International Journal of Engineering Applied Sciences and Technology, 2022
Vol. 7, Issue 9, ISSN No. 2455-2143, Pages 171-176
Published Online January 2023 in IJEAST (http://www.ijeast.com)

A REVIEW ARTICLE: USE OF SENTIMENT ANALYSIS IN SOCIAL MEDIA

Mohd Ahad
Department of Civil Engineering
Zakir Husain College of Engineering and Technology, Aligarh Muslim University
Aligarh (202002), Uttar Pradesh, India

Abstract— **As a result of the growth of online social networking platforms and applications, a sizeable amount of user-generated text content is created daily in the form of comments, reviews, and short text messages. Users can write messages, share them, and add images and videos to social networking sites like Twitter, Facebook, and others. Consequently, a significant volume of sentiment-rich data is generated. Sentiment analysis then comes into play in this scenario, which evaluates opinions as positive, neutral, or negative by extracting, recognizing, or representing them from various sources, including social media, news, articles, and blogs. This study aims to analyze the results from different sentiment analysis models and technologies that combine natural language processing. Case studies of various industries that can benefit or have been benefiting from sentiment analysis are also discussed to provide an approachable pathway for anyone who wishes to go more deeply into this field. For example, the business world has used it to learn what customers think of a certain company or brand. The impact of profanity on how readers interpret tweets and other social media messages are studied in sociology and psychology. Political scientists are trying to anticipate election results based on tweets to evaluate answers, among other things, and to look for trends, ideological bias, and opinions. Researchers have previously evaluated numerous models using well-known techniques like Naive Bayes, support vector machines, etc., and the findings have been compared with promising outcomes.**

Keywords— **Sentiment Analysis, Natural Language Processing, Twitter, Social Media Analysis, Machine Learning, Naive Bayes, SVM**

I. INTRODUCTION

People now communicate their ideas on issues and things differently thanks to the Internet's development and use. This has been improved by several platforms, including social media and email. For instance, social media has developed into a potent tool for exchanging information and facilitating conversation online. It offers a place, a way, and a platform for establishing new connections and openly sharing information. People may also communicate by posting brief messages on online forums, discussion boards, and product review websites. In recent years, blogs have also been popular as a tool to assess the preferences of the general public. because if a blog becomes popular on one platform, numerous others with related topics may start to appear.

Thanks to Natural Language Processing and especially sentiment analysis businesses can monitor how well their goods and services are performing using comments from social media. To help future product and service enhancement, they can acquire intelligence and business insights, separate potential clients from the general public, and carry out market segmentation. Omuya et al. (2022) [1] described in their research how social network mining applications of business analytics have not yet been completely investigated. For example, when someone wants to purchase a product, they would go online and check the reviews made on that product before making a decision. If the number of reviews generated is too many, it would be very challenging for either an individual or an organization to analyze. This process can thus be automated using sentiment analysis tools proposed by Yang et al. (2021) [2].

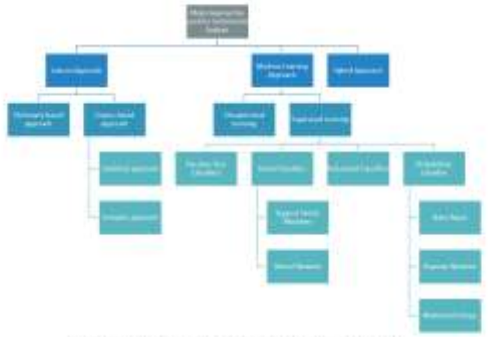
Sentiment analysis is a discipline that employs artificial intelligence (AI) and natural language processing (NLP) to ascertain how a certain demographic feels about a problem or a product. In the fields of political science, sociology, and psychology, sentiment analysis has also been used to examine patterns, ideological bias, opinions, and gauge responses, among other things, concluded research by Rezapour et al. (2020) [3]. In their research, Agarwal et al. (2018) [4] examined how people frequently utilize microblogs to express their feelings and sentiments about various events, including natural disasters, earthquakes, sports, and travel. Additionally, multimodal sentiment analysis has been used recently to assess the polarity of feelings in various contexts.

According to Cachola et al. (2018), [5] many disciplines must understand how profanity is expressed in naturally occurring text, including linguistics, which aims to better understand the pragmatics of profanity, computer science, which can explicitly model profanity in downstream NLP applications, and psychology, which studies the sociocultural factors of profanity. Social media is the ideal platform for observing and researching the use of vulgarity since the language there is highly expressive of thoughts, opinions, and feelings. As a result, researchers have access to vast volumes of user-generated information that spontaneously occurs on social media. As a result, it is believed that being able to understand vulgarity would be advantageous for sentiment research on social media.

Sharma et al. (2016) [6] determined how Hindi Twitter users felt about each of the Indian political parties under consideration. to determine popular support for these parties and forecast election outcomes. Later, Nugroho (2021) [7] used the VADER sentimental analysis model to successfully forecast the outcomes of the 2020 United States presidential elections.

In this paper, the author describes the sentiment analysis models and the statistical techniques for natural language processing concerning their application areas.

II. SENTIMENT ANALYSIS APPROACHES



Sentiment Analysis approaches. Source: Devopedia 2018.

Sentiment analysis seeks to ascertain the viewpoint held by the text's author or speaker concerning the subject matter or the document's overall contextual polarity (Mejova 2009) [8]. The major approaches used for sentimental analysis are

A. Lexicon Approach–

According to Gupta et al., (2020) [9] This approach determines the sentiment orientations of the entire text or group of sentences based on the semantic orientation of lexicons.

a. Dictionary-based Approach - This method starts by selecting a small number of words to use as a dictionary. Then, by adding synonyms and antonyms for those terms, a thesaurus, online dictionary, or WordNet may be used to extend that vocabulary. The dictionary is continually enlarged until no more words can be added.

b. Corpus-based Approach - Gupta et al. (2020) [9] concluded that this approach revealed the sentiment orientation of context-specific terms. There are two ways to use this strategy:

  i. Statistical Approach - Positive polarity is thought to exist in words that exhibit irregular positive activity. They have negative polarity if they exhibit negative recurrence in negative text. When a term appears equally often in both positive and negative literature, it is said to have neutral polarity.

  ii. Semantic Approach -By locating synonyms and antonyms for a phrase, this method assigns sentiment values to the word in question as well as terms that are semantically similar to it.

B. Machine Learning Approach –

It makes use of training sets to develop a rather simpler approach than the Lexicon approach. It aims to solve the problem of having huge amounts of data with many variables and is commonly used in areas such as pattern recognition (speech, images), financial algorithms (credit scoring, algorithmic trading) (Nuti et al. 2011) [10], energy forecasting (load, price) and biology (tumour detection, drug discovery). This system is subdivided into

A. Unsupervised Learning - When there is a lack of training data, it is implemented. Based on the data supplied, it creates predictive models.

B. Supervised Learning - When there are training data available, it is applied. Using data from both the input and the result, it creates a predictive model.

a. Decision trees - It continually divides the data based on a certain parameter. Decision nodes and leaves are the two components that may be used to explain the tree. The choices or results are represented by the leaves. The data is divided at the decision nodes.

  i. Linear Classifiers - The value of a linear combination of the features is used by a linear classifier to make classification decisions. When the problem is linearly separable, it functions better.

1. Support Vector Machines (SVM) - These are supervised learning models with corresponding learning algorithms used for classification and regression analysis that scan the data and identify patterns.

2. Neural Network -They enable computer programs to identify patterns and resolve common issues by mimicking the behaviour of the human brain.

  ii. Rule-based Classifier - The class determination is made via rule-based classifiers based on a variety of "if. Else" rules. These classifiers are typically used to produce descriptive models since the rules are simple to understand.

  iii. Probabilistic Classifier - Instead of just producing the class that the observation is most likely to belong to, a probabilistic classifier can predict, given an observation of an input, a probability distribution over several classes.

1. **Naive Bayes (NB)** - It was described as a straightforward probabilistic classifier by Treleaven et al. (2015) [11], based on the Bayes theorem application with strong (naive) independence assumptions, such as when features are independent of one another inside each class.

The resulting probability model is combined with a decision rule via the Naive Bayes classifier. The selection of the hypothesis with the highest likelihood is a typical practice.

iv. Bayesian Network - A joint probability distribution is represented by a compact, adaptable, and understandable Bayesian network. Due to the ability of directed acyclic networks to show causal relationships between variables, it is also a helpful tool in knowledge discovery.

1. Maximum Entropy - This probability distribution has the highest information-theoretical entropy and best captures the current state of knowledge.

C. Hybrid Approach –

It combines the use of both Lexicon and Machine Learning approaches to optimize the accuracy of the classification of feelings.

III. CASE STUDIES

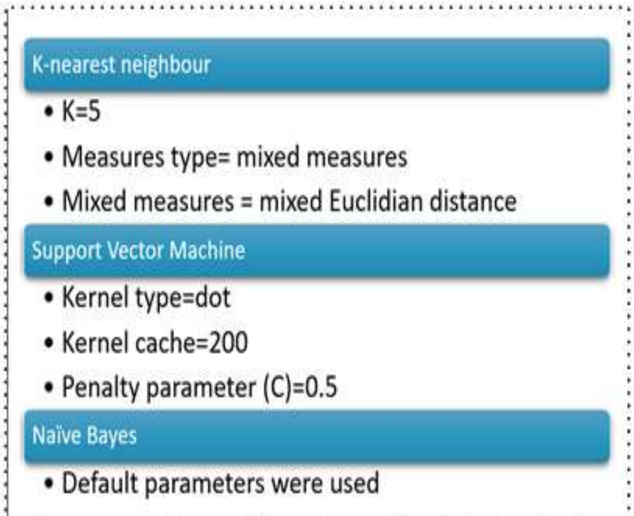
A. For Business Intelligence

In 2022, Omuya et al [1]. suggested a model for sentiment analysis employing machine learning algorithms, including NB, SVM, and K-nearest neighbour, using social media data and other data sets. The model's performance was examined and contrasted with that of other cutting-edge models. The utilization of diverse sections of speech, training the model on pre-processed data sets, and lowering dimensions all significantly enhance the performance of sentiment analysis models, according to experimental results. Their suggested model can be used, particularly in business intelligence, to comprehend the arbitrary reasons why consumers are or are not responding to something, which can be the reasons why consumers are buying a particular product; what the customers think of the user experience for the products. or the utilized services; whether the customer service support exceeded their expectations; and so forth.

a. Employed Approach

The Twitter data utilized for the tests in this paper came from the open-source sentiment140 social media data repository, which was developed by Alec et al (2018) [12]. Both the cross-validation approach, which used 10-folds with an automatic sampling type, and the split method, which divided the data into 30% for training and 70% for testing, were used for validation. The model was subsequently subjected to sentiment analysis using the NB, SVM, and K-nearest neighbor machine learning algorithms. The performance of these algorithms was examined using four performance metrics: accuracy, precision, recall, and F-measure.

Their model's performance was compared to the findings from the other two models used for sentiment analysis. The first model proposed by Fouad et al. (2018) [13] used feature selection and classifier ensemble. The second model proposed by Zafar et al. (2018) [14] is for sentiment analysis in a short text.



K-nearest neighbour
- K=5
- Measures type= mixed measures
- Mixed measures = mixed Euclidian distance

Support Vector Machine
- Kernel type=dot
- Kernel cache=200
- Penalty parameter (C)=0.5

Naïve Bayes
- Default parameters were used

The parameters that Omuya EO and colleagues utilized in their experiment when proposing ML algorithms.

b. Results

In terms of accuracy, the suggested model performed much better than the other models that made use of the same data set. Additionally, it was able to clean data, decrease data dimensions, and eliminate noise, which improved accuracy and performance. It was also typically reliable and consistent. Cross-validation and the split technique both showed excellent results for the NB and K-nearest neighbour machine learning algorithms.

B. In Politics

In an effort to forecast political election outcomes based on Twitter, experts have been working on this project for more than ten years. Tumasjan et al. (2010) [15] and Sanders and Van den Bosch (2013) [16] reported some encouraging early findings, but shortly articles expressing scepticism about the accuracy of election outcome prediction based on tweets surfaced (Gayo-Avello, 2012) [17]. Despite these results, which lend credence to the scepticism, studies on attempts to predict elections using tweets continue to be published, frequently incorporating sentiment analysis (Nugroho, 2021 [7]; Batra et al., 2020 [18]; Rao et al., 2020[19]).

a. Employed Approach

In this part, the work of Nugroho (2021) [7], who managed to forecast the results of the US presidential elections held on November 2020, will be examined. His outcomes will be contrasted with the real poll findings.

According to B. Liu (2010) [20], a sentiment lexicon is a collection of lexical skills (such as words) that may be categorized as either positive or negative depending on their semantic orientation. This definition was used in the study by the author. The following is a discussion of the author's analysis process:

i. Data Collection: It is done to extract the content from user-generated tweets on the elections. and the information was gathered as follows:

a. The creator sets up Twitter developer authentication for the application being used

b. A database was created to house the twitter data that had been gathered. The database was categorized as follows:

i. Sentiment - It noted the data's positivity, negativity, neutrality, compoundness, etc.

ii. Tweet - It has information like the user id, tweet id, etc.

iii. Last scraping - This table was used to determine how frequently the information was obtained from Twitter.

iv. User - This table served as a repository for the user's public information, including name, surname, followers, location, etc.

c. When storing data and tweets, checks are made to prevent data duplication.

d. Data was gathered one week before the commencement of the general election in the United States.

ii. Data Pre-processing: The pre-processing of data was conducted in two stages:

a. Data Wrangling - To convert the unformatted raw data collected from Twitter into a comprehensible format.

b. Data Cleaning - Characters that are not necessary for analysis, such as punctuation, digits, whitespace, and capital letters, were eliminated.

iii. Data Mapping: To ensure that only tweets from US citizens were examined, areas were mapped according to user location. To do this, the Tweepy API was used, and the necessary adjustments were implemented.

iv. Sentiment Analysis: Nugroho (2021) [7] utilized the VADER model to conduct lexicon-based emotional analysis on tweets. VADER (Valence Aware Dictionary and sentiment Reasoner), according to Hutto et al. (2014) [21], is a tool-based sentiment that is especially linked with social media feelings. In addition to classifying emotions as good or bad, VADER also indicates how positive or negative they are. The sentiment analysis was carried on the following manner:

a. The public's perceptions of both Joe Biden and Donald Trump as presidential contenders were assessed using the VADER model.

b. The findings of the sentiment analysis were categorized by state.

c. To categorize candidate successes based on the parties from each state, the percentage of emotion by the state was calculated.

d. The information is then categorized based on candidate-bearing parties for mapping the state following specified criteria, such as whether the state is Solid Republican, Solid Democrat, Lean Republican, and so on.

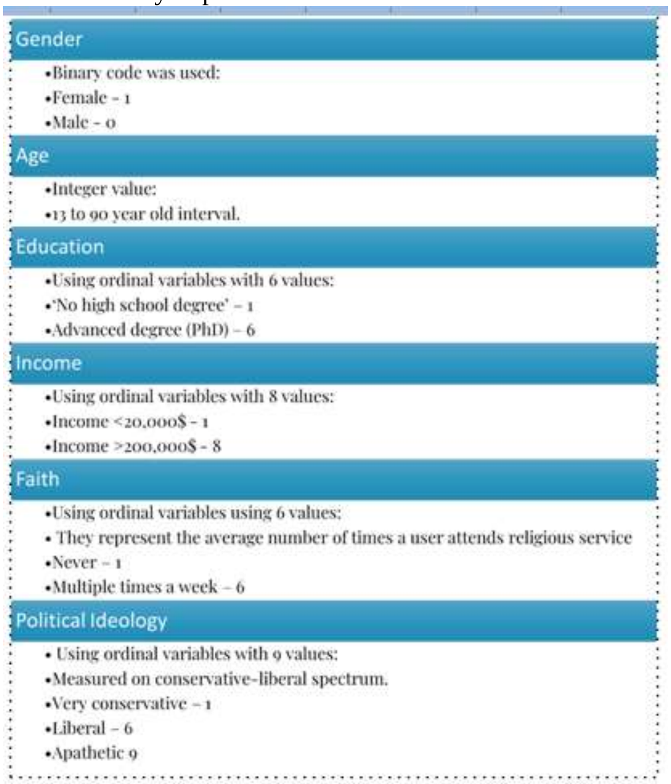
v. Lastly, the predicted data and the actual outcomes are compared.

b Results

According to an emotional analysis model developed by Nugroho (2021) [7]. The Democratic Party was expected to obtain 22 votes, surpassing the Republican Party's 19 votes. According to the BBC's results, the Democratic Party received 24 votes to the Republican Party's 20, giving them the victory. These findings indicate that the VADER sentiment analysis model was able to forecast the outcome of the US presidential elections.

C. In Sociocultural Matters

Given that most ideas can be rephrased so as not to contain vulgarity, the use of vulgar words denotes a deliberate attempt to fulfil a specific function, such as an intensifier for personal opinions, a way to offend or express hate speech toward others, a way to describe immoral behaviour, or a way to indicate an informal conversation. Thus, it is crucial to comprehend how obscenity is expressed in the real-world text since doing so may help with linguistic and psychological processes as well as subsequent NLP tools like sentiment analysis. In their work, Cachola et al. (2018) [5] used data from social media to conduct an extensive, data-driven empirical investigation of coarse language. Using tweets from people with known demographics, they examined the sociocultural and pragmatic components of profanity. Additionally, they gathered sentiment scores for derogatory tweets to research the connection between the usage of derogatory language and perceived sentiment and demonstrate how explicitly modelling derogatory language may improve sentiment analysis performance.



Gender

- •Binary code was used:
- •Female - 1
- •Male - 0

Age

- •Integer value:
- •13 to 90 year old interval.

Education

- •Using ordinal variables with 6 values:
- •'No high school degree' – 1
- •Advanced degree (PhD) – 6

Income

- •Using ordinal variables with 8 values:
- •Income <20,000$ - 1
- •Income >200,000$ - 8

Faith

- •Using ordinal variables using 6 values:
- • They represent the average number of times a user attends religious service
- •Never – 1
- •Multiple times a week – 6

Political Ideology

- • Using ordinal variables with 9 values:
- •Measured on conservative-liberal spectrum.
- •Very conservative – 1
- •Liberal – 6
- •Apathetic 9

a Employed Approach

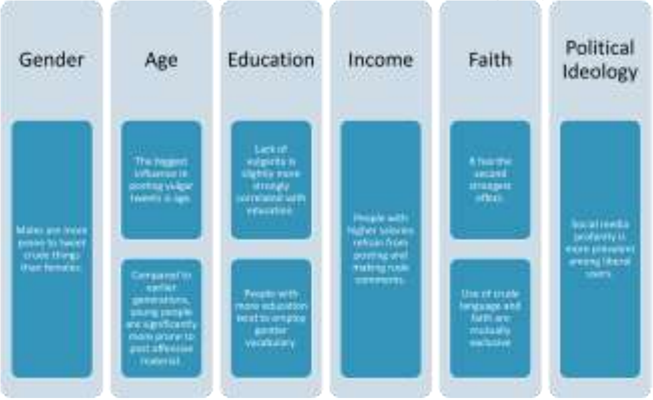
They compile a novel corpus of tweets, which were purposefully chosen among those sent by individuals who had self-reported certain sociodemographic characteristics in an online survey.

i. They began by using the profanity vocabulary found at www.noswearing.com to identify tweets that contained vulgarity.
ii. They used tokens to substitute usernames and URLs to maintain anonymity. Additionally, all words were lowercase, and punctuation was eliminated.
iii. The Amazon Mechanical Turk (MTurk) network was used to request human sentiment assessments.
iv. Their job structure and rules are in accordance with SemEval 2016 Task 4 subtask C. (Nakov et al., 2016 [22]). Tweet sentiment was rated by annotators on a scale of 1 to 5: very negative, moderately negative, neutral, somewhat positive, and very positive.
v. They employed partial Pearson correlation, where the dependent variable is the percentage of offensive tweets among all of the user's tweets.

They took into account gender and age as fundamental features for all analyses and used both variables as partial correlation controls to account for data skew.

b Results

The following conclusions are drawn from an examination of sociocultural and pragmatic components of vulgarity



Males are more prone to tweet crude things than females

The biggest influence in posting vulgar tweets is age.

Compared to earlier generations, young people are significantly more prone to post offensive material.

Lack of vulgarity is slightly more strongly correlated with education

People with more education tend to employ gentler vocabulary

People with higher salaries refrain from posting and making rude comments.

It has the second strongest effect.

Use of crude language and faith are mutually exclusive

Social media profanity is more prevalent among liberal users

IV.CONCLUSION

In the opening to this review article, an overview of how the development of social networking platforms and apps results in the production of a sizeable amount of sentiment-rich data in the form of tweets, comments, reviews, brief text messages, etc is provided. And how sentiment analysis is applied to extract, recognize, or portray these opinions from these social media sites, articles, blogs, etc. to categorize them as positive, neutral, or negative. Therefore, a rising number of application sectors are highlighted to show how sentiment analysis has an impact on society, business, politics, etc. The vast array of methods used in sentiment analysis is then briefly summarized (except for the Naive Bayes Classifier, which was presented in length due to its widespread use). To not only present a clear roadmap for someone thinking about beginning their research or profession in sentimental analysis for the relevant sector, but also to further elaborate its scope, case studies of numerous industries for which various effective models have been constructed are reviewed in detail.

The author believes that this study will be beneficial to researchers as well as freshers in the field of Natural Language Processing, primarily in Sentimental Analysis.

V. REFERENCE

[1]. Omuya EO, Okeyo G, Kimwele M. (2022) Sentiment analysis on social media tweets using dimensionality reduction and natural language processing. Engineering Reports. 2022;e12579. doi: 10.1002/eng2.12579

[2]. Yang S, Xing L, Li Y, Chang Z. (2021) Implicit sentiment analysis based on graph attention neural network. Eng Rep. 2021;4:e12452. doi:10.1002/ eng2.12452

[3]. Rezapour, M.(2021) Sentiment classification of skewed shoppers’ reviews using machine learning techniques, examining the textual features. Eng Rep 2021; 3:e12280. doi:10.1002/eng2.12280

[4]. A.Agarwal and D. Toshniwal, (2018) "Application of Lexicon Based Approach in Sentiment Analysis for short Tweets," 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE), 2018, pp. 189-193, doi: 10.1109/ICACCE.2018.8441696.

[5]. Isabel Cachola, Eric Holgate, Daniel Preoţiuc-Pietro, and Junyi Jessy Li. (2018). Expressively vulgar: The socio-dynamics of vulgarity and its effects on sentiment analysis in social media. In Proceedings of the 27th International Conference on Computational Linguistics, pages 2927–2938, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[6]. P. Sharma and T. -S. Moh (2016), "Prediction of Indian election using sentiment analysis on Hindi Twitter," 2016 IEEE International Conference on Big Data (Big Data), 2016, pp. 1966-1971, doi: 10.1109/BigData.2016.7840818.

[7]. D.K. Nugroho, (2021) "US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), 2021, pp. 136-141, doi: 10.1109/Confluence51648.2021.9377201.

[8]. Mejova Y (2009) Sentiment analysis: an overview, pp 1–34. http:// www.academia.edu/291678/Sentiment_Analysis_An _Overview. Accessed 4 Nov 2013

[9]. Neha Gupta, Rashmi Agrawal,(2020) Chapter 1 - Application and techniques of opinion mining, Editor(s): Siddhartha Bhattacharyya, Václav Snášel, Deepak Gupta, Ashish Khanna, In Hybrid Computational Intelligence for Pattern Analysis and Understanding, Hybrid Computational Intelligence, Academic Press, 2020, Pages 1-23, ISBN 9780128186992, https://doi.org/10.1016/B978-0-12-818699-2.00001-9.

[10]. G. Nuti, M. Mirghaemi, P. Treleaven and C. Yingsaeree, (2011) "Algorithmic Trading," in Computer, vol. 44, no. 11, pp. 61-69, Nov. 2011, doi: 10.1109/MC.2011.31.

[11]. Batrinca, B., Treleaven,(2015) P.C. Social media analytics: a survey of techniques, tools and platforms. AI & Soc 30, 89–116 (2015). https://doi.org/10.1007/s00146-014-0549-4

[12]. Alec, G, Bhayani, R, Lei H. (2018) Sentiment140 Repository. Stanford; 2018. http://help.sentiment140.com/for-students.

[13]. Fouad M, Gharib T, Mashat A. (2018) Efficient Twitter Sentiment Analysis System with Feature Selection and Classifier Ensemble; 2018. doi:10.1007/978-3-319-74690-6_51

[14]. Zafar L, Afzal M, Ahmed U. (2018) Exploiting polarity features for developing sentiment analysis tool. EMSASW; 2018

[15]. Tumasjan, A., Sprenger, T., Sandner, P., & Welpe, I. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. Proceedings of the International AAAI Conference on Web and Social Media, 4(1), 178-185. https://doi.org/10.1609/icwsm.v4i1.14009

[16]. Sanders, E. and Van den Bosch, A. (2013). Relating Political Party Mentions on Twitter with Polls and Election Results. In Proceedings of DIR-2013.

[17]. A.Gayo-Avello, (2012) "No, You Cannot Predict Elections with Twitter," in IEEE Internet Computing, vol. 16, no. 6, pp. 91-94, Nov.-Dec. 2012, doi: 10.1109/MIC.2012.137.

[18]. P. Khurana Batra, A. Saxena, Shruti and C. Goel, (2020) "Election Result Prediction Using Twitter Sentiments Analysis," 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC), 2020, pp. 182-185, doi: 10.1109/PDGC50313.2020.9315789.

[19]. Rao, Dr. D Rajeswara and Usha, S and Krishna, S Sri and Ramya, M Sai and Charan, G Sri and Jeevan, U, (2020) Result Prediction for Political Parties Using Twitter Sentiment Analysis (July 10, 2020). International Journal of Computer Engineering and Technology 11(4), 2020, pp. 1-6., Available at SSRN: https://ssrn.com/abstract=3648001

[20]. Liu, (2010) “Sentiment analysis and subjectivity,” Handb. Nat. Lang. Process. Second Ed., pp. 627-666, 2010.

[21]. Hutto and E. Gilbert, (2014) “VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text”, ICWSM, vol. 8, no. 1, pp. 216-225, May 2014.

[22]. Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. (2016). Semeval-2016 task 4: Sentiment analysis in twitter. In Proceedings of the 10th International Workshop on Semantic Evaluation, pages 1–18. https://doi.org/10.48550/arXiv.1912.01973

[23]. Devopedia. (2022). "Sentiment Analysis." Version 52, January 26. Accessed 2022-10-09. https://devopedia.org/sentiment-analysis