visitors each month. In fact, Reddit users vote on the forum more than 24 million times a day.

Every article has a comment thread connected with it, and Reddit users vote for each message, creating a net score, or "Karma". The analysis of people sound optimistic, negative or neutral. A feeling is often expressed in a text in simple or complex ways. The online users can use a number of different available methods to express their emotions. In addition, they may mix subjective and objective information on a particular topic. In addition, data collected from the internet often contain a lot of noise. Indeed, for all the aforementioned reasons, the task of feeling recognition (automatic) in the online text becomes more difficult[3][4].

2. SOFTWARE REQUIREMENT ANALYSIS

2.1 GENERAL DESCRIPTION

Social Media platforms have become essential for user-generated content and consumer opinions. The Reddit API and software were accessed so that posts from Reddit can be scrapped quickly. To expand on the recent reviews have been done by using framework tools like python, R, PRAW, matplotlib and NLTK. People are

sharing their experiences, opinions, or just talking online about anything that concerns them. A vast amount of data available allows designers of the program to research automated mining and evaluation[5-10].

3. DESIGN

3.1 SYSTEM DESIGN

Sentiment analysis of online reviews has become a popular field of research. Sentiment analysis experiments focus primarily on frame and lexicon design, abstraction of characteristics, and determination of polarity[15-18].

This study provides an aggregate overview of the three main sentiment analysis research fields: context, function extraction and sentiment analysis, summarizing and reviewing the latest growth, and offering a thorough guide to its implementation in industry and blogs[1-4]. Despite the current immaturity of related research, online review sentiment analysis has taken its place as an innovative frontline study that takes advantage in many fields, such as text mining, natural language processing, internet mining, and machine learning. Many problems still need to be explored further and solved as follows:

- (1) Insufficient empirical language data and platform. So far, no experimental platform has been released for public use. For lack of an open experiment platform and standardized benchmark, it is difficult to assess the effectiveness of all the methods.
- (2) No advancement in the study of literary emotions. Study of deep sentiment ultimately requires linguistic decoding, as well as translating sentiment into documents, so there is no advancement in studying deep-structure semantic sentiment and analyzing textual sentiment.
- (3) The commercial value of online product reviews is not investigated. Online reviews have a profound impact on consumer buying behaviour.

3.2 Scrapping Of Data

Data scraping or data extraction is a method used to retrieve large amounts of data from websites by downloading data and transferring it to a specific directory on one's machine or to a table (spreadsheet) type server or excel template.

It includes the use of complex arithmetical algorithms. Screen/web scrapping is a method for extracting word- based characters from the web so that they could be analysed. Commonly, it is used to extract characters or paragraphs of interest from the websites (web scraping), though not exclusively.

The first step in our research is to build a simple crawler that scrapes data from Reddit. This will be done using PRAW in Python. To scrape Reddit, we will need to obtain a valid API key. The key was first generated on the Reddit profile which gives the access to the crawler

bot to crawl the Reddit page legally while binding to all the rules and regulations of the site in order that the sites working is not affected.

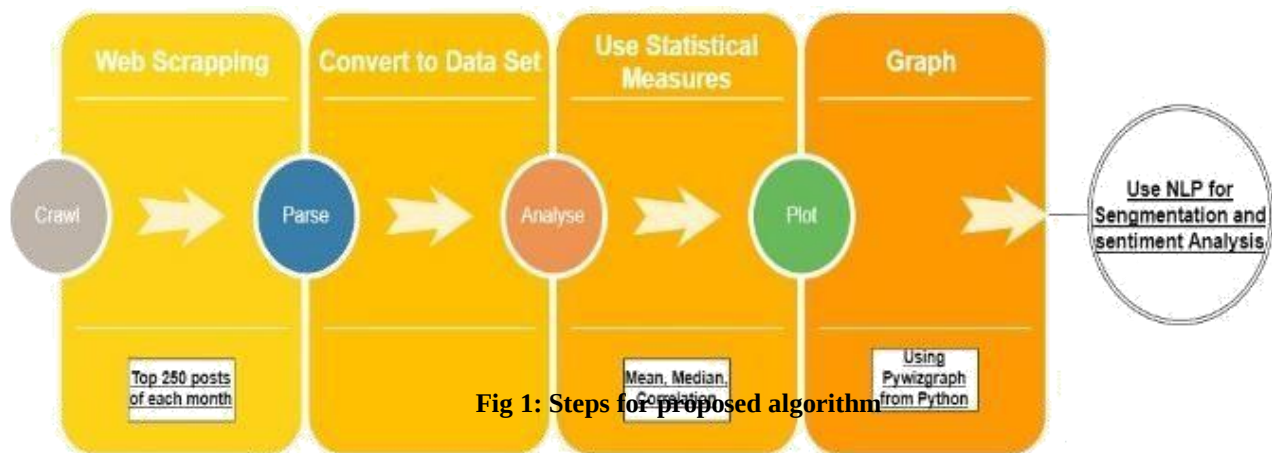


Fig 1: Steps for proposed algorithm

Data Scraping and Web Mining of Reddit Using the algorithm Proposed below:

Step 1: Create a reddit account.

Step 2: Navigate to preferences → apps.

Step 3: Get the client id and client secret keys by filling in the details.

Step 4: Create an init file with the following contents –

Reddit username, password, client id, client secret

Step 5: Create an empty file called commented.txt

Step 6: Authenticate –

Step 6.1: Read data from init file.

Step 6.2: Build json object and pass data to reddit.

Step 6.3: Authenticate the bot and store the returned data in an object.

Step 7: Fetch data –

Step 7.1: Get new comments from the subreddit 'India'.

Step 7.2: If comment text matches a regex of our interested topic,

Step 7.2.1: Save comment id.

Step 7.2.2: Add id to commented text file via file handling.

Step 7.2.3: Save comment body.

Step 7.2.4: Else, go to next comment.

Step 8: Set schedule –

Step 8.1: To work around Reddit API limits, set scraping schedule to few minutes.

Step 8.2: Call fetch data

Step 8.3: Wait for few seconds.

Step 8.4: Repeat

Step 9: Run analysis

Step 9.1: Get total comments parsed.

Step 9.2: Get total comments that are relevant.

Step 9.3: Get percent of relevant comments.

Step 9.4: Run analysis on the data collected.

Step 10: Stop

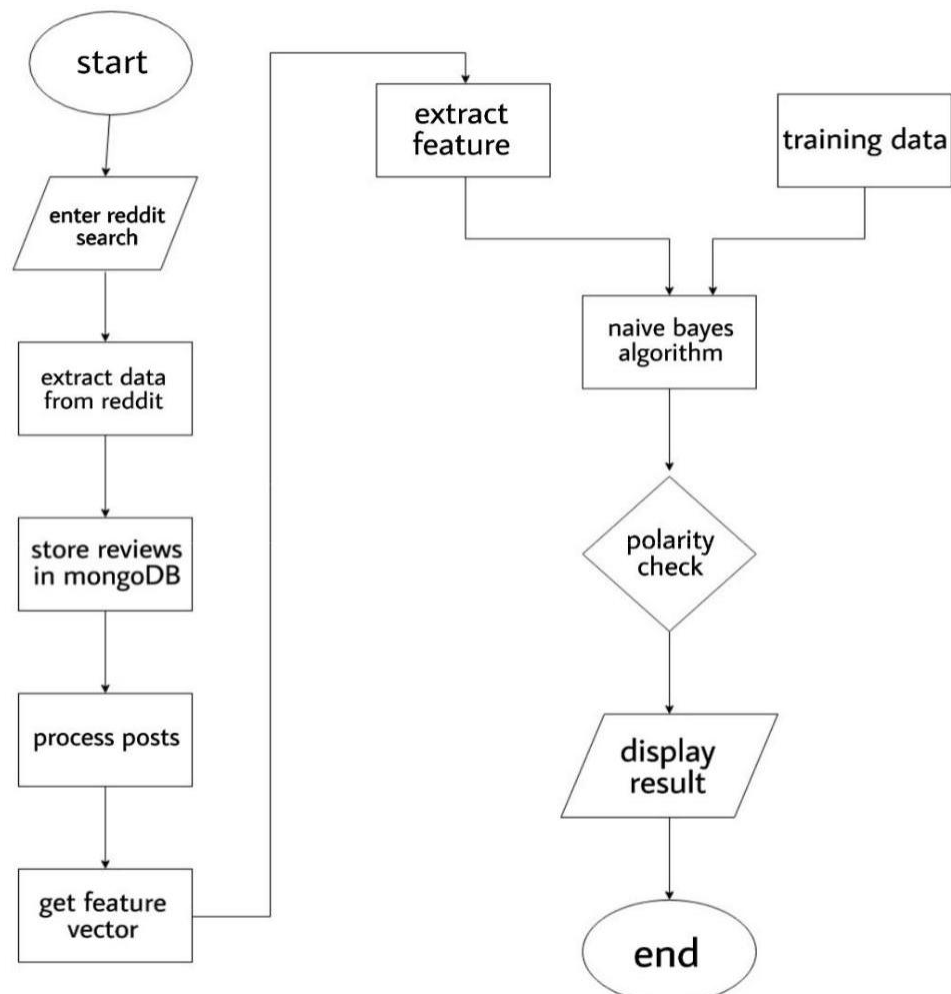


Fig 2: Flow chart for proposed algorithm

4. IMPLEMENTATION

4.1 Implementation of the project for performing sentimental analysis

-

Step 1: Generating the Key for the PRAW by means of an existing account on Reddit as a developer.

Step 2: Create an .ini file having the format

Step 3: Run the Python Code to fetch the data from the past year (scrap data) and to calculate and plot the data for further analysis

Step 4: The data set taken can be used to find the sentiments of users generally as well as for specific topics. It can also be used to find frequency distribution of words appearing in the corpus.

Step 5: Plot the values to have a graphical overview of the analysis that is just the visualization part of the process.

Step 6: Use tools like NLTK to find the positive response, negative or the neutral responses of the users and that analysis in turn depict the nature of the topic.

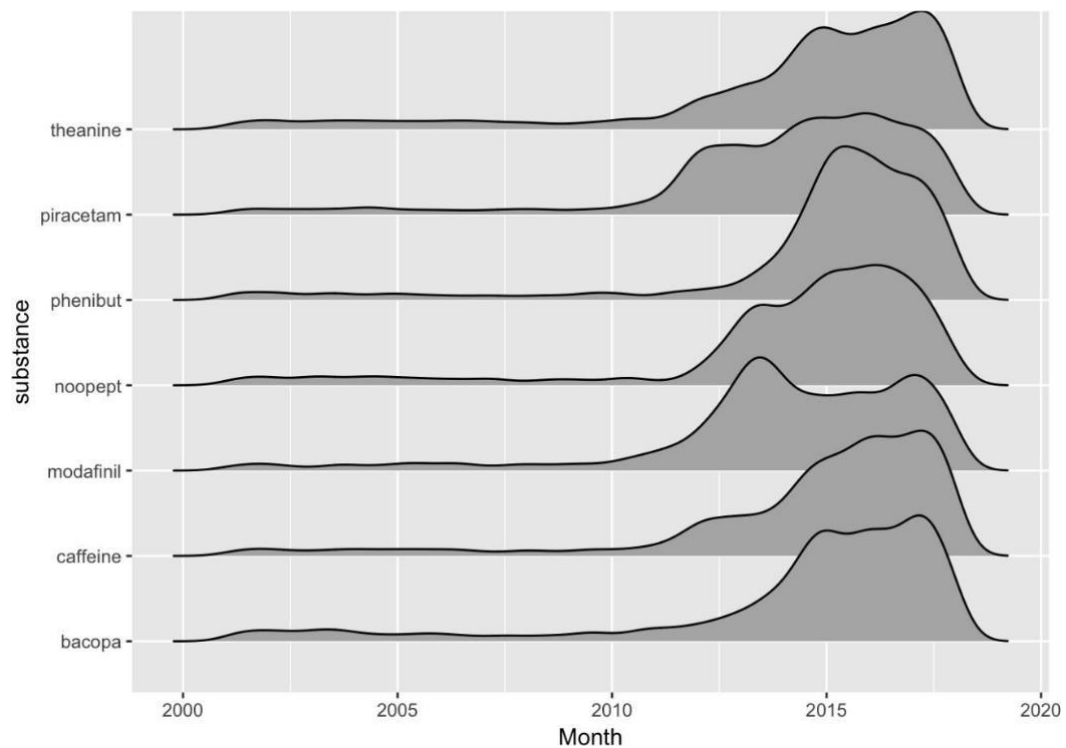
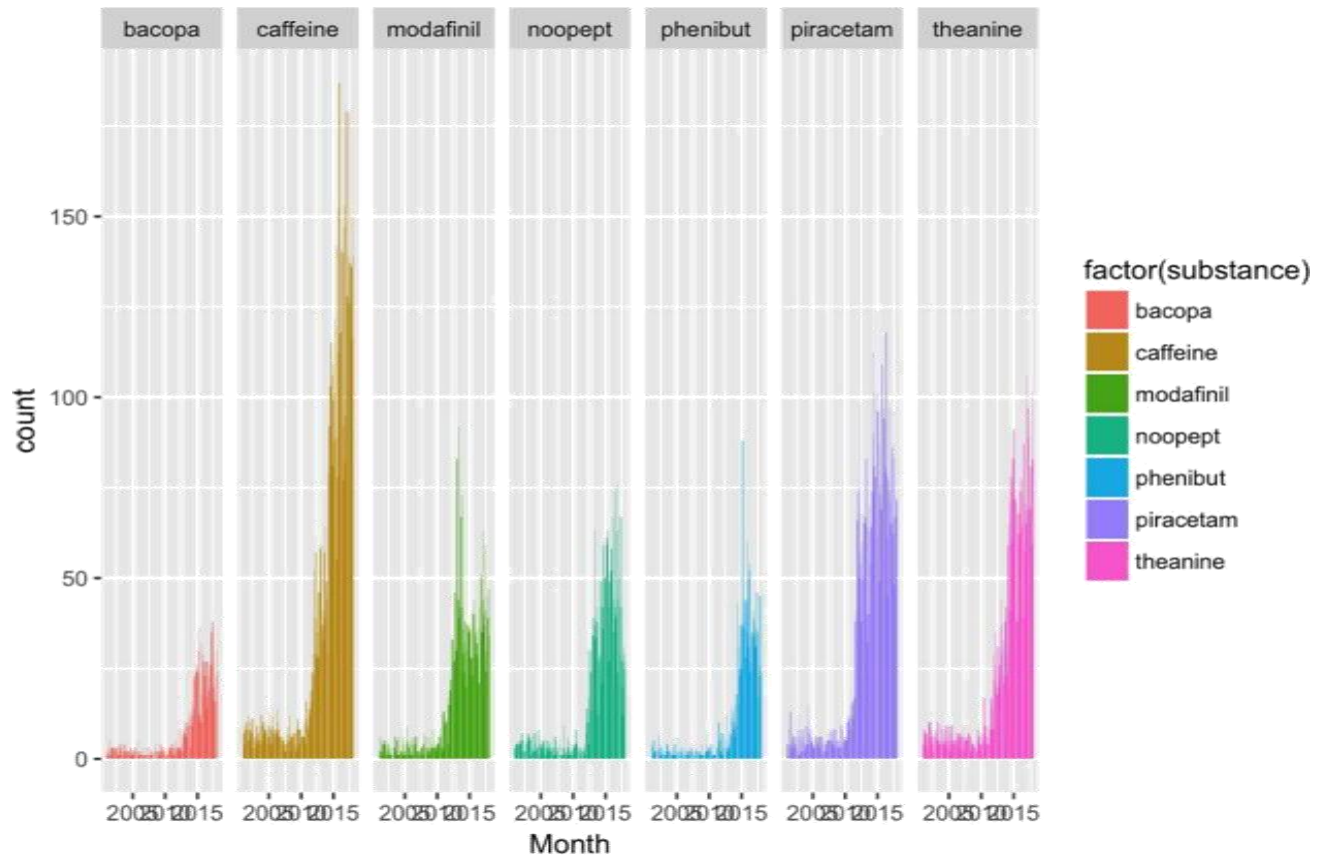
4.2 Subreddit Visualization using R

- Extract dataset from reddit using Google Big-Query and the user information to form network graphs. The graph includes the source, destination and the weight.
- The network graph was plotted using

5. POST-IMPLEMENTATION AND SOFTWARE MAINTENANCE

Post implementations steps for the project analysis includes the conclusion about reddit discussion:

1. Scraped using the Bot algorithm designed along with PRAW and Python.
2. Convert the fetched or parsed data into a dataset for ease of analysing.
3. Analyze the data set using software as NLTK used in Natural Language Processing (NLP) and create plots according to the result outputs.
4. Draw conclusion with respect to the analysis about the topic of the Reddit discussion among the users and the community.



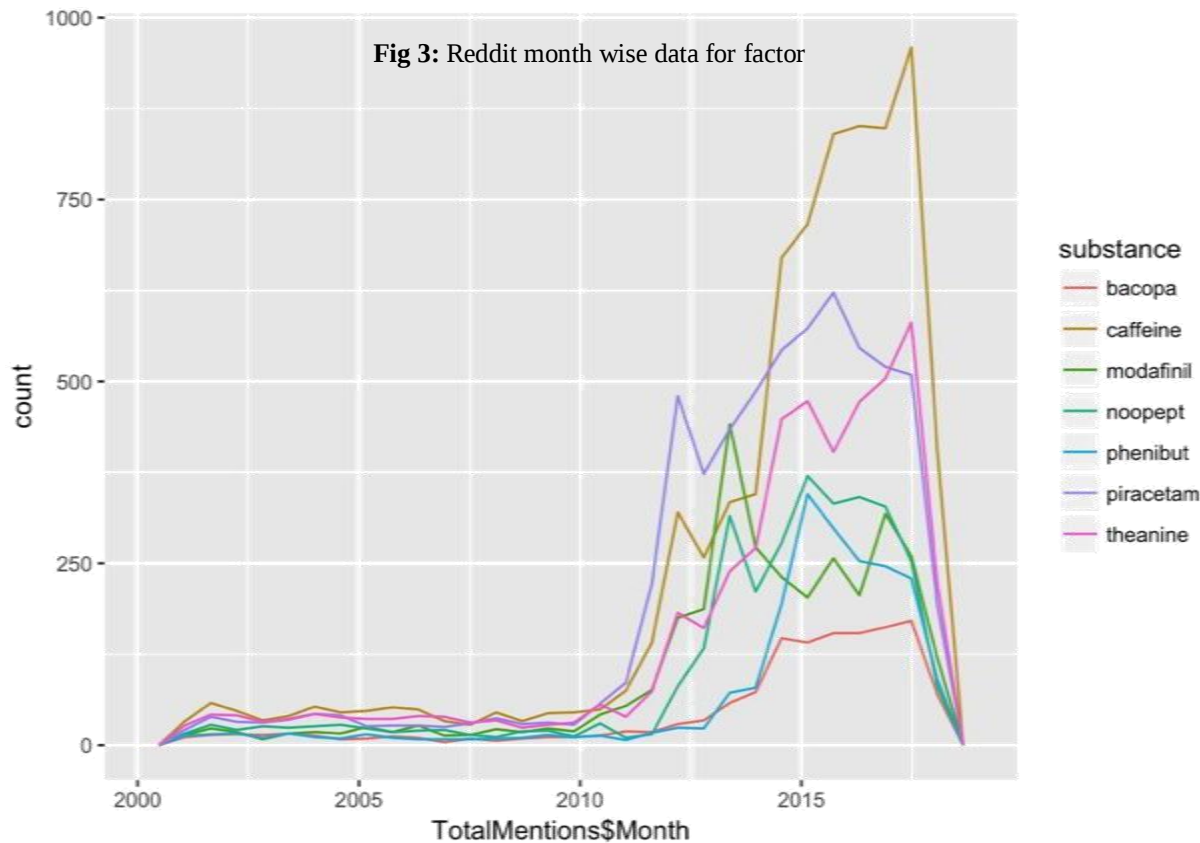


Fig 4: Reddit month wise graphical report for different parameters

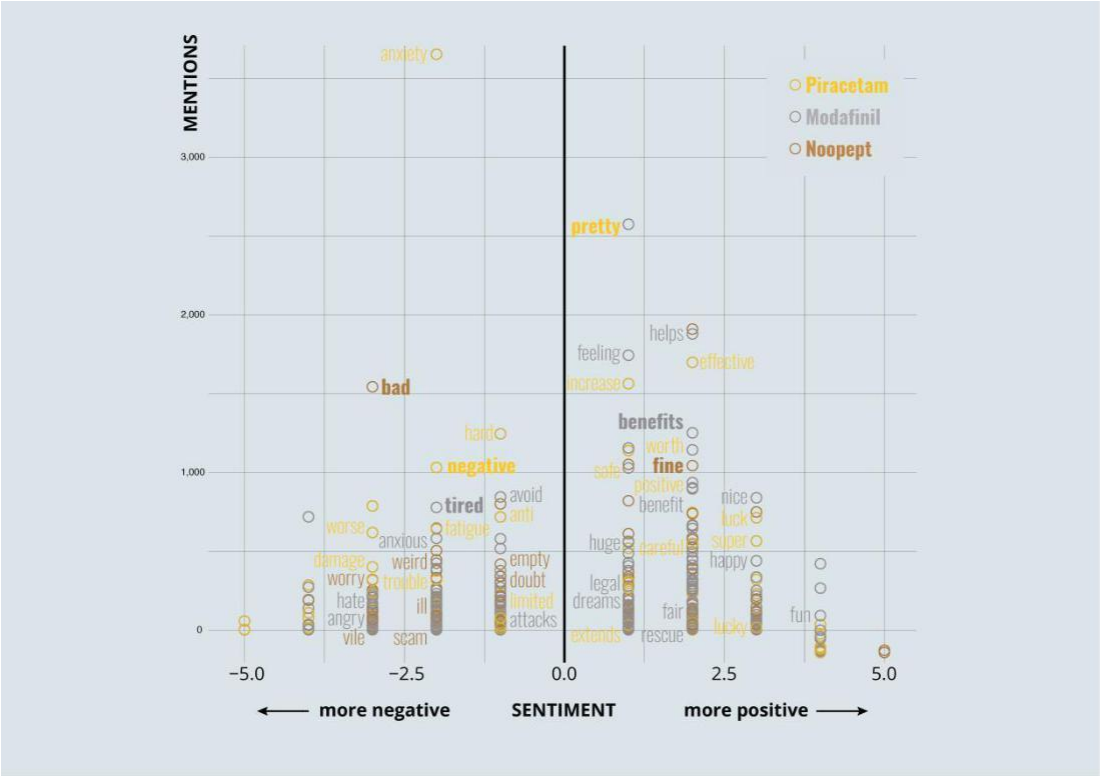


Fig 5: Line diagram for different parameters

Conclusion

This research focuses on the sentimental analysis of the reddit website and we may extend this research to different websites and different approaches.

References

- [1] L. Dey and S. M. Haque, "Opinion mining from noisy text data," *Int. J. Doc. Anal. Recognit.*, vol. 12, no. 3, pp. 205–226, 2009.
- [2] H. S. Behera and D. P. Mohapatra, "Computational intelligence in data mining—volume 1: Proceedings of the international conference on CIDM, 5-6 december 2015," *Adv. Intell. Syst. Comput.*, vol. 410, pp. 171–179, 2016.
- [3] "55.pdf."
- [4] S. Ahmed and A. Danti, "A novel approach for Sentimental Analysis and Opinion Mining based on SentiWordNet using web data," *Int. Conf. Trends Autom. Commun. Comput. Technol. I-TACT 2015*, pp. 0–4, 2016.
- [5] R. Dehkharghani, H. Mercan, A. Javeed, and Y. Saygin, "Sentimental causal rule discovery from Twitter," *Expert Syst. Appl.*, vol. 41, no. 10, pp. 4950–4958, 2014.
- [6] "5.pdf."
- [7] C. Lin, Y. He, R. Everson, and S. Rüger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Trans. Knowl. Data Eng.*, vol. 24, no. 6, pp. 1134–1145, 2012.
- [8] Fauziah, E. P. Wibowo, S. Madenda, and Hustinawati, "Identification of hand motion using background subtraction method and extraction of image binary with backpropagation neural network on skeleton model," *J. Phys. Conf. Ser.*, vol. 978, no. 1, 2018.
- [9] L. Gatti, M. Guerini, and M. Turchi, "SentiWords: Deriving a High Precision and High Coverage Lexicon for Sentiment Analysis," *IEEE Trans. Affect. Comput.*, vol. 7, no. 4, pp. 409–421, 2016.
- [10] A. Tripathy, A. Anand, and S. K. Rath, "Document-level sentiment classification using hybrid machine learning approach," *Knowl. Inf. Syst.*, vol. 53, no. 3, pp. 805–831, 2017.
- [11] U. Krcadinac, P. Pasquier, J. Jovanovic, and V. Devedzic, "Synesketech: An Open Source Library for Sentence-Based Emotion Recognition," *IEEE Trans. Affect. Comput.*, vol. 4, no. 3, pp. 312–325, 2013.
- [12] D. Gurkhe, N. Pal, and R. Bhatia, "Effective Sentiment Analysis of Social Media Datasets using Naive Bayesian Classification," *Int. J. Comput. Appl.*, vol. 99, no. 13, pp. 1–4, 2014.
- [13] D. K. Tayal and S. K. Yadav, "Fast retrieval approach of sentimental analysis with implementation of bloom filter on Hadoop," *2016 Int. Conf. Comput. Tech. Inf. Commun. Technol. ICCTICT 2016 - Proc.*, pp. 14–18, 2016.
- [14] R. M. Duwairi and M. Alfaqeeh, "RUM Extractor: A Facebook Extractor for Data Analysis," *Proc. - 2015 Int. Conf. Futur. Internet Things Cloud, FiCloud 2015 2015 Int. Conf. Open Big Data, OBD 2015*, pp. 709–713, 2015.
- [15] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Found. Trends Inf. Retr.*, vol. 2, no. 1–2, pp. 1–135, 2008.
- [16] M. Meire, M. Ballings, and D. Van den Poel, "The added value of auxiliary data in sentiment analysis of Facebook posts," *Decis. Support Syst.*, vol. 89, pp. 98–112, 2016.
- [17] A. Chandra Pandey, D. Singh Rajpoot, and M. Saraswat, "Twitter sentiment analysis using hybrid cuckoo search method," *Inf. Process. Manag.*, vol. 53, no. 4, pp. 764–779, 2017.
- [18] M. Stratford, "Welcome to Mendeley: Quick Start Guide," 2018.