Sentiment Analysis of Twitter Data to Detect and Predict Political Leniency Using Natural Language Processing

V. V. Sai Kowsik[1†], L. Yashwanth[1†], Srivatsan Harish[1†], A. Kishore[1†], Renji S[1†], Arun Cyril Jose[1*†]

[1*]Department of Cyber Security, Indian Institute of Information Technology, , Valavoor, Kottayam - 686635, Kerala, India.

*Corresponding author(s). E-mail(s): aruncyril@iiitkottayam.ac.in ;
Contributing authors: saikowsik2019 @ iiitkottayam. ac. in; lakavathyashwanth2019 @ iiitkottayam. ac. in; srivatsan2019@ iiitkottayam. ac. in; agirishettykishore2019@ iiitkottayam. ac. in; renjiarun.22phd11006@iiitkottayam.ac.in;
† These authors contributed equally to this work.

Abstract

This paper analyses twitter data to detect the political lean of a profile by extracting and classifying sentiments expressed through tweets. The work utilizes natural language processing, augmented with sentiment analysis algorithms and machine learning techniques, to classify specific keywords of interest as either positive or negative, based on the sentiment score of each keyword. The proposed methodology initially performs data pre-processing, followed by multi-aspect sentiment analysis for computing the sentiment score of the extracted keywords, which are then used for precisely classifying users into various clusters based on similarity score with respect to a sample user in each cluster. The proposed technique also predicts the sentiment of a profile towards unknown keywords and gauges the bias of an unidentified user towards political events or social issues. The proposed technique was tested on Twitter dataset with 1.72 million tweets taken from over 10,000 profiles, and was able to successfully identify the political leniency of the user profiles with 99% confidence level. The paper could also identify the impact of political decisions on various clusters, by analyzing the shift in number of users belonging to the different clusters.

Keywords: Clustering Methods, Keyword Search, Natural Language Processing (NLP), Machine Learning (ML), Sentiment Analysis, Social Networks, Twitter.

1 Introduction

The evolution of online social media (OSM) over the past few decades has witnessed a myriad of advancements in the field of information technology. The meta reports [1], first quarter (March 2023), reveals the average daily active users (DAUs) as well as the monthly active users (MAUs) on Facebook to be 2.04 and 2.99 billion, indicating an increase of 4% and 2% users over the previous year. Human interconnectivity in OSM can be understood through various layers [2] or concentric circles that represent different aspects of online interaction. These layers highlight the enormous ways in which OSM enables interconnectivity and information diffusion in online social networks (OSN)[3], ranging from trust in personal relationships [4] to professional networks, interest-based communities [5], information sharing through collective intelligence(CI) [6], influencer and brand communities, and public discourse. Each layer offers unique opportunities for individuals to connect, engage, and participate in the digital landscape.

Twitter's unique characteristics and widespread usage [7] make it a prominent information-sharing platform. Opinion formation [8][9] in online social networks burgeons as an important tool for individuals, business organizations, and public figures to connect, express themselves, and engage with a global audience in real-time [10]. Political bias can indeed be present in Twitter tweets [11] due to the subjective nature of political opinions [12] and the diverse range of users on the platform. Hence, online social media platforms are a significant source of big data[13], which serves as the key to understanding and predicting user perceptions[14], leading to the generation of online communities of interest[15].

The key part of any tweet is the hashtag [16] which plays a significant role in categorizing the publicly available content, thereby making it readily searchable for users. Hashtags serve as an automated mechanism for identifying the profile's disposition toward the subject or the overall contextual polarity by utilizing Natural Language Processing (NLP) [17] capabilities. Automatic keyword extraction[18] and text summarization task of NLP breaks the text into constituent parts, categorizes it, and grabs the essence of the topic to predict the user's lean towards a particular topic, as to whether positive or negative. Sentiment analysis using NLP[19] is often used to understand the emotional tone behind texts, such as tweets, feedback or comments, and reviews published by users on social media forums [20][21].

The proposed methodology analyses the political lean[22] of a profile by extracting keywords from Twitter tweets extracted from the Kaggle dataset during the time of the 2020 US election. Multi-aspect-based sentiment analysis is performed on the extracted keywords and profile classification is carried out based on similarity score. Sentiment scores are generated for classification on a positive vs. negative scale. The methodology also predicts the sentiments of unknown aspects of a person through

his published tweets. The contributions of this paper are as follows:

- identifying the optimal number of clusters to improve the performance of the classification process. The methodology proves that four-cluster classification gives better performance than two-cluster classification.
- shift in the number of republicans and democrats with keywords nullified or policy shift, which could predict the impact of future political decisions on the society.
- predicts the political lineancy of a person through his published tweets.

The rest of this paper is organized as follows: Section 2 focuses on related works closely aligned with the proposed work. Section 3 elaborates on the methodology, followed by the experimental setup in Section 4. Results and discussions are dealt with in Section 5, and the Conclusion with a focus on future works is summarized in Section 6.

2 Related Work

M. Wongka et al. [21] conducted a thorough investigation into how the public feels about the candidates for Indonesia's presidential election, in 2019. They extracted tweets from the Twitter platform using a data crawler and text mining was implemented using the Naive Bayes (NB) classifier. Text analysis and tokenization were applied to convert the sentences to simple words, on which text mining was carried out that could achieve an accuracy of 80.90%.
The technique proposed by Fagni et al. [22] learns latent political ideologies using a deep neural network. Users are grouped into a superficial ideology space and clustered. The cluster to which a user belongs determines their political inclination. They obtained the best results among all unsupervised techniques and paved the way for the future development of meticulous political emotion prediction using unsupervised approaches.
S.Kayiki[23] proposed a new method called SenDemonNet to gain insight into the community sentiment through Twitter tweets, regarding the demonetization policy implemented by the Indian Government in November 2016. They analyzed the tweets which are samples of societal opinions utilizing the "Twitter package" in R together with Twitter API. Initially, they performed tweet pre-processing, followed by feature extraction and weighted feature selection using a hybrid Forest-Whale Optimization algorithm (F-WOA)[24] with heuristic deep neural networks. SenDemonNet outperforms its competitors in terms of classification accuracy. However, the classification of unknown keywords nor the shift in classification is not considered, which is addressed in this proposed work. M. Wankhade et al. [25] investigated the various classification methods for sentiment analysis at its various levels, along with the procedures for data collection and feature selection. They discussed the pros and cons of each classification technique with benchmark techniques such as NB and support vector machine (SVM). Q. You et al.[26] identifies users' attributes and interests from images posted by them on social networks, which they posit to enhance the performance of the recommender system.
A.Ligthart et al. [27] carried out a tertiary study of sentiment analysis to provide a

comprehensive overview of a plethora of methods, including deep learning algorithms and different datasets conducive to sentiment analysis. P. Berka[28] carried out sentiment analysis using rule-based and case-based reasoning to evaluate the strength of opinion mining on unstructured data. To filter out spam tweets in Twitter datasets, S. Sedhai and A. Sun[29] proposed a framework that learns the patterns of new spam activities that involves the detection of blacklisted URLs, near duplicates of confidentially labeled tweets, spam word detection and multi-classifier-based detectors. S. M. Park and Y. G. Kim [30] proposed a method that analyses the causal relationship between the sentiment words. They analyzed the root cause of negative opinion in sentiment analysis.
S. M. Nagarajan and U. D. Gandhi [31] proposed a methodology using a hybridization technique that uses optimization methods and machine learning classifiers for analyzing Twitter data. They could detect spammers and protect legitimate users from unwanted urls and irrelevant messages through data pre-processing. N. Zainuddin et al. [32] suggest a hybrid sentiment classification technique for Twitter datasets from different domains by embedding a feature selection method for aspect-based sentiment analysis. H. Liu et al. [33] and L. Luceri et al.[34], performs a comparative analysis of various deep learning methods for aspect-based sentiment analysis influenced by social media on public opinions. S. Stieglitz et al. [35] present the challenges faced by researchers during various phases, such as data discovery, collection, and preparation, suggesting potential solutions that could aid the researchers in effectively critiquing social media data. A. R. Pathak et al. [36] propose a technique for sentiment analysis that works at the sentence level and uses a neural network for performing semantic analysis.
F. Cena et al.[37], highlights the negative bias of users for generating recommendations that prove to have more impact than positive preferences. Their proposed methodology is fruitful when only a few positive recommendations are available, but it is harder to acquire. L.M. De Campos et al.[38], put forward the technique to identify the politicians, and expertise in various fields based on their profile as well as from their political speeches, which proved to be much better than the expert recommendation task. However, they did not consider the shift in political thoughts of the politicians with changing government policies, which is addressed in our work by predicting the political lineancy of a person with chaniging policies on key issues.
S. Abdi et al.[39] introduced a methodology to learn users' emotions and to predict their future opinions, in order to better understand the users feeling which proved to be critical in the decision-making process. However, categorizing users based on their emotions and identifying the strength of users who could contribute positively to changing organizational policies has not been addressed in their work.
However, the proposed methodology predicts the user sentiments, based on keywords extracted from tweets and also predicts the bias of unknown users towards a particular topic of interest. The paper proposes an efficient method to detect the bias of a profile towards a topic or group of topics based on their activity on Twitter, which relies on Natural Language Processing (NLP) to perform this task. Our work classifies the users into four categories, namely, the extreme Republicans, moderate Republicans, extreme Democrats, and moderate Democrats for carrying out the analysis.

3 Methodology

The architecture consists of four main components: Data set collection and pre-processing, topic extraction and sentiment analysis, client-based analysis, and sentiment prediction of new user aspects, as elaborated in Fig. 1.

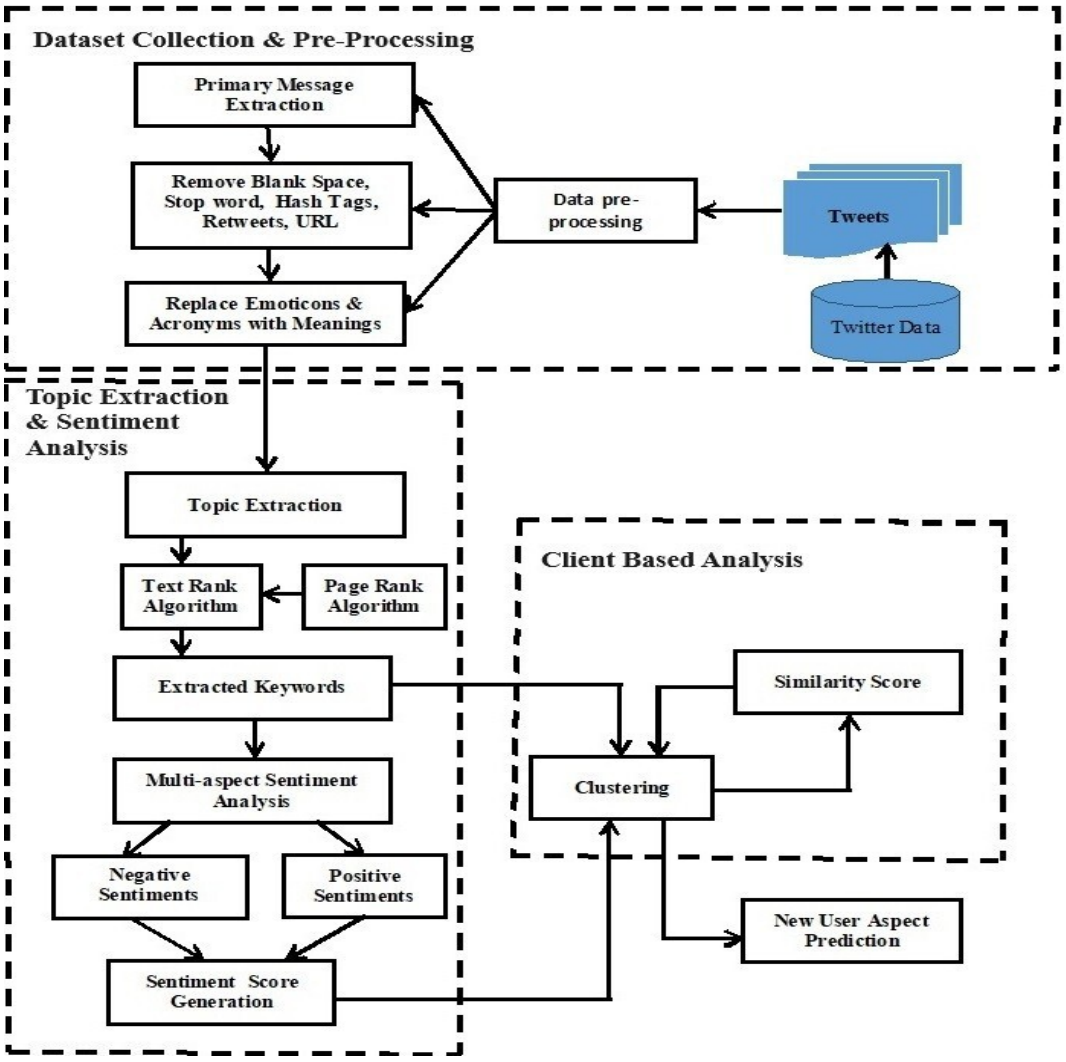

Fig. 1 Architecture diagram

3.1 Dataset Collection and Data Pre-processing

Twitter tweets related to a particular topic or group of topics along with multiple users related to different tweets are collected and pre-processed by removing the empty space, special characters, stop words, emoticons, hashtags, timestamps, and URLs, using Natural Language Toolkit (NLTK)[40], a commonly used Python library for NLP. In data pre-processing, the primary message of each tweet is extracted from which blank spaces, stop words, hashtags, repetitive words, and URLs are removed. Emoticons and acronyms are then replaced with their meanings.

3.2 Topic Extraction and Sentiment Analysis

Topic Extraction distils the main theme from tweets using NLP techniques such as keyword extraction for determining the sentiment of each topic using sentiment analysis or opinion mining [8].

Keyword extraction [18] filters the frequently used and most important words and expressions from text, thereby summarizing [41] the key contents of the text. In this work, the TextRank algorithm is used, which is a graph based ranking algorithm that applies the principle of PageRank algorithm for keyword extraction. PageRank [42] calculates the weights for web pages, where all web pages are considered as nodes in a directed graph. In TextRank algorithm, a document is divided into multiple sentences and only words with specific Parts of Speech (POS) tags such as, NOUN, PROPN, or VERB POS tags [43] are considered for ranking. Each word is a node in TextRank and the window size, 'k' determines the co-occurrence relationship between words or sentences. Words within a distance of 'k' are said to be connected in the graph indicating their proximity. The TextRank algorithm takes the text as input, along with the desired POS tags, window size, lowercase option, and stop words if any, that are to be filtered out. It then pre-processes the text, builds the graph, calculates the node weights and returns the keywords based on their importance in the text input.

Multi-aspect sentiment analysis is performed on each keyword using Valence Aware Dictionary and Sentiment Reasoner (VADER)[44] and NLTK libraries, to determine the polarity of a profile towards a particular keyword. Multi-aspect sentiment analysis [45] analyses text data to determine the sentiment towards multiple aspects or topics. VADER is a vocabulary and rule-based sentiment analysis tool[46], customized precisely to the perceptions of users expressed in online social networks. VADER makes use of a variety of sentiment lexicons, which are collections of lexical elements (such as words), often classified as either positive or negative depending on their semantic inclination. The various steps involved in the keyword extraction process is listed in Algorithm 1.

NLTK is a versatile Python library with a wide range of utilities and functionalities for tasks such as tokenization, stemming, lemmatization, POS tagging and syntax parsing that effectively manipulates and analyses linguistic data. NLTK includes pre-trained models for POS tagging, which assigns grammatical tags to words in a sentence.

The algorithm for multi-aspect sentiment analysis is as follows:

The pre-processed data is used to perform sentiment analysis by creating a data

Algorithm 1 Keyword Extraction(Input_Text)

Require: Input: Text
Ensure: Return (keywords)

1: Read the text
2: Performs spell check for the words in the text
3: Create an instance 'x', of TextRank class.
4: Invoke Analyse function of TextRankclass using instance x, with Input_Text as parameter.
5: Call the x.get_keywords(c), where 'c' represents the count of keywords to be generated.
6: Retain all missing hashtags in the original text by adding them to the list of keywords generated in step 5
7: Return(keywords)

Algorithm 2 MultiAspectSentimentAnalysis ( )

1: Tokenize input text to a list of sentences.
2: Create an empty list called aspect_scores_list to store the sentiment scores of each sentence containing the aspect.
3: Loop until the last statement in text:
4: Check if the aspect (in lowercase) is present in the sentence using the 'in' operator.
5: If the aspect is present, calculate the sentiment score, aspect_score.
6: Append the aspect_score to the aspect_score_list.
7: End

frame, grouped by username. Topic extraction is done on every tweet, and the sentiment score of each topic is evaluated as mentioned in Algorithm 2. Here, multiple topics may get repeated throughout the tweets of a user resulting in multiple sentiment scores, in which case the average of all the sentiments obtained becomes the overall sentiment score for the user profile.

3.3 Client Based Analysis

Client-Based analyses perform clustering on the extracted keywords and their corresponding sentiments to determine which group of topics the profile lean towards. Clustering groups keywords with similar sentiments to determine a profile's inclination towards to a topic for determining the bias based on their activity on Twitter.
In this work, clustering is performed based on similarity score between users and an arbitrary number of clusters initialized with a sample user placed in each cluster, representing the overall characteristics of users in that cluster. The sample acts as a dictionary, with keywords as keys and the sentiment scores as values. The average of all the sentiment scores represent the overall sentiment score of the different keywords used at various instances in the tweets of the user. Users are assigned their respective clusters based on the highest similarity score with respect to the sample users of each cluster.The highest value of similarity score refers to the least value of

Euclidean distance with respect to the sample users assigned to each cluster, which means that the respective sentiments are similar or same, and the similarity score is high with more matching keyword.
Similarity score computation is carried out by comparing two users based on their common interest, where each user is a dictionary whose keys are a set of pre-determined aspects subjective to the user. The similarity score of a particular topic for a particular user is a value ranging from 0 to 1, where 0 represents very high dissimilarity and 1, a high similarity index. The proposed methodology identifies similar interests of users by key intersection procedure that determines their common interests based on a set of pre-determined keys in their dictionary. If there are no common interests, the score is set to 0, otherwise, Algorithm 2 computes the Euclidean distance [47] between the sentiment scores of the two users quantified across their common interests. An aggregate similarity score of 1 indicates identical sentiments across their common interests, whereas 0 represents dissimilar interests.
To compute the similarity score, $S_i, 1 \leq i \leq 4$, for a new user 'N' with respect to the sample users, $T_i, 1 \leq i \leq 4$, in each of the 4 clusters, the following mathematical model has been adopted:

Let W_i be the set of keywords for sample users in the respective clusters such that:

$W_i = k_{i1}, k_{i2}, k_{i3}, \ldots, k_{in_i}$, where, $1 \leq i \leq 4$, and n_i is the total number of keywords of sample users in each cluster.
For the new user 'N', let 'K' be the keyword set for which the sentiment score is to be found:
$K = K_1, K_2, K_3, \ldots, K_m, where, m \geq 1$
Common topics form the intersection of the two sets W_i and K, represented by,
$I_i = W_i \cap K$, where, $1 \leq i \leq 4, n \geq 1$
I_i is then used to compute the similarity score as follows:
$S_i = SentimentScore(I_i)$
$S_k = SentimentScore(K)$
$C_i = Min(Euclidean\_distance(S_i, S_k)), where 1 \leq i \leq 4, n \geq 1$
where the Euclidean distance between S_i and S_k, for $1 \leq i \leq 4$ is:

$$Euclidean\_distance(S_i, S_k) = \sqrt{(S_i - S_k)^2} \tag{1}$$

The Euclidean distance is computed using equation(1), and the new user 'N' is assigned to cluster C_i, which is the minimum of all four Euclidean distances of the user 'N' with respect to the sample users in each cluster. The Euclidean distance computed is normalized to a range of 0 to 1 to get the similarity score as follows:

$$Similarity\_Score = \frac{1}{(1 + Euclidean\_distance(S_i, S_k))} \tag{2}$$

The confidence score of the new user 'N' is computed from the similarity score obtained from equation(2) as follows:
$CS_N = Avg(S_i)$, for all U_i with respect to N, where $S_i \geq 0.9$. If there are no common keywords, the similarity score returned will be zero.

3.4 Sentiment Prediction of New User Aspect

To examine the effect of nullified keywords on a topic or a group and their impact on cluster formation, clustering is performed by removing the keywords from the list of keywords suggested by expert recommendations and analyzing their impact on the clustering process. To predict the sentiment of a particular topic, not present in the profile, the proposed technique first identifies the cluster to which the particular user could belong to based on his existing keyword list. Similarity scores are then evaluated with respect to each user in the currently assigned cluster. The scores are then sorted and the aggregate sentiment score of the top ten similar users are computed for the specified topic, which represents the overall sentiment of the considered user towards the topic.

4 Experimental Setup

The proposed methodology was experimented on Twitter dataset collected from Kaggle [48]. The dataset contains 1.72 Million tweets collected during the time span of 15.10.2020 to 08.11.2020. Relevant data vital for analysis has been extracted from the dataset, "hashtag_donaldtrump Version 19" [48], and pre-processed to get a data frame that contains two attributes namely, 'tweet' and 'username'.
In this study, a data frame has been created with the first 100,000 rows extracted from the original data frame, and a new data frame is moulded by grouping the 'username' attribute. Keyword extraction is performed for the first 1000 unique usernames and for each user tweet as mentioned in Algorithm 1. A subsidiary data frame with two attributes namely, username and dictionary is created for mapping between users and their related keywords. The sentiment scores for each keyword used across all the tweets of a particular username at different scenarios are computed as mentioned in Algorithm 2, from which the aggregate sentiment score is generated.
The primary focus of this paper is to compute the confidence score by taking the average similarity score of users in the cluster with a similarity score greater than 0.9, who are considered to be most similar to the current user that is being assigned to the cluster. Since only the highest matching scores were used, the methodology was designed by considering zero errors in classification and the primary focus was on classifying unknown users based on their tweets which has resulted in a very high confidence score of 99%, which proves the efficiency of the methodology.

5 Results and Discussions

5.1 Classification of Tweets

The sentiment score of 1,000 users towards the topic 'trump' has been considered by categorizing them into two groups. The first group contained 311 users with positive sentiment toward the topic 'trump', whereas the second had 492 users with negative sentiment toward the topic using the multi-aspect sentiment analysis technique mentioned in Algorithm2.

Fig. 2. represents clustering of users based on the topic 'trump'. In the graph, each

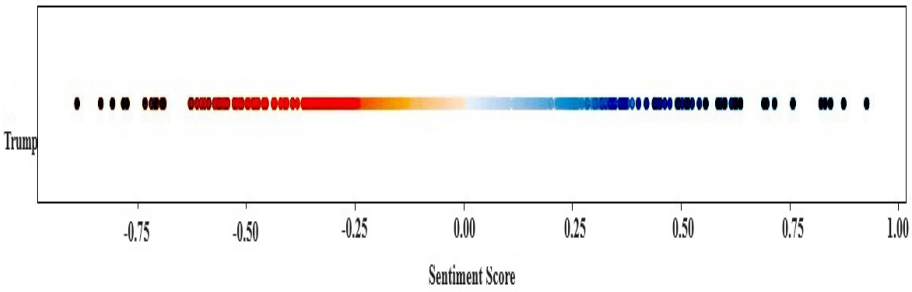

Fig. 2 Positive Vs. negative sentiment score of 'Trump'

dot represents a single user and x-axis represents the sentiment score of users computed by Algorithm 2, extracting the keyword 'trump' from tweets given as input to the algorithm.

Fig. 3. represents users who have positive sentiment towards topic 'trump', x-axis

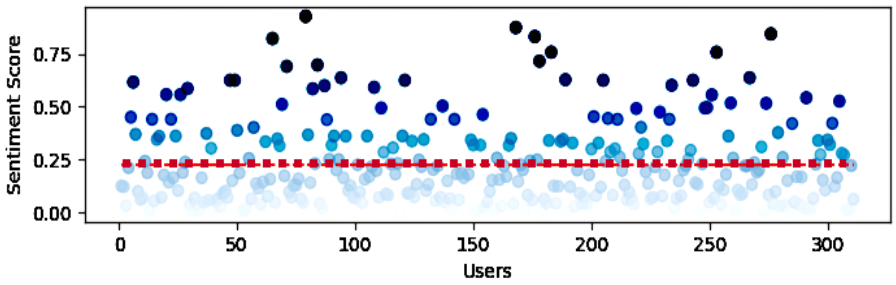

Fig. 3 Positive sentiment towards 'Trump'

shows the number of users and y-axis represents the sentiment score. Dense dotted line on the graph represents the average sentiment of all users who are positive towards the topic 'Trump'.

Fig. 4. represents the plot of the users who have negative sentiment towards topic 'trump', where x-axis denotes the number of users and y-axis, the sentiment score.

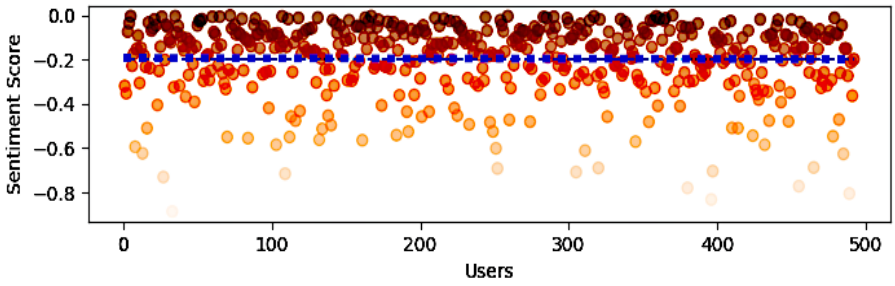

Fig. 4 Negative sentiment towards 'Trump'

The dense dotted line represents the average sentiment of all users who are negative towards the topic 'trump'.

5.2 Multi-Aspect Sentiment Analysis

Multi-aspect sentiment analysis is performed on a set of 10,000 users with most frequently used relevant keywords identified from the particular domain of interest. The count of keywords across all user profiles was computed to generate a list of top twenty-five keywords. It was refined once again and eleven keywords were chosen from the generated list for further analysis. These keywords that were identified for analysis relates to the United States (US) presidential election 2020 and was recommended by experts in the field. The proposed work considers the following keywords from among the recommended words: 'trump', 'biden', 'joebiden', 'realdonaldtrump', 'gop', 'maga', 'democrats', 'obama', 'rallies', 'foxnews', and 'fauci'. From the set of 10,000 users the users with no sentiments towards any of the topics taken as aspects were removed and remaining users were considered for further analysis. Clustering was carried out with the remaining 8095 users, initially based on two clusters, followed by four clusters to determine the cluster variance with respect to the number of clusters. The clusters were named as 'republicans' and 'democrats' in case of two clusters, and as 'extreme republicans', 'moderate republicans', 'extreme democrats' and 'moderate democrats' in case of four clusters.

As shown in Fig. 5(a), for the two clusters that were formed, 4605 users were classified as republican and 3490 as democrats. In the case of four clusters, as shown in Fig. 5(b), the results obtained shows 320 users in the extreme republican cluster, 3736 users in the moderate republicans cluster, 3624 users in moderate democrats cluster and 415 users in extreme democrats cluster.

Table 1. shows two clusters, the republicans and democrats, along with the variance computed for each keyword in both the clusters as mentioned in sub-section 3.3 of the methodology.

Re-clustering was done with four clusters and the effect of each keyword on clustering is as shown in Table 2, which reveals that the variance among the keywords is significantly lower for the four clusters when compared to two, resulting in well-defined clusters with more accurate sentiment prediction. Hence four clusters were used for

Table 1 Two-cluster variance

Keyword	Republicans	Democrats
trump	0.07527045063119472	0.08069756419821233
biden	0.10744352137425661	0.10692885030554354
rallies	0.14961832738459332	0.16242220446545205
democrats	0.1792316872206129	0.144455114837359
obama	0.17028611606491886	0.144713952759956
fauci	0.1581834747710887	0.1696663267281974
gop	0.11046998001187537	0.13241083783824564
maga	0.13424636260372025	0.12166526707427407
foxnews	0.13414535972054642	0.14572074847083566
realdonaldtrump	0.11684174405209016	0.13764916166050126
joebiden	0.1190163907394011	0.11909738587243356

Table 2 Four-cluster variance

Keyword	Extreme_Republicans	Moderate_Republicans	Extreme_Democrats	Moderate_Democrats
trump	0.041159650915479525	0.07216544602080319	0.04849170406941834	0.025991129389720865
biden	0.029078952623644985	0.11456754408167354	0.09680202633169523	0.044098486833727954
rallies	0.005894376398809523	0.20665859767623618	0.15941641870711928	0.015890527749999998
democrats	0.02125698288888889	0.1530412406726054	0.18048061431109358	0.0057423970000000015
obama	0.021011843333333335	0.15449071673941922	0.17361834140046906	0.03573801124999999
fauci	-	0.10121909291612387	0.1130102949779045	0.014740166666666662
gop	0.015931046666666664	0.13929858060331815	0.11149961500250127	0.03192487641865079
maga	0.02261088339616402	0.1348644777029969	0.13569128352157453	0.039391558576998054
foxnews	0.010950837666666666	0.16058714195048643	0.13155930971244506	0.014744995
realdonaldtrump	0.013343146750329939	0.14191311064352952	0.1134143458153826	0.01571681333728816
joebiden	0.023499343523809522	0.1289106863607285	0.11390494759144396	0.01642387702991454

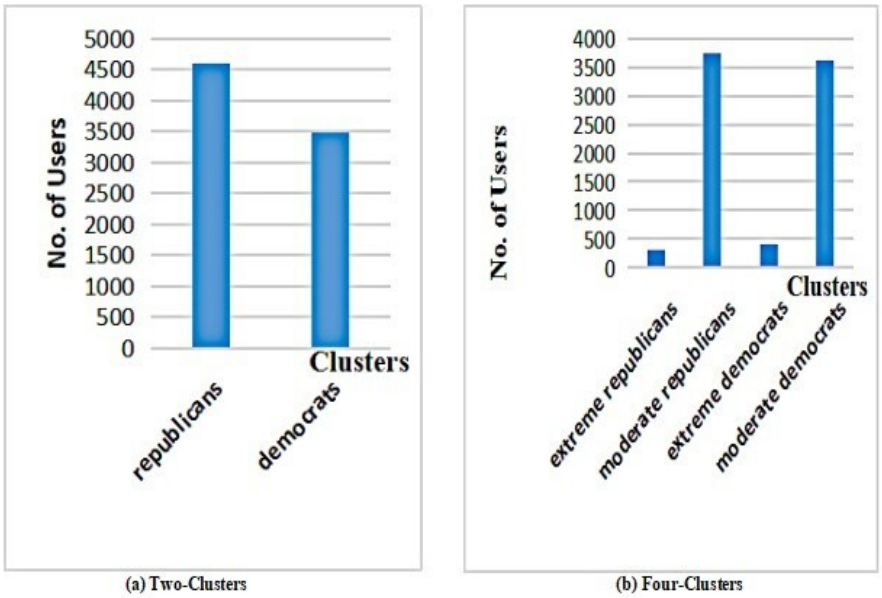

Fig. 5 Clusters for Classification

further analysis in this work.

Fig. 6. represents the similarity score against the sample user of each cluster and the number of users that belong to each cluster. The sentiment of the keyword 'trump' is predicted by removing it from the keyword list and recalculating the similarity score with respect to the sample users of all clusters. The variation in the number of users is shown in Table 3.

Keyword analysis was carried out for predicting the sentiment of an unmentioned

Table 3 User variation

Keyword	Users in Cluster 1	Users in Cluster 2	Users in Cluster 3	Users in Cluster 4
trump	320	3799	3549	427
biden	320	3801	3552	422
rallies	320	3833	3527	415
democrats	320	3821	3527	415
obama	320	3826	3531	418
fauci	320	3800	3561	414
gop	320	3773	3588	414
maga	320	3784	3574	417
foxnews	320	3777	3582	416
realdonaldtrump	320	3796	3565	414
joebiden	320	3753	3607	415

topic by selecting fourteen of the randomly chosen user profiles. Four users from the

Table 4 Predicted Vs. actual sentiment score

User	Assigned Cluster	Actual Sentiment	Predicted Sentiment	Confidence
user1	2	-0.14643405797101447	-0.034475696969696976	0.9945264403597461
user2	2	-0.10683372093023254	-0.11478616666666666	0.9948373996964752
user3	2	-0.17132936893203887	-0.14801571428571428	0.9957717644774652
user4	1	0.06251138211382114	0.12628024898373985	0.9989324739753187

Table 5 Predicted sentiment score

Keyword	User1	User2	User3	User4
trump	-0.14643405797101447	-0.10683372093023254	-0.17132936893203887	0.06251138211382114
biden	-0.5536800000000001	-0.1914730769230769	-0.12271171874999998	-0.052307692307692284
rallies	-0.4648	-	0.32115	0.1738666666666667
democrats	-	-0.4199	0.08499999999999999	0.008968750000000003
obama	0.128	0.3612	-	-0.4297500000000001
fauci	-	-	-0.026050000000000018	-
gop	-	-0.021788888888888924	-0.24704047619047614	-
maga	-0.7901	-0.56975	-0.5106	-
foxnews	-0.0577	-0.41035	-0.4404	-0.17215000000000003
realdonaldtrump	-0.04201500000000001	0.005113333333333318	0.3182	-
joebiden	-0.8728	-0.7184	-0.09738478260869564	-0.2235666666666667

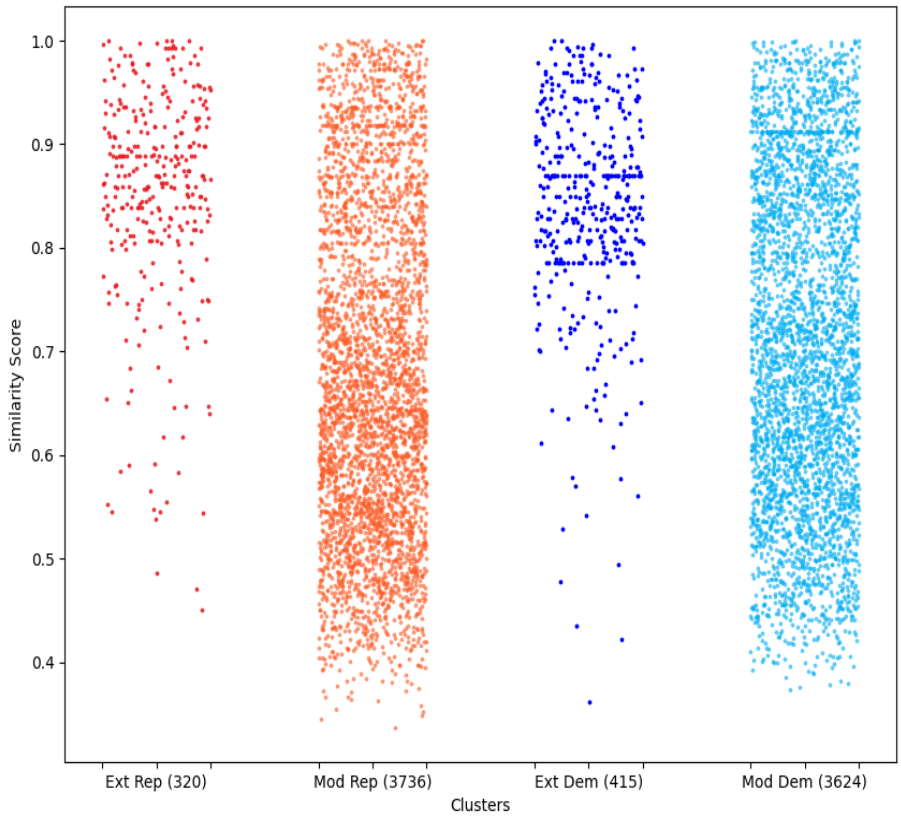

Fig. 6 Similarity score for clusters

datasets, not in any of the clusters, and with a minimum match of at least four keyword in the keyword list were chosen to analyze their sentiment towards the new topic. The user profile list is shown in Table 4, along with the predicted and actual sentiment score of users towards a particular keyword and the confidence level of each user in their respective clusters.

Actual sentiment represents the sentiment score before removal of the keyword for which the sentiment is to be predicted. Predicted sentiment in Table 4 is the sentiment score after excluding the keyword. Assigned cluster field shows the cluster to which each user belongs, after computing the similarity score with sample users of each cluster. In the case of neutral sentiment, the user can be assigned to any of the four clusters according to the current classification, since such a user have negligible or minimal impact on the decision making process. Confidence level is obtained by taking the aggregate sentiment score evaluated using Algorithm 2, of those users with similarity score greater than 0.9, which is the threshold value that ascertains the sentiment with 90% confidence level that the user belongs to the assigned cluster.

Users with more than 0.9 similarity score and have the missing keyword are identified and considered for predicting the sentiment score of each user with respect to the

missing keyword. The average sentiment score of the keywords for the identified users are then assigned as the predicted sentiment score, as shown in Table 5.

Fig. 7. shows the graphical representation of the count of users in each cluster with

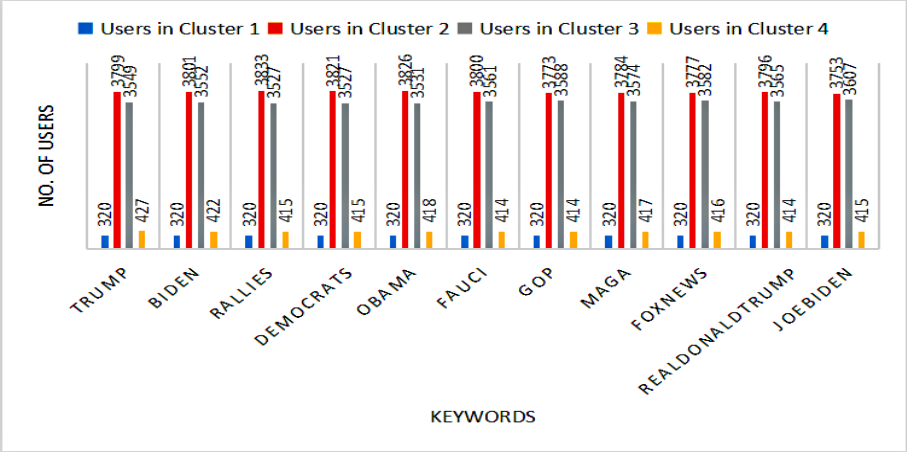

Fig. 7 Clusters for Classification

the expert-recommended keywords (those shown in the x-axis), when not considered to be present in the user profile.

Fig. 8. shows the graphical representation of the variation in the number of users in

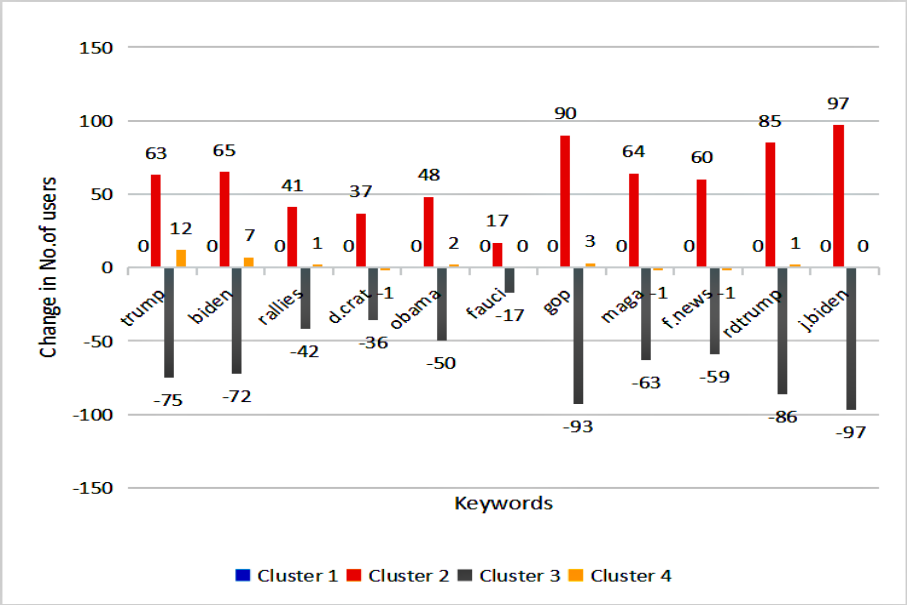

Fig. 8 Clusters for Classification

each cluster with the assumption that the keywords, as shown along the x-axis has not been considered as a part of the user profile.
The graph shown in Fig.8 illustrates the effect that happens when the keyword 'trump' is not considered for analysing the sentiment score. It is evident that there is a shift of 63 users from moderate democrats to moderate republicans and 12 users moved from extreme democrats to moderate republicans. Cluster 3 had lost 75 users and Cluster 1 remains unaffected.
The proposed methodology makes valuable contributions to various industrial and research areas for predicting the bias of unknown users compared to other existing works mentioned in [21][39]. Our work extracts keywords from tweets using the TextRank algorithm more efficiently compared to the text analysis and tokenization method adopted in[21]. The methodology proposed in [39] tries to learn user emotions for future predictions, but does not classify the user's shift with policy changes, which is addressed in our paper.
The above results clearly reveal that the polarity of an unknown user, not belonging to any of the clusters could be predicted using the proposed methodology. The methodology serves as a novel contribution in the realm of sentiment analysis that contributes to improve or assist brand/political parties to analyse the impact of their current decisions based on user sentiments and tailor their policies accordingly to forecast future plans. The work further goes on to predict the impact of removing certain keywords that can have on user sentiments regarding a topic and its effect on the standing of the political party as a whole.

6 Conclusion

In general, this paper focuses on various methods such as topic extraction and data pre-processing from twitter tweets, multi-aspect sentiment analysis, clustering users based on common topics of interest and predicting sentiment of users to unknown topics based on their tweet history. By introducing a multi-aspect sentiment analysis approach, we were able to predict user sentiments and provide valuable insights about the users attitude and behaviour related to the context of the topic being discussed. We also implemented keyword extraction using the NLTK and VADER sentiment libraries to enhance the overall prediction capability of the proposed methodology and classify users based on their similarity score. The methodology also predicts the sentiment of an unknown user with 90% confidence level for a specific topic of interest. The paper also describes the shift in the strength of users belonging to various clusters with nullified keyword, which can be utilized for formulating party manifestos and their real-time impact among the public.
The proposed analysis technique can be extrapolated to multiple domains such as brand monitoring, business forecasting, reputation management, trend analysis, customer service enhancement etc. Also, extending the current work to multiple languages enables to capture the sentiment of users independent of their language and potentially expand the user base for analysis.

Acknowledgement: Not Applicable

Statements and Declarations:

Funding and/or Conflicts of Interests/Competing Interests

Funding: The authors have no relevant financial or non-financial interests to disclose.

Conflict of Interest: The authors declare that they have no conflict of interest

Statement of Responsibility: All authors reviewed the manuscript and contributed equally to this work.

Data Availability Statement: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study. The datasets used were publicly available and have been referenced in the manuscript.

Code Availability: The experimental code is available on request.

Ethical Approval: Not Applicable.

Authors' contributions: All authors contributed equally to this work.

References

[1] D. Wehner, "Meta Reports First Quarter 2023 Results," pp. 1–10, March 2023. https://s21.q4cdn.com/399680738/files/doc_news/Meta-Reports-First-Quarter-2023.

[2] M. Toprak, C. Boldrini, A. Passarella, and M. Conti, "Harnessing the Power of Ego Network Layers for Link Prediction in Online Social Networks," IEEE Transactions on Computational Social Systems, vol.10, no. 1, pp. 48-60, March 2022. doi: 10.1109/TCSS.2022.3155946

[3] S. Kumar, M. Saini, M. Goel, and B. S. Panda, "Modeling information diffusion in online social networks using a modified forest-fire model," J. Intell. Inf. Syst., vol. 56, no. 2, pp. 355–377, Springer 2021, doi: 10.1007/s10844-020-00623-8.

[4] M. Salehan, D. J. Kim, and C. Koo, "A study of the effect of social trust, trust in social networking services, and sharing attitude, on two dimensions of personal information sharing behavior," The Journal of Supercomputing, vol. 74, no. 8, pp. 3596–3619, Springer Nature 2018, doi: 10.1007/s11227-016-1790-z.

[5] A. Crisci, V. Grasso, P. Nesi, G Pantaleo, I. Paoli, and I. Zaza, "Predicting TV program audience by using Twitter based metrics," Multimedia Tools and Applications, vol. 77, no. 3, pp. 12203–12232, Springer 2018.

[6] N. Thanh, E. Szczerbicki, and B. Trawi, "Collective intelligence in information systems," J. Intell. Inf. Syst., vol. 37, pp. 7113–7115, Springer 2019, doi: 10.3233/JIFS-179324.

[7] Z. Deng, M. Yan, J. Sang, and C. Xu, "Twitter is faster: Personalized Time-aware Video Recommendation from Twitter to YouTube," ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), vol. 11, no. 2, p. 31, 2015.

[8] R. Das, J. kamruzzaman, and G. Karmakar, " Opinion Formation in Online Social Networks: Exploiting Predisposition, Interaction, and Credibility," IEEE Transactions on Computational Social Systems, vol.6, no.3, pp. 554-566, 2019.

[9] D. Xue, S. Hirche, and M. Cao, "Opinion Behavior Analysis in Social Networks under the Influence of Coopetitive Media," IEEE Transactions on Network Science & Engineering, vol. 7, no. 3, pp. 961–974, 2020, doi: 10.1109/ TNSE. 2019.2894565.

[10] A. Ouertatani, G. Gasmi, and C. Latiri, "Parsing argued opinion structure in Twitter content," J. Intell. Inf. Syst., Springer Nature, September, 2020. https://doi.org/10.1007/s10844-020-00620-x.

[11] S. Brito, R. Luiz, C. Silva, P. Jorge, and L. Adeodato, "A Systematic Review of Predicting Elections Based on Social Media Data: Research Challenges and Future Directions," IEEE Transactions on Computational Social Systems, pp. 819–843, 2021.

[12] P. Stefanov, K. Darwish, A. Atanasov, and P. Nakov, "Predicting the topical stance and political leaning of media using tweets," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020.

[13] M. Ianni, E. Masciari, and G. Sperl, "A survey of Big Data dimensions vs Social Networks analysis," J. Intell. Inf. Syst., pp. 73–100, Springer 2021.

[14] C. Ahmed, A. Elkorany, and E. Elsayed, "Prediction of customer ' s perception in social networks by integrating sentiment analysis and machine learning," J. Intell. Inf. Syst., vol. 60, no. 3, pp. 829–851, Springer 2023, doi: 10.1007/ s10844-022-00756-y.

[15] N. Chouchani and M. Abed, "Online social network analysis: detection of communities of interest," J. Intell. Inf. Syst., Springer Nature, 2018.

[16] F. Nazir, M. A. Ghazanfar, M. Maqsood, and F. Aadil, "Social media signal detection using tweets volume, hashtag, and sentiment analysis," Multimedia Tools and Applications, Springer Nature 2018. https: //doi.org/ 10.1007/ s11042-018-6437-z

[17] M. Trupthi, S. Pabboju, and N. Gugulotu, " Deep Sentiment Extraction for Consumer Products Using NLP-Based Technique," Soft Computing & Signal Processing, pp.191-201, 2019 Springer Singapore. doi: 10.1007/978-981-13-3393-4.

[18] Z. Nasar, S. W. Jaffry, and M. K. Malik, "Textual Keyword Extraction and Summarization: State-of-the-art," Information Processing & Management, vol. 56, no. 6, Springer 2019, doi: 10.1016/j.ipm.2019.102088.

[19] K. Chakraborty, S. Bhattacharyya, S. Member, and R. Bag, "A Survey of Sentiment Analysis from Social Media Data," IEEE Transactions on Computational Social Systems, vol. PP, pp. 1–15, 2020, doi: 10.1109/TCSS.2019.2956957.

[20] K. P. Vidyashree and A. B. Rajendra, "An Improvised Sentiment Analysis Model on Twitter Data Using Stochastic Gradient Descent (SGD) Optimization Algorithm in Stochastic Gate Neural Network (SGNN)," SN Computer Science, vol. 4, no. 2, pp. 1– 11, 2023, doi: 10.1007/s42979-022-01607-x.

[21] M. Wongka and A. Angdresey, "Sentiment Analysis using Naive Bayes Algorithm of the Data Crawle: Twitter," In 2019 Fourth International Conference on Informatics and Computing (ICIC), pp. 1–5, 2019.

[22] Fagni, Tiziano, and Stefano Cresci, "Fine-Grained Prediction of Political Leaning on Social Media with Unsupervised Deep Learning," Journal of Artificial Intelligence Research, vol.73, pp. 633-672, 2022.

[23] S. Kayiki, "SenDemonNet: Sentiment Analysis for Demonetization Tweets using Heuristic Deep Neural Network," Multimedia Tools & Applications, vol. 81, no. 8, pp. 11341–11378, Springer 2022. doi: 10.1007/s11042-022-11929-w.

[24] Y. Zheng, Y Li, G Wang, Y. Chen, Q Xu, J. fan, and X. Cui, "A Novel Hybrid Algorithm for Feature Selection Based on Whale Optimization Algorithm," IEEE Access, vol. 7, pp. 14908–14923, 2019, doi: 10.1109/ACCESS.2018.2879848.

[25] M. Wankhade, A. C. S. Rao, and C. Kulkarni, "A survey on sentiment analysis methods, applications, and challenges," Artificial Intelligence Review, vol. 55, no. 7. Springer Netherlands, 2022. doi: 10.1007/s10462-022-10144-1.

[26] Q. You, S. Bhatia, and J. Luo, "A picture tells a thousand words - About you! User interest profiling from user-generated visual content", Signal Processing, vol. 124, pp.45–53, Elsevier, 2016. doi: 10.1016 j.sigpro. 2015.10.032.

[27] A. Ligthart, C. Catal, and B. Tekinerdogan, "Systematic reviews in sentiment analysis: a tertiary study," Artificial Intelligence Review, vol. 54, no. 7. Springer Netherlands 2021. doi: 10.1007/s10462-021-09973-3.

[28] P. Berka, "Sentiment analysis using rule-based and case-based reasoning," J. Intell. Inf. Syst., Springer Nature 2020. https://doi.org/10.1007/s10844-019-00591-8.

[29] S. Sedhai and A. Sun, "Semi-Supervised Spam Detection in Twitter Stream," IEEE Transactions on Computational Social Systems, vol. 5, no. 1, pp. 169–175, 2018, doi: 10.1109/TCSS.2017.2773581.

[30] S. M. Park and Y. G. Kim, "Root Cause Analysis Based on Relations Among Sentiment Words," Cognitive Computation, vol. 13, no. 4, pp. 903–918, Springer 2021, doi: 10.1007/s12559-021-09872-3.

[31] S. M. Nagarajan and U. D. Gandhi, "Classifying streaming of Twitter data based on sentiment analysis using hybridization," Neural Computing and Applications, vol. 31, no. 5, pp. 1425–1433, Springer 2019, doi: 10.1007/s00521-018-3476-3.

[32] N. Zainuddin, A. Selamat, and R. Ibrahim, "Hybrid sentiment classification on twitter aspect-based sentiment analysis," Applied Intelligence, vol. 48, no. 5, pp. 1218–1232, Springer 2018, doi: 10.1007/s10489-017-1098-6.

[33] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu, and A. Abusorrah, "Aspect-Based Sentiment Analysis: A Survey of Deep Learning Methods", IEEE Transactions on Computational Social Systems, vol. 7, no. 6, pp. 1358–1375, 2020, doi: 10.1109/TCSS. 2020.3033302.

[34] L. Luceri, T. Braun, and S. Giordano, "Analyzing and inferring human real-life behavior through online social networks with social influence deep learning," Applied Network Science, vol. 4, no. 1, Springer 2019, doi: 10.1007/s41109-019-0134-3.

[35] S. Stieglitz, M. Mirbabaie, B. Ross, and C. Neuberger, "Social media analytics – Challenges in topic discovery, data collection, and data preparation," International Journal of Information Management, vol. 39, no. December 2017, pp. 156–168, Elsevier 2018, doi: 10.1016/j.ijinfomgt.2017.12.002.

[36] A. R. Pathak, M. Pandey, and S. Rautaray, "Topic-level sentiment analysis of social media data using deep learning," Journal of Applied Soft Computing, vol. 108, p. 107440, Elsevier 2021, doi: 10.1016/j.asoc.2021.107440.

[37] F. Cena, L. Console, and F. Vernero, "How to Deal with Negative Preferences in Recommender Systems: a Theoretical Framework," J. Intell. Inf. Syst., pp. 23–47, Springer Nature 2022. https://doi.org/10.1007/s10844-022-00705-9

[38] L. M. De Campos, J. M. Fern, L. Redondo-exp, and J. F. Huete, "LDA-based term profiles for expert finding in a political," J. Intell. Inf. Syst., Springer Nature 2021. https://doi.org/10.1007/s10844-021-00636-x

[39] S. Abdi, J. Bagherzadeh, G. Gholami, and M. S. Tajbakhsh, "Using an auxiliary dataset to improve emotion estimation in users ' opinions," J. Intell. Inf. Syst., Springer Nature 2021. https://doi.org/10.1007/s10844-021-00643-y

[40] A. Petukhova and N. Fachada," TextCL: A Python package for NLP preprocessing tasks," SoftwareX, vol. 19, p. 101122, Elsevier 2022. doi: 10.1016/j.softx.2022.101122.

[41] A. Pramita, S. Rustad, G. F. Shidik, E. Noersasongko, A. Syukur, A. Affandy, and D.R.I.M Setiadi, "Review of automatic text summarization techniques & methods," J. King Saud Univ. – Computer and Information Science, vol. 34, no. 4, pp. 1029–1046, Elsevier 2022, doi: 10.1016/j.jksuci.2020.05.006.

[42] P. Sagar, S. Divakar, and Y. Pankaj, "A systematic review on page ranking algorithms," Int. J. Inf. Technol., Springer 2020, doi: 10.1007/s41870-020-00439-3.

[43] A. Chiche and B. Yitagesu, "Part of speech tagging: a systematic review of deep learning and machine learning approaches," J. Big Data, vol.9, Springer 2022, doi: 10.1186/s40537-022-00561-y.

[44] S. Elbagir and J. Yang, "Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment," Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 0958, 2019.

[45] L. Sun, J. Guo, and Y. Zhu, "A multi- aspect user- interest model based on sentiment analysis and uncertainty theory for recommender systems," Electronic Commerce Research, no. 0123456789, Springer 2018, doi: 10.1007/s10660-018-9319-6.

[46] L. G. Singh and S. R. Singh, "Empirical study of sentiment analysis tools and techniques on societal topics," J. Intell. Inf. Syst., Springer Nature 2020. https://doi.org/10.1007/s10844-020-00616-7.

[47] D. Zhao, X. Hu, S. Xiong, J. Tian, J. Xiang, J Zhou, and H. Li, "k-means clustering and kNN classification based on negative databases," Applied Soft Computing, vol. 110, p.107732, Elsevier 2021, doi: 10.1016/j.asoc.2021.107732.

[48] https://www.kaggle.com/datasets/manchunhui/us-election - 2020-tweets.