Clustering Algorithms in Sentiment Analysis Techniques in Social Media – A Rapid Literature Review

Vasile Daniel Păvăloaia
Accounting, Information Systems and Statistics Department, Alexandru Ioan Cuza University of Iasi, Iasi, Romania

***Abstract*—Based on the high dynamic of Sentiment Analysis (SA) topic among the latest publication landscape, the current review attempts to fill a research gap. Consequently, the paper elaborates on the most recent body of literature to extract and analyze the papers that elaborate on the clustering algorithms applied on social media datasets for performing SA. The current rapid review attempts to answer the research questions by analyzing a pool of 46 articles published in between Dec 2020 – Dec 2023. The manuscripts were thoroughly selected from Scopus (Sco) and WebOf-Science (WoS) databases and, after filtering the initial pool of 164 articles, the final results (46) were extracted and read in full.**

***Keywords*—Clustering algorithms; K-means; HAC; DBSCAN; sentiment analysis; natural language processing techniques; social media datasets; Twitter/X**

I. Introduction

The demand for seamless and simple contact between humans and machines has long been desired, since Turing test [1], and in the last years has grown significantly in a society that is getting more and more digitalized.

Social networks offer internet communities where users may simulate human social interactions. One of the most well-known [2] micro blogging sites is Twitter [3], [4], rebranded in X since July 2023. With 500 million daily tweets, 152 million active users per day, and 330 million active users per month [5], enables users to submit real-time, succinct messages (maximum 280 characters) on diverse social and personal topics. Every three days, more than one billion new Tweets are published on Twitter/X [6]. Researchers have extensively examined Twitter/X data to answer a variety of research problems, including detection of sentiments [7]. Twitter/X data analysis for sentiment/emotion/mood/opinion extraction is considered a difficult challenge in human computing. However, because tweets are limited to 280 characters, individuals tend to use casual language, which makes it difficult to understand the true mood behind tweets [8]. Also, due to the high number of total registered users (650 million) and instant notifications [9], [10] over a broad range of mobile equipment's, Twitter provide useful datasets for research to help better understand public behaviors, opinions, and sentiments [11]. This review is built on Social Media (SM) datasets where Twitter/X was found to be the most prominent for many reasons, such as: high data volume, public data availability, hashtags (relevant for clustering analysis), text-based posts, real-time analysis and abundant recent publications which are conducive to performing a comprehensive investigation.

Natural Language Processing (NLP), translates human language into machine language to facilitate interactions between humans and machines, was born out of this need. SA, a component of NLP, is employed in SM and other online environments to analyze and understand the emotions, opinions, and attitudes expressed in text. This endeavor can be done through a variety of methods, including Machine learning (ML) [12], [13], [14], NLP [15] [16], and text analytics [17], [18]. To evaluate whether a text's overall sentiment is positive, negative, or neutral is the ultimate goal of SA. The outcome is often a score or a label that describes the text's sentiment. Applications for this kind of analysis include SM monitoring, marketing, and customer service.

Overall, SA employs both ML and NLP methods. The models are used to estimate the sentiment of observed text after being trained on labeled data that comprises text and the appropriate sentiment (positive, negative, or neutral). While labeled data is often used in SA, there are several techniques that can be used to estimate the sentiment of text without prior labeled data, depending on the specific use case and available resources. It can be mentioned here Unsupervised SA [2], [19], Lexicon-based SA [9], [19], Transfer learning [20] and Active learning [21]. The goals of SA are accomplished through a number of phases. These actions may consist of data collection, data preprocessing, feature extraction, model training, model evaluation, model deployment. In this process, the most representative task is the choice of the algorithm which mainly depends on the goal and the resources of the project.

Algorithms are sets of instructions or rules that are followed in a specific order to accomplish a specific task or solve a specific problem. They are crucial in SA as they are used to automatically process [22], [23], [24] and analyze text data to determine the sentiment or emotional tone. Without algorithms, SA would be a manual and time-consuming process. Algorithms in SA can be used to classify text into positive, negative, or neutral sentiment categories, to generate a sentiment score or to create clusters based on similar patterns.

Due to the dynamics in this topic (SA), the current manuscript's aim is to analyze the literature in the period Dec 2020 – Dec 2023 for extracting the approach of the articles that deal with SA clustering algorithms applied on Twitter/X datasets. The investigation highlights the domains where the clustering algorithms are being employed, the most relevant

methods as well as the newly developed algorithms. The accuracy comparison will be displayed where this information is available. The contribution of such an investigation is relevant as it provides the best practice for anyone interested in matching the algorithm with the applicative sector/s by emphasizing the new discoveries. Although SA topic has abundant literature and many reviews exist, most of them refer to classification algorithms while those dealing with clustering have different approaches than the current paper. The current review is structured as follows: Section I presents the general background for the topic, Section II illustrates the selection process of manuscripts as well as the inclusion and exclusion criterions, Section III presents the results, by describing the SA algorithms with an emphasize on the clustering situation and Section IV presents the Discussion on results while Section V details the conclusions and future research paths.

II. Materials and Methods

A. Research Questions

The current study attempts to identify the most popular (1) clustering algorithms, the domains (2), and their performances (3) by quickly reviewing the relevant literature. After review-ing the literature analysis and using the key phrases to search the WoS and Sco databases, the intermediate findings show that there is a large amount of research on SA (mainly Twitter/X) dataset and the number of papers produced each year is growing exponentially. Despite that there are many reviews, only a small number of them address clustering algorithms as the majority focus on classification algorithms within Twitter/X datasets. In addition, no review was identified starting with 2020 that has a similar scope to the one in this research. As a result, the article aims to provide answers to the following research questions:

RQ1 – Which are the most employed clustering algorithms within the researches that perform SA using Twitter/X datasets, since 2020 to date, and what are their benefits and performances?;

RQ2 – What are the sectors of activity where the clustering algorithms were used within the selected literature?

In order to answer the research questions, WoS and Sco databases of article were used as the primary data source since they have the most pertinent papers that have been published in reputable journals which follow the peer review process.

B. Research Methodology

The decision to perform a rapid review was founded on the advantages it brings, namely because it offers accurate information while promoting the exploration of original perspectives [25, 26]. In order to build the current review, it was employed the truncation strategy for all the three phrases to include all of the expression's variants in the search [25]. As it can be seen in Fig. 1, the essential search combination was ("sentiment analys*" AND cluster* AND (“social media” OR “social network*”)) applied on Titles, Abstracts and Author’s Keywords while the publication years were set between 2020 and 2023. Upon completion the first phase 164 results (from WoS (61) and Sco(103)) were obtained, and two exclusion phases were further employed: Firstly, there were removed all document types except Articles and Reviews as most conference articles are the short and incomplete version of the Articles (1); the results which were not yet published (2) and manuscripts elaborated in other languages that English (3) have been eliminated as well. Further, the obtained references were merged into the same file and duplicates from the two databases were removed; Secondly, within the 77 intermediary results, 31 manuscripts were removed as the author/s did not employ clustering analysis but just listed a form of the keyword “cluster”. The remaining 46 manuscripts (44 articles and 2 reviews) were read in full and extracted the most relevant information required for answering the RQs. The findings are presented in the Results Section.

EndNote Online appropriately manages the references, removes duplicate sources, saves, organizes, and cites the list of references for a study. VOSViewer program was employed in this study as it is recognized to give cutting-edge methods for network layout and network clustering [27], to better extract the key domains and key associated relevant terms [28] from the list of selected articles (N=46). From Fig. 2 can be depicted the two clusters, one related to SA, SM and dataset (Twitter/X) and the second which is technological-related and contains the AI technologies and other clustering related concepts.

While cluster 1 - red contain SA keyword with the highest frequency of appearance (based on the association strengths), cluster 2 - green illustrate the most frequently used clustering algorithms found within the results (N=46), namely K-means, Hierarchical (mostly HAC) and DBSCAN.

The network map (see Fig. 2) is created based on the association strengths and highlights the clustering technologies used in the SA on Twitter/X dataset and contributes to answering the RQ1. The dominant keyword for all articles included in the study, as shown in Fig. 3, is SA, while closely related terms like Twitter/X, datasets, and SM come in second and third, respectively, with respect to their intensity. It can also be noticed that Covid-19 topic is a very common term within the selected papers, and this is expected as the analysis include manuscripts published within Dec 2020 – Dec 2023.

The clusters and density visualization are accessible in Fig. 3 where it can easily be observed the two formed clusters, highlighted with red and green background colors. The green cluster illustrates the highest weight (importance) on the association between the links for the keywords ML, NLP, clustering algorithms, topic modelling which validates the references included in the selection. The visual representation of clusters in Fig. 3 may be considered a confirmation for the proper selection of the articles, which is in line with the declared keywords: SA, Twitter/X dataset and clustering algorithms.

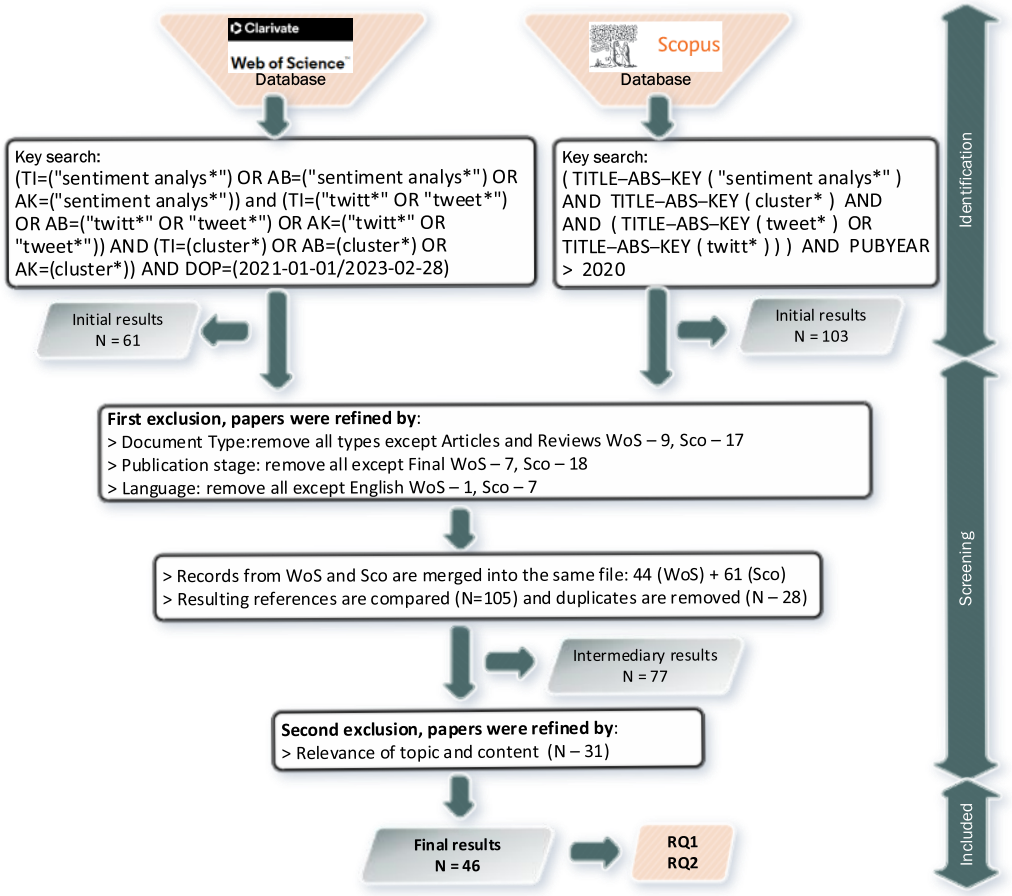

Fig. 1. The literature selection methodology based on PRISMA framework.

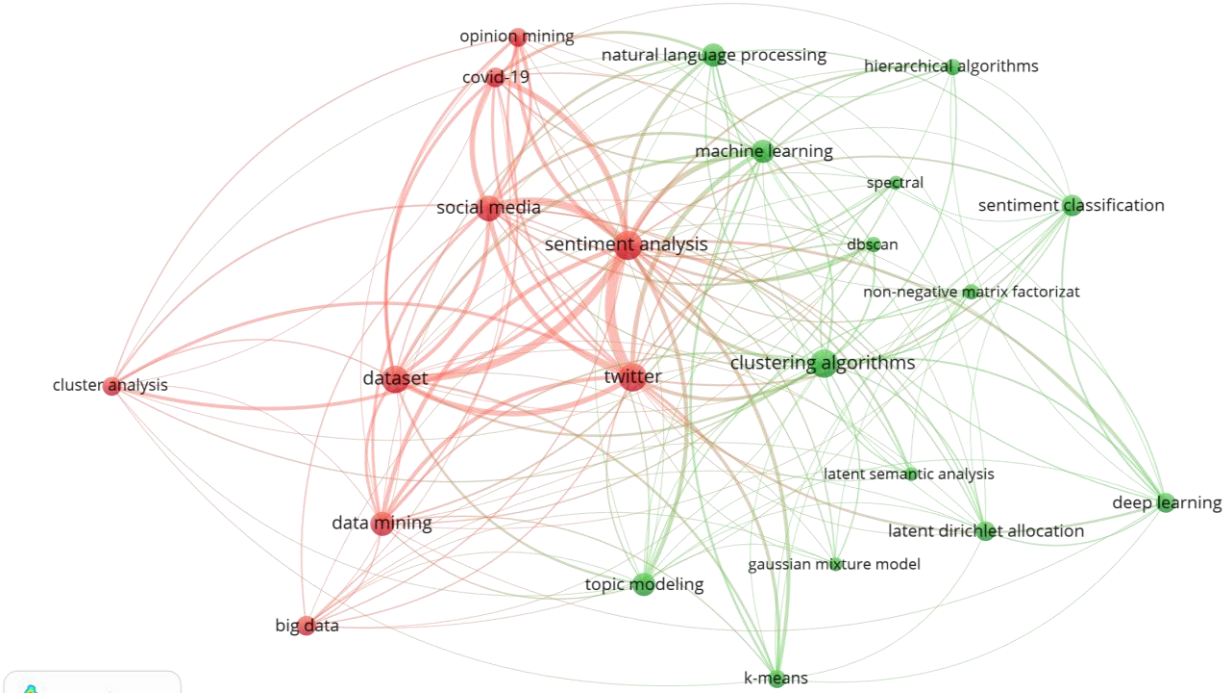

Fig. 2. The keywords network map.

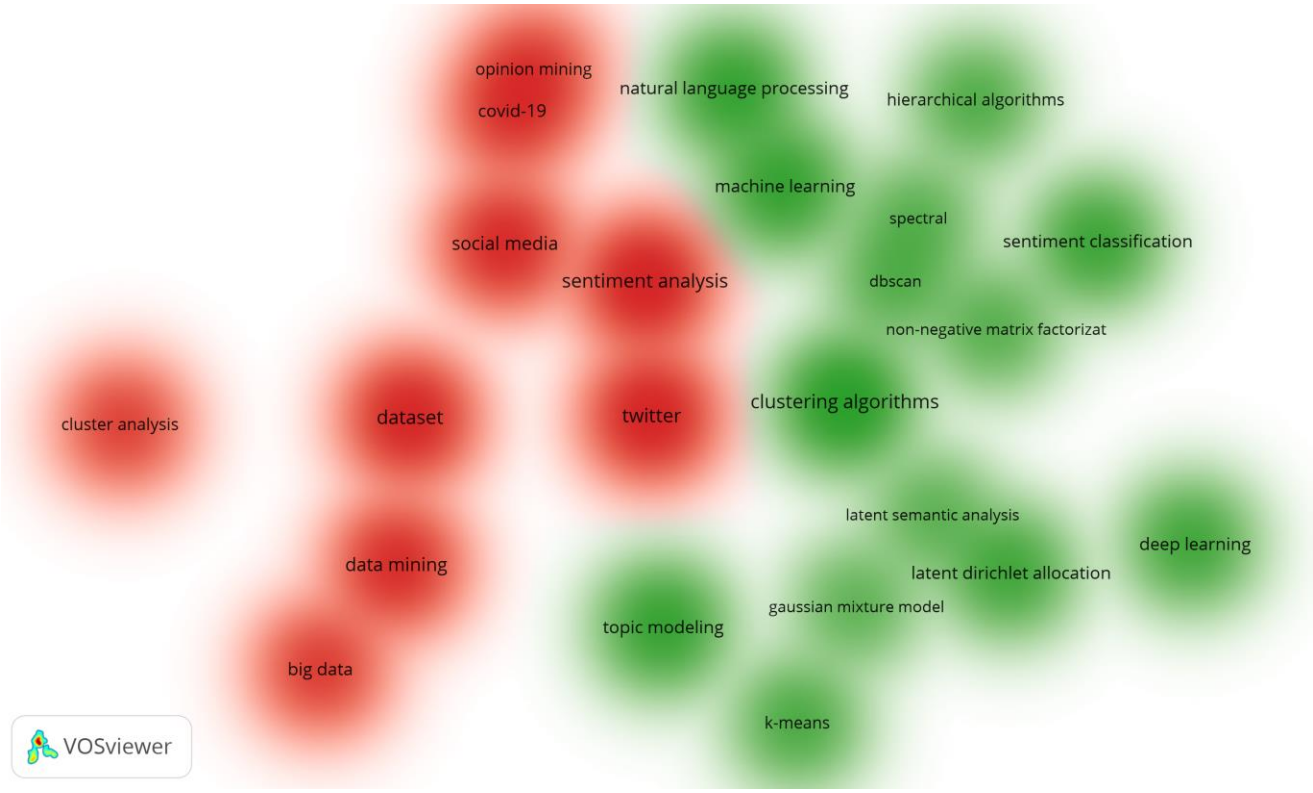

Fig. 3. Density visualization by clusters.

III. Results

Algorithms are used in SA either to process or analyze text data to determine the sentiment or emotional tone of a text (classification) or to group related entries in similar feature groups (clustering). In this sense, the above actions can include techniques such as NLP, ML, and Deep Learning (DL). SA algorithms can be employed for a variety of applications, such as SM monitoring, customer feedback analysis, and opinion mining analysis. This review's main aim is to discover the answer to the RQs. In this endeavor, the results extracted from the analyzed manuscripts and presented below.

A. The Most Employed Clustering Algorithms, Their Benefits and Performances (RQ1)

Algorithm wise, there are different options that can be used for performing the SA, such as:

- Rule-based methods [23], [29]. This method uses a set of predefined rules or dictionaries to classify the sentiment of text. It can be simple but less accurate.
- Lexicon-based methods [19]. This approach uses a lexicon (a collection of words and their associated sentiments) to classify the sentiment of text. It can be simple but less accurate as well.
- ML-based methods [30], [31], [19]. In this case, it trains a model using labeled data, where the labeled data contains text and the corresponding sentiment (positive, negative, neutral) and then the model is used to predict the sentiment of un-seen text. It can be more accurate than the above two methods.
- DL-based methods [32], [33], [34]. It's a type of ML-based methods, but it uses deep neural network architectures such as Long Short-Term Memory (LSTM) [35], [36], Convolutional Neural Network (CNN) [22] and Bidirectional Encoder Representations from Transformers (BERT) [37], [24], etc. It can achieve better results than traditional ML-based methods.

Algorithms are an essential component of any SA endeavors. The SA algorithms are used to classify the polarity of a text as positive, negative, or neutral, based on the sentiment expressed in the text. Among their various contributions to SA, the specialized literature mentions:

- Text classification [38], [31], [22]: SA algorithms are often used to classify text into different sentiment categories, such as positive, negative, or neutral. This is typically done using supervised learning algorithms, like Naive Bayes [2], [39], Support Vector Machine (SVM) [40], [41], Logistic Regression [41], [42] or DL algorithms like BERT, LSTM and CNN.
- Opinion mining [31], [33], [22]: SA algorithms can also be employed for extracting and understanding opinions and sentiments from text. This is typically done using NLP techniques, such as sentiment lexicons, sentiment ontologies, and sentiment-bearing terms.
- Emotion detection [43], [19]: SA algorithms identify emotions in text, such as happiness, sadness, anger, and surprise. Similarly with Opinion mining this endeavor is pursued by NLP techniques as well.

- Sentiment summarization [44], [35], [45]: SA algorithms summarize the overall sentiment of a text, such as a product review, a tweet, or a news article. Consequently, this action is performed by analyzing the sentiment of individual sentences, paragraphs, or even the whole documents.
- Opinion Spam detection [19, 38]: SA algorithms may be employed to detect fake or biased reviews or opinion from text. This can be achieved by comparing the sentiment of different text and detect any suspicious patterns.
- Aspect-based SA: SA algorithms are also used to extract and understand opinions and feelings about specific aspects of a text, such as a product, a service, or a person. This action is also pursued with the help of NLP techniques (sentiment lexicons and/or ontologies).
- Clustering text [46]: These algorithms are a type of unsupervised learning algorithms that are used in SA to group similar text samples together based on their sentiment. Clustering algorithms can be useful for tasks such as discovering latent themes/topics [15] within a dataset of text or grouping similar text samples together for further analysis. In line with the above concepts, [9] use Latent Dirichlet Allocation (LDA) and K-means to extract themes among the topics dis-cussed on Twitter/X posts in relation to natural disaster. The authors identify different themes with several emotions associated with it, to cluster people's reactions by time and location, during natural disasters. Positive and negative sentiments have both been subjected to text clustering by [17] in order to identify the main concerns that individuals have with regards to AI Ethical challenges. Other researchers [31] employ text clustering in a novel approach to extract agrarians' recommendations to boosting crop yields by informing farmers via SA on the most recent agricultural inputs. Overall, the specific algorithm used in SA depends on the problem, the resources available and the desired accuracy.

The main objective of clustering algorithms in SA is to group together [46], [17], [34] reviews or texts that express similar opinions, attitudes, or emotions. This can be useful in identifying common themes or topics in a set of reviews, understanding how different sentiments are distributed across a dataset, and identifying outliers or abnormal observations.

Clustering algorithms used in SA typically work by analyzing a set of features extracted from the text, such as word frequency [13], sentiment scores [40], [22] or other metrics [47]. These features are then used to calculate the similarity between different reviews, which is used to group similar reviews together into clusters.

Some examples of clustering algorithms employed in SA research endeavors include K-means, Hierarchical Clustering, Density-Based Spatial Clustering of Applications with Noise (DBSCAN), Expectation-Maximization (EM), Gaussian Mixture Model (GMM), Affinity Propagation (AP), Spectral Clustering and Self-Organizing Map (SOM). Some authors use these algorithms in different combinations [38], [48] to achieve the best results, depending on the nature of the data and the specific requirements of the task.

1) Topic modeling techniques in NLP: Topic modeling techniques are often used in conjunction with other elements, such as text preprocessing algorithms, feature extraction algorithms, and classification algorithms, to create a complete NLP system. The best choice of topic modelling option will depend on the specifics of the task, the size and quality of the dataset, and the computational resources available. Within the selected papers (N=46), the results illustrate that LDA has the highest (69%) utilization within the analyzed manuscripts, followed by LSA(20%) and NMF(11%).

Although the literature points out several topic modeling techniques in NLP, the majority of authors (69%) in the current pool of articles [54], [9] [55], [30] and many others have used LDA for topic modeling [29], [47]. In the light of the above, LDA is a generative probabilistic model of a corpus [54] used as a topic modeling algorithm that can discover the underlying themes in a collection of documents. LDA can automatically identify latent topics [15] in a set of reviews and offer a method to comprehend how various sentiments are distributed across a dataset.

2) SA clustering algorithms: Clustering algorithms used in SA are a set of unsupervised ML techniques that group similar texts (comments, reviews, posts) based on their predominant sentiment. Without the use of predetermined labels or categories, these algorithms are designed to identify patterns in the data and group related objects together. Although their majority is unsupervised, indicating they do not rely on labeled data, some can be semi-supervised, meaning they use a small amount of labeled data to guide the clustering process.

The selected body of literature is analyzed in line with the selection methodology presented in Fig. 1 and most used clustering algorithms, according with the network map in Fig. 3 used are K-means, Hierarchical (mostly HAC) and DBSCAN.

a) K-means: K-means is a popular unsupervised learning algorithm that is frequently used in SA for grouping purposes. The algorithm works by partitioning a dataset of texts into k clusters, where k is a user-specified parameter. Each cluster represents a group of texts that are similar to one another in terms of the features used to represent them. K-means' fundamental principle is to form spherical clusters, where each cluster is determined by the mean of points within the cluster. It starts with a random initialization of k centroids, one for each cluster. Each text is then assigned to the cluster with the closest centroid. The centroid of each cluster is then again calculated as the mean of all the points inside the cluster. This procedure is repeated until convergence, or a stopping condition is met. In SA, K-means algorithm takes a set of reviews as input, and for each review, a set of features are extracted such as word frequency, sentiment scores, or other metrics aiming at grouping similar reviews together into clusters. Ease of use and scalability are two of the main advantages of using K-means in SA. This fast algorithm can handle large datasets, and it is relatively easy to interpret the obtained results. It does,

however, have significant limitations, notably when working with datasets with different densities or non-spherical clusters. Additionally, it calls for previous knowledge of the number of clusters, which may not be known.

The Pension and Funds Administration's (AFP) goal is to shield the elderly population from the threat of poverty while enabling residents to save up money for their retirement. This study uses ML models to categorize and examine the sentiments of Twitter/X users (affiliates) utilizing the hashtag #afp. With the aid of the K-Means algorithm and the unsupervised learning technique, [13] were able to count the number of clusters using the elbow approach. Last but not least, despite the fact that data normalization was performed, the SA and the clusters created show that there is a very noticeable dispersion. This is one of the few research projects that display the employed precision index (IP) formula. In this research, the IP was used to gauge the effectiveness of clustering, where IP= $\sum c\ k$=1 n(ck)/n and ck stands for the number of data points necessary to achieve the proper clustering for cluster k and n is the overall number of data points. The performance of clustering improves with increasing accuracy index.

In order to decide which papers should be associated with which T topics, this work [54] examines two distinct clustering approaches. These techniques combine a genetic algorithm with a local convergence process and the K-Means clustering algorithm. The approach for assessing customer service interactions given in this article may be used to understand user satisfaction with this service and the major issues that consumers are concerned about. As K-means has the highest coverage among the algorithms extract from the literature, the following paragraphs will highlight the particularities of this algorithm for several sectors of activity.

Health & Medicine (Covid-19), customer preferences and society issue related research in-volving K-means clustering solutions

By crawling Twitter/X tweets, the authors [12] conduct a study employing cluster analysis on Covid-19 Outbreak Sentiments with K-means. The tweets are clustered into k groups using the K-means algorithm. By using t-Distributed Stochastic Neighbor Embedding (t-SNE) approach, the findings of each cluster are presented. Lexi-con-based SA has been employed to determine the sentiments of these clusters, and word clouds are used to examine the clusters' dominating subjects. The findings revealed nine clusters with diverse subjects, with the maximum positivity score of 83.25% and the lowest negative score of 16.75%. Word clouds are used to examine dominant subjects, and the outcomes of clustering are assessed using the 0.0070 Silhouette coefficient.

In their research [30], the authors emphasized the impact of COVID-19 and lockdowns on the agriculture sector and its related domains. The study performs SA and use K-means algorithms for clusters discovery in data. Among the findings, the most obvious one is that COVID-19 highly impacted the socioeconomic life of the people that work in agriculture as well as the agricultural sector itself.

The research in [4] contributes to the literature with the idea of amalgamation of extensive feature engineering and negation modelling with the unsupervised K-mean clustering approach for classification of large unlabeled Twitter/X corpus based on the tag #Lockdown. The novel framework of authors performs real-time labelling of Twitter/X datasets into three classes Positive, Negative, or Neutral for the textual based Twitter/X SA. The model was evaluated by inertia and silhouette score – two known evaluation metrics used to measure cluster quality – which show that the developed automatic labeling technique, applied in this context, achieves significant benefits.

In the manuscript [50], divide anomalous data into clusters using K-means to decide the anomaly type. The authors developed a framework for anomaly detection on two case studies (Corona Virus Tweet Dataset and Russia Ukraine Tweet Dataset) from Twitter/X, pre-processing of data, topic modelling, collection of the most frequently used words by applying topic modelling with LDA and Non-negative matrix factorization (NMF) [53].

Another study that performs a cluster analysis using K-means is done by [56] and analyses the polarity of Airlines Sentiment dataset to depict the customers' sentiments regarding the airline's services. The performance obtained on the dataset is displayed in Table I. On the same topic of analyzing customer's sentiments, this time to-wards products, K-means is used by [58] in the Twitter/X datasets and divides it into various types of clusters base on customer's emotions. The model has been instructed on how to compare various products and different sentiments. The k-means classification approach is applied while considering the cluster of customers who have a good opinion of the product. As a consequence, it offers a simple strategy for addressing a receptive audience directly and may reflect the growth and quality of a corporation.

TABLE I. MODELS' PERFORMANCES AND DOMAINS WHERE APPLIED

Dataset	Domain	Combined	Precision	F1 Score	Accuracy	Reference
Corona Virus	Medicine	No	0.8210	0.8093	0.7857	[42]
Russia Ukraine	Politics	No	0.5635	0.665	0.7019	
US Airlines	Customer review	No	0.840	0.730	0.890	[18]
Disaster	Natural disasters	no	0.957	0.963	0.997	[69,70]

To determine the ideal number of clusters every year, data visualization techniques such as term frequency, inverse document frequency, k-means clustering, and PCA were utilized by [55]. Authors used interpretation models to enable within-year (or within-cluster) comparisons after data visualization. The mate-rial inside each cluster for a particular year was examined using LDA topic modeling. To investigate the effect per cluster every year, Valence Aware Dictionary and Sentiment Reason-er SA were utilized. The average bot score per cluster each year was calculated using the Botometer automatic account check. The average number of likes and retweets per cluster was used by the authors to conduct correlations with other interesting outcome variables in order to gauge user involvement with the Dry January - a temporary alcohol abstinence campaign.

Climate change and natural disasters, involving K-means clustering solutions

Researchers [59] use correspondence analysis and the bisecting k-means algorithm to cluster tweets based on phrases that represent people's opinions. This reveals the fundamental determinants of discourses connected to carbon prices in Europe, the USA, South Africa, Canada, and Australia. The findings, which are presented in five clusters, demonstrate that views of the effects of taxes on people and companies, as well as faith in the government, are the key motivating factors for attitudes regarding carbon taxes.

Other authors [57] created and built a brand-new disaster intelligence system that automatically runs SA, automated K-Means, and AI-based translation to produce AI-driven insights for disaster strategists. The method provided crucial data for catastrophe planners or strategists, such as the natural disaster clusters that were most strongly correlated with negative feelings.

The authors [19] suggest a cuckoo search clustering technique for SA based on a roulette wheel. The proposed clustering method identifies optimal centroids from emotional datasets to determine document sentiment polarity. Tested on nine datasets, including Twitter/X and reviews of spam, its effectiveness surpasses K-means using a roulette wheel cuckoo search approach. According to their findings, the recommended strategies provide the best average precision, recall, and accuracy over 80% of the datasets.

Other researchers [9] apply K-means clustering algorithm to identify the underlying themes in the tweets for obtaining topic clusters on natural disasters dataset. During the cluster algorithm selection, authors compared it with HCA, and the results show that K-means performs better. They divided the data into three groups since it was determined that this would display the data most effectively. Cluster 0 denotes the phase of panic, during which people are distributing warning signs. Cluster 1 denotes the reactive stage, during which individuals discuss charity, wealth, and prayers. The larger of the two groups, Cluster 2, covers the stabilization period, where people were expressing gratitude for the care they had received. The results of the cluster performances reside in word clouds and charts, with no numeric values mentioned.

The research [60] identify typical daily morning congestion patterns for each route in the network to enhance the morning traffic prediction on a daily basis by clustering analysis using K-means on the reduced P-dimension matrix. The elbow approach is used to choose the optimal cluster size K. A vector of is created by a daily tweeting profile. Finally, to determine typical sleep-wake patterns observed in tweets, identical K-means clustering sets are utilized. In general, authors find that the earlier people go to bed, the more crowded the roads will be the following morning.

Society & Political views (Elections) and other subjects related research involving K-means clustering solutions

In order to differentiate political positions (left, center, right), authors in [61] applied the developed algorithm to obtain the scores of 882 ballots cast in the first stage of the convention (4 July to 29 September 2021). Then, they used k-means to identify three clusters containing right-wing, center, and left-wing positions. Our results may help us to better understand political behavior in constitutional processes.

Other authors in [62] used the k-means++ clustering method, a version of the k-means clustering algorithm that employs a smart centroid initialization strategy, to cluster the tweets (as vectors). The number of clustering iterations necessary and the regularity of the clusters are both influenced by the initial choice of centroids. The k-means++ clustering method was selected for its simplicity, effectiveness, and speed. The elbow curve method determined the optimal number of clusters for each dataset, forming topic-level clusters with similar information. Dense clusters - indicating widely shared information during an election period - were identified, with their geo-locations helping to map the topics geographically. The study analyzes user types and information patterns to observe how tweeting behavior related to the scheme changed during the election.

Authors in study [3] combine the Spider-Monkey Optimization (SMO) with K-means clustering, forming a hybrid approach (SMOK) to overcome early clustering termination issues seen in K-means. By utilizing K-means cluster outcomes to initialize the SMO population, SMOK enhances cluster quality, leading to faster convergence and superior results. It notably outperforms other algorithms like Particle-Swarm, Genetic algorithm, and Differential Evolution in computation time on Twitter/X datasets.

In conclusion, the K-Means clustering method was proven to be efficient in topics such as natural disasters [9], [57], climate change [59] the electoral campaigns [62], and politic opinion [52] and their analysis [61], traffic control (Yao & Qian, 2021), alcohol consumption [55], consumer behavior [56], medicine and Covid-19 [52], [12] analysis on the pension funds [13] domains and was analyzed the datasets from the following countries India [62], [30] USA (Pittsburg) [60], UK [55], Chile [61] and others with best results.

b) Hierarchical clustering: Hierarchical Clustering is an unsupervised learning method that is often used in SA to group similar texts. This is a tree-based clustering algorithm that builds a hierarchy of clusters by merging or splitting existing clusters. Starting with individual data points, it uses a bottom-up strategy to combine them into bigger clusters until every data point is in a single cluster. The two main approaches to hierarchical clustering are Agglomerative (bottom-up) and

Divisive (top-down) subcategories [38]. The divisive hierarchical method starts by iteratively dividing the dataset into multiple clusters. In the case of Hierarchical Agglomerative Clustering (HAC), each entry is initially clustered as a single point, which subsequently combines the smaller clusters into bigger clusters. The linking criteria of the method is used to assess cluster similarity. The following includes possible linking criteria: single linkage (SL); complete linkage (CL); average linkage (AL). The distance between two clusters in the HAC is determined by the lowest (SL), maximum (CL), or average (AL) distance between any two points in one cluster. HAC is used in SA to compile related reviews depending on their sentiment. A collection of reviews is provided as input to the algorithm, and for each review, a set of at-tributes such as word frequency, sentiment ratings, or other metrics are retrieved. The similarity between various reviews is then determined using these criteria, and lastly, comparable reviews are grouped together into clusters. The ability to handle non-spherical clusters and construct a hierarchical structure of clusters that can be used to understand how various attitudes are distributed across a dataset is one of the key benefits of utilizing HAC in sentiment research. Additionally, it enables the user to view the clustering outcomes as a dendrogram. However, when used on big datasets, it can be computationally costly and necessitates the selection of a link-age criterion.

The literature reviewed in the current project displays several innovative uses of this algorithm, either alone or in combination with other techniques, for performances improvements using Twitter/X datasets.

Topic modeling is essential to comprehend the tweets and group them into manage-able categories. As traditional methodologies are unable to effectively handle noise, high volume, dimensionality, and short text sparseness, some authors [10] rely on topic modelling approaches to cluster the tweets (or short text messages) to groups. Their original solution uses a hierarchical two-stage clustering technique and can address the problem of data sparsity in short text. Based on their statement, their technique performed better than other algorithms based on the results of standard datasets analysis.

In study [2], the authors propose a hierarchical method to extract the important words that people talk about during the coronavirus pandemic outbreak. Thus, the most used five words repeated in the people's posts on Twitter/X (using Coronavirus dataset) are included in each obtained cluster. Their findings demonstrate that the proposed model is capable of classifying and analyzing viewpoints presented in short text.

An original unsupervised ensemble/cooperative framework built on concept-based and HAC for Twitter/X SA is developed in [2]. The authors use four Twitter/X Dataset - Health Care Reform (HCR), Sentiment Strength (SS), Stanford Twitter/X Sentiment Test Set (STS-Test) and NewTweets (NT) - delegated to three popular HAC (SL, CL, and AL) combined with CBA in a serial ensemble manner to cluster tweets into two groups (positive and negative). Further, different feature representation methods are also examined and better performance of TF-IDF is revealed as compared to the Boolean method. The authors conclude by suggesting that CBA+CL ensemble can be the best choice among the selected clustering algorithms. According to the authors, their proposed framework is original as it has never been investigated before.

The research in [48] developed an original and simple clustering technique known as YAC2 and in study [38] extends its efficacy using three Twitter/X datasets. The technique of YAC2 is comparable to the divisive hierarchical clustering method, which divides a single cluster repeatedly into other clusters until no further clustering is possible. The efficacy of YAC2 has been demonstrated in study [48] by comparing its performance with well-established clustering algorithms (K-means, DBSCAN) on several datasets. The advantages of YAC2 include low theoretical complexity, handling of heterogeneous data, dynamic generation of cluster splits and proven high performances in comparison with DBSCAN and Spectral clustering algorithms.

c) Density-Based Spatial Clustering of Applications with Noise (DBSCAN): DBSCAN is an unsupervised learning algorithm that is often used in SA to group similar texts or reviews based on their sentiment. It is a density-based clustering algorithm that groups together data points that are closely packed together. DBSCAN algorithm is based on the idea of density reachability, which means that a point p is density-reachable from a point q if there exists a set of points which are all mutually density reachable from q and p. The algorithm defines two types of points: core points and non-core points. A core point is a point that has at least a minimum number of points (MinPts) within a distance ε (eps) from it. A non-core point is a point that is not a core point but is density-reachable from a core point. In SA, DBSCAN is used to group similar reviews together based on their predominant sentiment. The algorithm takes a set of reviews as input, and for each review, a set of features are extracted such as word frequency, sentiment scores, or other metrics. These features are then used to calculate the similarity between different reviews, which is used to group similar reviews together into clusters.

One of the main advantages of using DBSCAN in SA is that it can handle datasets with varying [63], and it does not require the number of clusters to be specified in advance (like in the case of K-means). It can also identify clusters of reviews that have similar sentiments and are close together in the feature space. However, it can be sensitive to the choice of parameters eps and MinPts and it doesn't perform well with high-dimensional data.

The research projects where this clustering algorithm was used are elaborated by [64] where authors propose a new methodology involving DBSCAN that had been applied to 7,014 tweets to identify regions of consumers sharing content about food trends. Grid maps were employed to investigate sub-regional variations and SA was utilized to address the attitude of their social representations. The study shows that the DBSCAN and SA-based technique is a legitimate research tool that may be used to identify communities with significantly diverse socio-psychological processes.

A study that applies cluster analysis with the DBSCAN algorithm is [65]. The manuscript assesses the spatial distribution of SM activities, aiming to define the concept of a

"district" through the geographical proximity of geotagged photos and texts on Instagram and Twitter/X. By setting the DBSCAN parameters to a minimum of five points and a threshold distance of 300 meters, they categorized SM posts as core (within a district), border (district edge), or outliers. DBSCAN's resistance to noise and flexibility with various cluster shapes made it ideal for this study, which examined the perception of city images in Poland's Tri-City Region using both "big data" and "small data" approaches, focusing on imageability and Lynchian features through SM analytics.

d) Other clustering methods: By applying a variety of cutting edge techniques, SA and the identification of significant users in social networks are enhanced in this research [32]. The tweets are grouped into topics using weighted partition around medoids (WPAM). Instead of using preset k values, an artificial cooperative search (ACS) is used to optimize the k values of WPAM. Outlier is nearly completely avoided in WPAM due to the dynamic selection of k values. As a result, it groups the tweets by subject using dynamic clustering (DC). After the dynamic clusters have been created, Stanford NLP is used to extract the subjects from each cluster.

The proposed automatic learning using CA-SVM based SA model reads the Twitter/X dataset [40]. The characteristics were then extracted from them in order to produce a collection of words. The tweets are grouped based on the phrases using TGS-K means clustering, which calculates Euclidean distance based on many variables, including semantic sentiment score (SSS), gazetteer and symbolic sentiment support (GSSS), and topical sentiment score (TSS). In comparison to the current works, the proposed model has a sentiment score of 92.05% and an accuracy score of 92.48%.

The Louvain Community Detection Algorithm (LCDA) was used by [11] to find semantic clusters. For topic modeling and semantic network clustering, the study employed the LDA method and the Louvain algorithm, respectively. The modularity score, which measures how well nodes are assigned to clusters, is maximized for each cluster using this method. The LDA approach unearths six themes, including veganism, food waste, organic food consumption, sustainable travel, sustainable transportation, and sustainable energy use. While the Louvain algorithm identifies four clusters: responsible consumption, energy consumption, lifestyle and climate change, and renewable energy. The Louvain method was also used to discover semantic clusters of latent issues since the study's goal is to find the themes and subjects linked to sustainable consumption. The study offers a novel viewpoint on several linked issues of sustainable consumption that help to sustainably level world consumption.

Due to its limited size, lack of organization, misspellings, use of slang and abbreviations, SA performed on Twitter/X datasets can prove to be a difficult task. To ease this process, Tweet Analyzing Model for Cluster Set Optimization with Unique Identifier Tagging (TAM-CSO-UIT) was developed by authors [66] utilizing prospects to assess the mood of tweets downloaded from Twitter/X. The suggested model TAM-CSO-UIT correctly analyzes and categorizes the tweets, according with the author's statements and the results reported.

Five computer nodes make up the Hadoop cluster in the study [22], namely one master node and four slave nodes. The authors have established ten evaluation measures to evaluate the experimental findings. Additionally, they executed their solution within the Hadoop cluster to avoid a lengthy execution time from their hybrid's developed Fuzzy Deep Learning Classifier (FDLC). To show the potency of their proposed classifier, an experimental comparison between our FDLC and some other ideas from the literature is conducted. The empirical findings shown that the proposed FDLC outperforms existing classifiers in terms of complexity, convergence, stability, true positive rate, true negative rate, false positive rate, error rate, accuracy, classification rate, kappa statistic, F1-score, and time consumption.

B. The Relevant Sectors of Activity where the Clustering Algorithms were Used (RQ2)

The current Section provides the answer to the second research question (RQ2). In the analyzed literature published in between 2020-2023, K-means, Hierarchical Clustering and DBSCAN proved to be the most used clustering algorithms. The references reveal that the algorithms are applicable in various sectors of activity, as well as in various situations: used alone or in combination with others. In order to provide the answer to RQ2, the manuscripts included in the pool of results were filtered based on the identified domain. To automatically extract the topic/sector of activity, Monkeylearn [67] tool proved to be very efficient. Based on this tool analysis, a new column was added to the database of papers, displaying the domain for each line. Therefore, for each paper in the database, the sector of activity was extracted by Monkeylearn from Title and Abstract variables, using a pretrained model.

1) Healthcare and medicine: Upon analysis 39% of the papers belong to the Health & Medicine sector of activity. As the review is based on the manuscripts published between 2020 and 2023, much research involving COVID-19 subject was performed. The following paragraphs display the insights extracted from the manuscripts included in the selection.

Several studies approach different hybrid clustering algorithms [47], [51], [68] while others develop new algorithms [36] for the purpose of grouping twits on similar topics related to COVID-19. Consequently, in order to understand city-level differences in emotions regarding COVID-19 vaccine-related subjects in the three biggest South African cities, [47] employ clustered geo-tagged Twitter postings. The study's findings demonstrated that clustered geo-tagged Twitter postings may be utilized to analyze the dynamics of emotions more effectively toward local discussions about infectious illnesses as COVID-19, malaria, or monkeypox. This can offer additional city-level data to health policy planners and decision-makers in planning and making decisions on vaccine reluctance for upcoming epidemics.

The authors [68] employ Uniform Manifold Approximation and Projection (UMAP) and Hierarchical Density-Based Spatial Clustering of Ap-plications with Noise (HDBSCAN). They combine the techniques to undertake a grid search for identifying the clustering model with the highest relative validity score representative for COVID-19. A total of 2666 hashtags

extracted from Twitter dataset related to COVID-19's prevention strategies and vaccination were grouped into 20 topics for the two main clusters including rural and urban users. The study developed by Liu reveals how clustering algorithms used on Twitter may help with the spatially targeted deployment of epidemic prevention and management activities.

Long COVID syndrome was first described as a set of fuzzy symptoms that persisted following COVID-19 recovery using patient-created vocabulary. According to the SA performed by the authors [49], opinions about the long COVID syndrome can be equally split between positive (19.90%) and negative (18.39%). Similarly, the study performed by [33] emphasizes how the Indian government's progressive unlocks and lockdowns of various areas during the Corona outburst were perceived by the general population.

According to [69], influential users are considerably more important to be examined as they can provide valuable knowledge within the tweets concerning opinions related to COVID-19 vaccines. Thus, the findings enabled researchers to thoroughly examine the ego networks of the three user clusters: pro-vaxxers, neutrals, and anti-vaxxers.

2) Climate and natural disasters: With regards to this domain, the analyzed authors approach subtopics related to natural disasters. Consequently, [57] created and built a brand-new disaster intelligence system that automatically runs SA, automated K-Means, and AI-based translation to produce AI-driven insights for disaster strategies. Others [19] suggest a cuckoo search clustering technique for SA based on a roulette wheel to extract the best cluster centroids from the emotional dataset's content. Other researchers [9] identify the underlying themes in the tweets for obtaining topic clusters on natural disasters dataset.

AI-Social Disaster is developed by [70] and represents a decision support system (DSS) for identifying and analyzing natural disasters like earthquakes, floods, and bushfires. The approach gives crucial details for catastrophe planners or strategists, such as which natural disaster clusters were linked to the most negative opinions. Further, [57] proposes another DSS that collects Tweets on natural disasters in 110 supported languages using a live Twitter feed. The aim is to produce AI-driven insights for catastrophe planners, the system automatically carries out AI-based translation, SA, and automated K-Means algorithm. There is proof that being exposed to weather hazards has a negative influence on people's physical and mental health, especially in places affected by heat islands and climate change. The study in [71] reveals how Twitter data may be used to measure urban heat stress in real time and serve as a quick signal of times when people are feeling more uncomfortable due to the heat and are more likely to be dissatisfied with the weather.

3) Environment: In their research, [11] identify four clusters: responsible consumption, energy usage, renewable energy, and lifestyle and climate change. The SA's findings indicate that users are more likely to have positive emotions than negative ones, and they offer a new angle on a number of interrelated issues of sustainable consumption that help to stabilize global consumption which exerts a high impact on Environment.

The study in [72] findings provide information on ways to improve knowledge sharing to improve carbon-neutral information sharing which provides policy and social implications for tackling environmental issues by analyzing social patterns on Twitter. Although carbon taxes are an efficient emission reduction strategy that benefits the environment, it is unpopular, and it is unclear why. This study [59] examines a sample of data from Twitter to identify the driving forces behind discourses about carbon prices in response to the scarcity of timely updates on people's perspectives of the relevant topic.

The research in [60], identify typical daily morning congestion patterns for each route in the network to enhance the morning traffic prediction and decrease the daily air pollution by clustering.

The goal of the author's [73] exploratory study is to locate, group, and assign an emotional value to tweets that contain the phrases "university" and "sustainable" in Spain relative to the rest of the globe. The findings highlight important aspects of the environment, research, and innovation via the lens of universities' contributions to local communities. They also offer an entrepreneurial perspective and highlight how academic knowledge is really used in the workplace.

4) Society and political views (Elections): The majority of studies include different tool developments and analyses performed during election campaigns. Consequently, a study demonstrates that during an election campaign, only the information that has been extensively disseminated is in the heart of the densest clusters. The geographic distribution of these clusters helps to group various topics together. Thus, [62] examines, in India, the types of users and information patterns to ascertain how the scheme-related tweeting patterns changed during the course of elections. The study reveals that a significant number of government-related tweets are generated during the voting periods and election length. The location of the voting phase, however, has no relevance to it. In future voting stages, the positive news outweighs the negative tweets and complaints that were generated in the initial voting phase. In order to gauge the emotional impact of the messages released by various Spanish Newspapers, NLP techniques and ML algorithms are used on this research [74] to discover the predominant topics linked to the elections as well as to highlight the candidates and political parties. The findings show the degree of attention given by the media to the regional election debates and campaign activities in Madrid.

The study in [61] aims to differentiate political positions (left, center, right) in Chile by developing an algorithm to obtain the scores of 882 ballots cast in the first stage of a convention. The authors employ k-means to identify three clusters containing right-wing, center, and left-wing positions and the results may prove to be efficient in the better understanding of political behavior within the constitutional processes.

In relation to Society, a study developed by [75] reveals that Twitter data offers a distinctive and practical source of

information for the analysis of significant civic movements, such as large-scale protests across numerous European nations. Additionally, such an approach might highlight significant spatiotemporal and emotional trends, which may also help to comprehend how protests escalate through space and time. Moreover, the inhabitants of a region may presently convey their own experiences with warm weather as well as their sentiments about it on SM. The public mood and health of an area may be reflected in the geotagged, time-stamped, and easily available SM databases, according to a recent study published by [71]. Further, research on the emotional data may be done using the Roulette wheel selection based cuckoo search clustering method [19]. The approach created by [19], [17] and [5] prove to have important and practical implications for creating a system that can produce accurate remarks on any societal issue with massive impact on the inhabitants.

IV. Discussion

SA algorithms have been advancing rapidly in recent years, thanks to breakthroughs in ML and NLP research. Within this section, several discussions shall be conducted on the key advancements encountered in SA algorithms. State-of-the-art performance in many NLP tasks, including SA, has been attained by DL models like BERT, Generative Pre-trained Transformer 3 (GPT-3), and XLNet. GPT-3, developed by OpenAI, can produce human-like prose on a variety of themes and was trained on a varied collection of online content. In 2019, researchers at Google AI created XLNet, another cutting-edge language model for NLP activities built on the transformer architecture, like BERT and GPT-2. Unlike GPT-3, XLNet uses an autoregressive language modeling approach, to predict each token in a sequence based on all the tokens that come before it. In the light of this information, XLNet captures complex dependencies and interactions between the tokens in a sequence, leading to higher performance on a wide range of NLP tasks, including SA. These new DL models (BERT, GPT-3, and XLNet) are able to recognize more sophisticated and subtle expressions of sentiment as well as the context and meaning of each word inside a phrase. Another advancement is the use of transfer learning, where a pre-trained model is fi-ne-tuned on a specific task, this allows the model to learn task-specific features while still retaining the general-purpose understanding of language learned during pre-training. This strategy can lead to enhancements in the SA model performance, particularly when training is scarce. Although the above technologies have made significant advancements in NLP and AI, they are still far from perfect and have limitations and challenges that need to be addressed. The existing AI technologies have been built over many years and have been re-fined and improved through countless iterations and experiments. Therefore, it is unlikely that they will be replaced overnight by new technologies, as there is a significant amount of knowledge and expertise that has gone into their development.

In the light of the above, it is still relevant to rely on the current technology and therefore, this review contributes to the domain. The recent research has focused on the existing technology to develop more robust models that can handle noise and outliers in the data, like sarcasm, irony, and emojis which can often be misleading, and also models that can handle multiple languages and cross-lingual SA analysis. Overall, in the light of the above mentioned advancements, it becomes obvious that the field of SA is rapidly advancing. The discoveries revealed within this review have led to more accurate and effective SA models, which can provide valuable insights into customer opinions, feedback, and attitudes, and support decision-making in a variety of industries, including marketing, healthcare, and finance.

A shortcoming is that SA models trained on general-purpose datasets may not per-form well on data from niched domains like product evaluations, or medical records. In order to solve this problem, domain-specific SA models have been created. These models are trained using domain-specific datasets, which improves accuracy and performance.

TABLE II. Comparative Analysis of Clustering Algorithms

Algorithm	Strengths	Weaknesses	Reference
K-Means	Easy to implement; Fast and scalable; Well suited for finding spherical clusters; Inductive learning.	Does not handle non-spherical clusters; The number of clusters should be specifield.	[77], [78], [76]
Hierarchical Clustering	Can handle non-spherical clusters and varying cluster sizes; Suitable for complex data structures; Transductive learning.	Can be computationally expensive for large datasets; Sensitive to the presence of noise and outliers in the data; No guarantee of optimality.	[77], [79], [80], [76]
DBSCAN	Automatically determines number of clusters; Robust to noise; Computationally efficient, suitable for large datasets; Transductive learning.	Sensitive to parameter choices; May not perform well in high-dimensional data; No probabilistic framework.	[79], [15], [14], [76]
Gaussian Mixture Model (GMM)	Ability to handle overlapping clusters; Good for density estimation; Inductive learning.	Sensitive to initial conditions; Does not scale well to large datasets.	[14], [76]

In line with the “traditional” clustering algorithms displayed by the papers included in the current review, a comparative analysis was performed and according to the literature investigation, clustering algorithms are used in conjunction with other algorithms, such as feature extraction algorithms and classification algorithms, to create a complete SA framework. The comparison analysis included in Table II is based on the aspects identified in the literature review (references are included in the last column). The analysis presents the strengths and weaknesses of each algorithm included in the current comparison. Among them, Hierarchical and DBSCAN offer, according with [76], transductive learning (the algorithm learns from a subset of the data and applies that knowledge to the entire dataset) while K-means and GMM support inductive learning

(the algorithm learns from the entire dataset and makes predictions on new, unseen data).

From the total of 46 manuscripts investigated, the authors of 19 papers opted for performing clustering analysis using K-means, 16 authors employed Hierarchical Clustering or other variant of Hierarchical algorithms, six used DBSCAN and the remaining papers (5) rely on using other algorithms (GMM, Spectral and others). Moreover, some authors combined the clustering algorithms with other techniques such as LDA, NMF and BERT [52] to detect anomalies and develop the clusters of most frequent words, while displaying the results using a word clouds and other visualization methods. The results show that applying K-means in a framework [52] can enhance the analysis in the considered dataset. An original development is revealed in a study of [57] that employed Automated K-Means clustering on Mobile Apps for the first time to uncover hidden knowledge, patterns, similarities, and differences contained among various types of catastrophe tweets.

Regarding HAC, [2] compares the performances of simple use of three agglomerative HAC (SL, CL, AL) and their combination with the concept-based methods. The results state that algorithms are encountering higher precision when used in combination with other algorithms or techniques in the form of frameworks. Based on the results from Table II, author personal perception and other source [76], the discussion regarding which algorithm to use is subject to the specific characteristics of the dataset and clustering task. Therefore, the observations in this research conduct the following insights: (1) K-means can handle well-defined clusters, but it may not be the best choice for large datasets. As the number of data points increases, the computational complexity of K-means also increases; (2) Hierarchical Clustering is suitable for complex datasets with non-spherical clusters and unknown number of clusters; (3) DBSCAN performs well in datasets with arbitrary shaped clusters and varying densities. When making a decision regarding the clustering algorithm to use, it is important to consider the strengths and weaknesses of each algorithm before selecting the most appropriate one.

V. Conclusions

To conclude, the algorithms play a critical role in SA, allowing for the automatic analysis of large volumes of textual data and providing valuable insights into customer opinions, feedback, and attitudes towards products, services and other topics extracted from SM environment (specifically from Twitter/X dataset on this research).

Considering the topic addressed, the results reveal that in the analysis period of Dec 2020 to Dec 2023 undertaken by this research, the most numerous articles treat different Coronavirus topic with subjects ranging from people's fears regarding Covid-19 [5],[11] [12] to their sentiments expressed towards different vaccination campaigns [47]. Further, the selection of studies based on the criteria formulated and displayed in Fig. 1, proved to employ "traditional" clustering algorithms, to develop new ones, as well as to use of different hybrid combinations between "traditional" and newly developed ones. None of the manuscripts included in the study refer to the state of the art DL models such as GPT-3 and XLNet while very few reference BERT. Thus, in a future study, the author intends to extend the current study by analyzing the impact of these modern models on SA techniques over a longer span of time and analyzing more SM channels. One limit of the study is that it does not cover the ethical aspects related to the use of AI and the algorithms in analyzing people's sentiments. This constitutes another future research path that will be fulfilled in the upcoming studies.

References

[1] Abayomi-Alli, A., Abayomi-Alli, O., Misra, S., Fernandez-Sanz, L. (2022). Study of the Yahoo-Yahoo Hash-Tag Tweets Using Sentiment Analysis and Opinion Mining Algorithms. Information ,13(3).

[2] Abuzayed, A., & Al-Khalifa, H. (2021). BERT for Arabic topic modeling: An experimental study on BERTopic technique. Procedia computer science, 189, 191-194.

[3] Ahmed, M. H., Tiun, S., Omar, N., & Sani, N. S. (2023). Short Text Clustering Algorithms, Application and Challenges: A Survey. Applied Sciences (Switzerland), 13(1).

[4] Aldaz, C. E. B., Duran-Rodas, D., & Hamón, L. A. S. (2021). What is the public opinion about universities and sustainability? A social media analysis among 'Spain' and across the world. International Journal of Innovation and Sustainable Development, 15(4), 438-457.

[5] Alhazmi, H. N. (2022). Text Mining in Online Social Networks: A Systematic Review [Review]. International Journal of Computer Science and Network Security, 22(3), 396-404.

[6] Asghar, M. Z., Khan, A., Bibi, A., Kundi, F. M., & Ahmad, H. (2017). Sentence-level emotion detection framework using rule-based classification. Cognitive Computation, 9(6), 868-894.

[7] Awoyemi, T., Ebili, U., Olusanya, A., Ogunniyi, K. E., & Adejumo, A. V. (2022). Twitter Sentiment Analysis of Long COVID Syndrome . Cureus Journal of Medical Science, 14(6), 13, Article e25901.

[8] Ayo, F. E., Folorunso, O., Ibharalu, F. T., Osinuga, I. A., & Abayomi-Alli, A. (2021). A probabilistic clustering model for hate speech classification in twitter. Expert Systems with Applications, 173.

[9] Babic, K., Petrovic, M., Beliga, S., Martincic-Ipsic, S., Matesic, M., & Mestrovic, A. (2021). Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model . Applied Sciences-Basel, 11(21), 22, Article 10442.

[10] Badi, H., Badi, I., El Moutaouakil, K., Khamjane, A., & Bahri, A. (2022). Sentiment Analysis And Prediction Of Polarity Vaccines Based On Twitter Data Using Deep Nlp Techniques. Radioelectronic and Computer Systems, 2022(4), 19-29.

[11] Bibi, M., Abbasi, W. A., Aziz, W., Khalil, S., Uddin, M., Iwendi, C., & Gadekallu, T. R. (2022). A novel unsupervised ensemble framework using concept-based linguistic methods and machine learning for twitter sentiment analysis . Pattern Recognition Letters, 158, 80-86.

[12] Bonifazi, G., Breve, B., Cirillo, S., Corradini, E., & Virgili, L. (2022). Investigating the COVID-19 vaccine discussions on Twitter through a multilayer network-based approach. Information Processing & Management, 59(6), 103095.

[13] Brzustewicz, P., & Singh, A. (2021). Sustainable consumption in consumer behavior in the time of covid-19: Topic modeling on twitter data using lda. Energies, 14(18).

[14] Camacho, K., Portelli, R., Shortridge, A., & Takahashi, B. (2021). Sentiment mapping: point pattern analysis of sentiment classified Twitter data. Cartography and Geographic Information Science, 48(3).

[15] Cardone, B., Di Martino, F., & Senatore, S. (2021). Improving the emotion-based classification by exploiting the fuzzy entropy in FCM clustering. International Journal of Intelligent Systems, 36(11).

[16] Cartaxo, B.,Pinto,G.,Soares, S. (2018). "The role of rapid reviews in supporting decision-making in software engineering practice Proceedings of the 22nd Conference EASE,Christchurch,New Zealand.

[17] Chauhan, N. S. (2022). DBSCAN Clustering Algorithm in Machine Learning. KDnuggets. Retrieved 20 ian 2024 from https://www.kdnuggets.com/2020/04/dbscan-clustering-algorithm-machine-learning.html

[18] Chihab, M., Chiny, M., Boussatta, N. M. H., Chihab, Y., Hadi, M. Y. (2022). BiLSTM and Multiple Linear Regression based Sentiment

Analysis Model using Polarity and Subjectivity of a Text. International Journal of Advanced Computer Science and Applications, 13(10).

[19] Cordoba-Cabus, A., Hidalgo-Arjona, M., & Lopez-Martin, A. (2021). Coverage of the 2021 Madrid regional election campaign by the main Spanish newspapers on Twitter: natural language processing and machine learning algorithms . Profesional De La Informacion, 30(6), 17.

[20] Cyril, C. P. D., Beulah, J. R., Subramani, N., Mohan, P., Harshavardhan, A., & Sivabalaselvamani, D. (2021). An automated learning model for sentiment analysis and data classification of Twitter data using balanced CA-SVM. Concurrent Engineering Research and Applications, 29(4).

[21] Dal Mas, F., Piccolo, D., Cobianchi, L., Edvinsson, L., Presch, G., Massaro, M., Bagnoli, C. (2019, Oct 31-Nov 01). The Effects of Artificial Intelligence, Robotics, and Industry 4.0 Technologies. Insights from the Healthcare Sector. [Proceedings of the european conference on the impact of artificial intelligence and robotics (eciair 2019)]. European Conference on the Impact of Artificial Intelligence and Robotics (ECIAIR), EM Normandie Business Sch, Oxford, ENGLAND.

[22] Dhiman, A., & Toshniwal, D. (2022). AI-based Twitter framework for assessing the involvement of government schemes in electoral campaigns. Expert Systems with Applications, 203.

[23] Dutta, R., Das, N., Majumder, M., & Jana, B. Aspect based sentiment analysis using multi-criteria decision-making and deep learning under COVID-19 pandemic in India [Article; Early Access]. Caai Transactions on Intelligence Technology, 16.

[24] Dwivedi, D. N., Mahanty, G., & Vemareddy, A. (2022). How Responsible Is AI? Identification of Key Public Concerns Using Sentiment Analysis and Topic Modeling . International Journal of Information Retrieval Research, 12(1), 14.

[25] Dzyuban, Y., Ching, G. N. Y., Yik, S. K., Tan, A. J., Crank, P. J., Banerjee, S., Chow, W. T. L. (2022). Sentiment Analysis of Weather-Related Tweets from Cities within Hot Climates. Weather, Climate, and Society, 14(4), 1133-1145.

[26] Es-Sabery, F., Hair, A., Qadir, J., Sainz-De-Abajo, B., Garcia-Zapirain, B., & Torre-Diez, I. (2021). Sentence-Level Classification Using Parallel Fuzzy Deep Learning Classifier. IEEE Access, 9.

[27] Ezugwu, A. E., Ikotun, A. M., Oyelade, O. O., Abualigah, L., Agushaka, J. O., Eke, C. I., & Akinyelu, A. A. (2022). A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects. Engineering Applications of Artificial Intelligence, 110.

[28] Ghiassi, M., Lee, S., & Gaikwad, S. R. (2022). Sentiment analysis and spam filtering using the YAC2 clustering algorithm with transferability . Computers and Industrial Engineering, 165, Article 107959.

[29] Ghiassi, M., Saidane, H., & Oswal, R. (2021). YAC2: An α-proximity based clustering algorithm. Expert Systems with Applications, 167.

[30] Gupta, B., Kulkarni, G., Kumar, A. R., Padmini, V. S., Uma, S. M., & Roy, D. R. (2021). Some enhancements in the choice of functionalities for data mining and their application in opinion mining. Journal of Nuclear Energy Science and Power Generation Technology, 10(9).

[31] Gupta, I., & Joshi, N. (2021). Real-time Twitter corpus labelling using automatic clustering approach. International Journal of Computing and Digital Systems, 10(1), 519-532.

[32] Hassan, A., Abbasi, A., & Zeng, D. (2013). Twitter sentiment analysis: A bootstrap ensemble framework. In 2013 International Conference On Social Computing

[33] Hayawi, K., Shahriar, S., Serhani, M. A., Taleb, I., & Mathew, S. S. (2022). ANTi-Vax: a novel Twitter dataset for COVID-19 vaccine misinformation detection. Public Health, 203, 23-30.

[34] Hennig, C. (2022). An empirical comparison and characterisation of nine popular clustering methods. Advances in Data Analysis and Classification, 16(1), 201-229.

[35] Huang, J., Obracht-Prondzynska, H., Kamrowska-Zaluska, D., Sun, Y., & Li, L. (2021). The image of the City on social media: A comparative study using "Big Data" and "Small Data" methods in the Tri-City Region in Poland. Landscape and Urban Planning, 206.

[36] Hussein, A., Ahmad, F. K., & Kamaruddin, S. S. (2021). Cluster Analysis on Covid-19 Outbreak Sentiments from Twitter Data using K-means Algorithm. Journal of System and Management Sciences, 11(4).

[37] Ibrahim, A. F., Hassaballah, M., Ali, A. A., Nam, Y., & Ibrahim, I. A. (2022). COVID19 outbreak: A hierarchical framework for user sentiment analysis. Computers, Materials and Continua, 70(2)

[38] Iparraguirre-Villanueva, O., Guevara-Ponce, V., Sierra-Liñan, F., Beltozar-Clemente, S., & Cabanillas-Carbonell, M. (2022). Sentiment Analysis of Tweets using Unsupervised Learning Techniques and the K-Means Algorithm . International Journal of Advanced Computer Science and Applications, 13(6), 571-578.

[39] Karaca, Y. E., & Aslan, S. (2021). Sentiment Analysis of Covid-19 Tweets by using LSTM

[40] Learning Model. Journal of Computer Science, IDAP-2021(Special), 366-374. https://doi.org/https://doi.org/10.53070/bbd.990421

[41] Kovacs, T., Kovacs-Gyori, A., & Resch, B. (2021). #AllforJan: How Twitter Users in Europe Reacted to the Murder of Jan Kuciak-Revealing Spatiotemporal Patterns through Sentiment Analysis and Topic Modeling . Isprs International Journal of Geo-Information, 10(9), 22.

[42] Kumar, S., Khan, M. B., Hasanat, M. H. A., Saudagar, A. K. J., AlTameem, A., & AlKhathami, M. (2022). An Anomaly Detection Framework for Twitter Data. Applied Sciences (Switzerland), 12(21).

[43] Lee, H. J., Lee, M., Lee, H., & Cruz, R. A. (2021). Mining service quality feedback from social media: A computational analytics method. Government Information Quarterly, 38(2), Article 101571.

[44] Li, C., Mao, K., Liang, L., Ren, D., Zhang, W., Yuan, Y., & Wang, G. (2021). Unsupervised active learning via subspace learning. In Proceedings of the AAAI Conference on Artificial Intelligence

[45] Liu, Y., Yin, Z., Ni, C., Yan, C., Wan, Z., & Malin, B. (2023). Examining Rural and Urban Sentiment Difference in COVID-19-Related Topics on Twitter: Word Embedding-Based Retrospective Study. Journal of medical Internet research, 25, e42985.

[46] Mendon, S., Dutta, P., Behl, A., & Lessmann, S. (2021). A Hybrid Approach of Machine Learning and Lexicons to Sentiment Analysis: Enhanced Insights from Twitter Data of Natural Disasters. Information Systems Frontiers, 23(5), 1145-1168.

[47] Mishra, R. K., Urolagin, S., Jothi, J. A. A., Neogi, A. S., Nawaz, N. (2021).Deep Learning-based Sentiment Analysis and Topic Modeling on Tourism During Covid-19 Pandemic. Frontiers in Computer Science.

[48] Monkeylearn. (2022). No-Code Text Analytics. Retrieved 20 Febr 2024 from https://monkeylearn.com

[49] Moreno, A., & Iglesias, C. A. (2021). Understanding customers' transport services with topic clustering and sentiment analysis. Applied Sciences (Switzerland), 11(21). https://doi.org/10.3390/app112110169

[50] Naeem, S., Mashwani, W. K., Ali, A., Uddin, M. I., Mahmoud, M., Jamal, F., & Chesneau, C. (2021). Machine Learning-based USD/PKR Exchange Rate Forecasting Using Sentiment Analysis of Twitter Data. Computers, Materials and Continua, 67(3), 3451-3461.

[51] Ogbuokiri, B., Ahmadi, A., Bragazzi, N. L., Movahedi Nia, Z., Mellado, B., Wu, J., Kong, J. (2022). Public sentiments toward COVID-19 vaccines in South African cities: An analysis of Twitter posts . Frontiers in Public Health, 10, Article 987376.

[52] Oyewole, G. J., & Thopil, G. A. (2022). Data clustering: application and trends. Artificial Intelligence Review.

[53] Pandey, A. C., Kulhari, A., & Shukla, D. S. (2022). Enhancing sentiment analysis using Roulette wheel selection based cuckoo search clustering method. Journal of Ambient Intelligence and Humanized Computing, 13(1).

[54] Pindado, E., & Barrena, R. (2021). Using Twitter to explore consumers' sentiments and their social representations towards new food trends. British Food Journal, 123(3), 1060-1082.

[55] Popescul, D., Radu, L. D., Păvăloaia, V. D., & Georgescu, M. R. (2020). Psychological Determinants of Investor Motivation in Social Media-Based Crowdfunding Projects: A Systematic Review. In Frontiers in Psychology (Vol. 11).

[56] Pradhan, R., & Sharma, D. K. (2022). A hierarchical topic modelling approach for short text clustering. International Journal of Information and Communication Technology, 20(4), 463-481.

[57] Prottasha, N. J., Sami, A. A., Kowsher, M., Murad, S. A., Bairagi, A. K., Masud, M., & Baz, M. (2022). Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning. Sensors, 22(11).

[58] Proudfoot, D. (2022). An Analysis of Turing's Criterion for 'Thinking'. Philosophies, 7(6 C7 - 124).

[59] Radu, L. D. (2020). Disruptive Technologies in Smart Cities: A Survey on Current Trends and Challenges. Smart Cities, 3(3), 1022-1038.

[60] Ramya, G. R., & Bagavathi Sivakumar, P. (2021). An incremental learning temporal influence model for identifying topical influencers on Twitter dataset. Social Network Analysis and Mining, 11(1).

[61] Rehman, M., Razzaq, A., Baig, I. A., Jabeen, J., Tahir, M. H. N., Ahmed, U. I., Abbas, T. (2022). Semantics Analysis of Agricultural Experts' Opinions for Crop Productivity through Machine Learning. Applied Artificial Intelligence, 36(1).

[62] Russell, A. M., Valdez, D., Chiang, S. C., Montemayor, B. N., Barry, A. E., Lin, H. C., & Massey, P. M. (2022). Using Natural Language Processing to Explore "Dry January" Posts on Twitter: Longitudinal Infodemiology Study. Journal of Medical Internet Research, 24(11).

[63] Ruz, G. A., Henríquez, P. A., & Mascareño, A. (2022). Bayesian Constitutionalization: Twitter Sentiment Analysis of the Chilean Constitutional Process through Bayesian Network Classifiers . Mathematics, 10(2), Article 166. https://doi.org/10.3390/math10020166

[64] Salhi, D. E., Tari, A., & Kechadi, M. T. (2021). Using E-reputation for sentiment analysis: Twitter as a case study. International Journal of Cloud Applications and Computing, 11(2), 32-47.

[65] Scikit-learn.Clustering performance evaluation. https://scikit-learn.org/stable/modules/clustering.html#clustering-performance-evaluation

[66] Shah, S. H., Iqbal, M. J., Bakhsh, M., & Iqbal, A. (2020). Analysis of different clustering algorithms for accurate knowledge extraction from popular datasets. Inf. Sci. Lett, 9(4), 21-31.

[67] Shekhawat, S. S., Shringi, S., & Sharma, H. (2021). Twitter sentiment analysis using hybrid Spider Monkey optimization method. Evolutionary Intelligence, 14(3), 1307-1316.

[68] Singh, M., Singh, A., Bharti, S., Singh, P., & Saini, M. (2022). Using Social Media Analytics and Machine Learning Approaches to Analyze the Behavioral Response of Agriculture Stakeholders during the COVID-19 Pandemic. Sustainability (Switzerland), 14(23).

[69] Sufi, F. (2022). A decision support system for extracting artificial intelligence-driven insights from live twitter feeds on natural disasters . Decision Analytics Journal, 5.

[70] Sufi, F. K. (2022). AI-SocialDisaster: An AI-based software for identifying and analyzing natural disasters from social media . Software Impacts, 13.

[71] Tania, M. H., Hossain, M. R., Jahanara, N., Andreev, I., & Clifton, D. A. (2022). Thinking Aloud or Screaming Inside: Exploratory Study of Sentiment Around Work. JMIR Formative Research, 6(9).

[72] Vanam, H., Jebersonretna Raj, R., & Janga, V. (2023). Novel cluster set optimization model with unique identifier tagging for twitter data analysis. Journal of Intelligent and Fuzzy Systems, 44(2), 2031-2039.

[73] VOSViewer. (2022). Visualizing scientific landscapes. Retrieved 20 Jan 2024 from https://www.vosviewer.com/features/highlights

[74] Yao, Q., Li, R. Y. M., & Song, L. X. (2022). Carbon neutrality vs. neutralite carbone: A comparative study on French and English users' perceptions and social capital on Twitter. Frontiers in Environmental Science, 10, 11.

[75] Yao, W., & Qian, S. (2021). From Twitter to traffic predictor: Next-day morning traffic prediction using social media data. Transportation Research Part C: Emerging Technologies, 124.

[76] Yao, Z., Yang, J., Liu, J., Keith, M., & Guan, C. (2021). Comparing tweet sentiments in megacities using machine learning techniques: In the midst of COVID-19. Cities, 116.

[77] Yenduri, G., Rajakumar, B. R., Praghash, K., & Binu, D. (2021). Heuristic-assisted bert for twitter sentiment analysis. International Journal of Computational Intelligence and Applications, 20.(03).

[78] Yousefinaghani, S., Dara, R., Mubareka, S., Papadopoulos, A., & Sharif, S. (2021). An analysis of COVID-19 vaccine sentiments and opinions on Twitter. International Journal of Infectious Diseases, 108, 256-262.

[79] Zhang, J., Wang, Y., Shi, M., & Wang, X. (2021). Factors driving the popularity and virality of Covid-19 vaccine discourse on Twitter: Text mining and data visualization study. JMIR Pub. Health and Surv., 7(12).

[80] Zhang, Y., Abbas, M., & Iqbal, W. (2021). Analyzing sentiments and attitudes toward carbon taxation in Europe, USA, South Africa, Canada and Australia. Sustainable Production and Consumption, 28, 241-253.