A Machine Learning Based Social Network Data Mining System for Better Search Engine Algorithm

Sheik Erfan Ahmed Himu[1], Arafat Ibne Ikram[2], Kazi Mohammed Abdullah[3],
Tasnia Fahrin Choity[4], Md. Shahazan Parves[5], Md. Rabiul Hasan[6], Md Imtiaz Uddin[7]
[1]*Dept. of Electrical and Mechanical Engineering, Nagoya Institute of Technology. Nagoya, Japan*
[2567]*Dept. of Electrical and Electronic Engineering, International Islamic University Chittagong. Chattogram, Bangladesh*
[3]*Dept. of Electronic and Telecommunication Engineering, International Islamic University Chittagong.Chattogram, Bangladesh*
[4]*Dept. Computer Science and Engineering, International Islamic University Chittagong. Chattogram, Bangladesh*
Email: [1]erfanhimu@gmail.com, [2]arafatibne.ikram@gmail.com, [3]kazitusher636@gmail.com,
[4]tasniachoity@gmail.com, [5]parveziiuc4708@gmail.com, [6]robiulhasan4245@gmail.com, [7]imtiazanam56@gmail.com

***Abstract*—Social networks provide access to a vast amount of information in a variety of ways. However, these are poorly organized. On the other hand, data search engines include limited and occasionally inaccurate data. So, it is possible to find the right analytical answer to a problem based on what people say on social media. Consequently, a machine learning-based social network data mining system will aid in the development of a superior search engine. Google is thought to be the best search engine in the world right now. People used to do SEO after building a website to improve its search engine ranking. Google's AI decides which search results to show based on how much traffic that website gets. But we still aren't getting the right results. But if we configure our search algorithm to utilize social network user-generated content based on their sentiment, we can obtain accurate search results. This paper proposed a machine learning-based data mining system from social networks where all data is collected from social networks using linear regression, polynomial regression, and percentile machine learning techniques and stores unstructured and pre-structured data in big data for data validation. With the help of some techniques, we can show that the information from social networks is a good way to solve our problems.**

***Index Terms*—Big data, Social Media Analysis, Machine Learning, Data Mining, Searching Algorithms, Search Engine**

I. INTRODUCTION

In today's world, social media has a significant presence and influence. Every social media platform, including Facebook, Twitter, WhatsApp, Telegram, Linked-in, Instagram, and others, comprises a diverse range of individuals. People used to talk about a wide range of subjects and exchange knowledge with one another [1]. Even though the majority of the information gathered through social media is unverified, there is little information that may be utilized for a more beneficial purpose. We can enhance stock analysis, medical data analysis, political analysis, and a variety of other tasks by using social media technologies [2]. However, to make appropriate use of them, we must first correctly gather them. For data collection, we offer several different options. Social media postings, such as Facebook page posts and Facebook group posts, Instagram feeds and Twitter trends, Linked-in activities, WhatsApp business APIs, and Telegram open groups, will assist us in effectively collecting data [3]. During the COVID-19 epidemic, we saw the most beneficial effects of social media networks in action. When there is a pandemic As of the 14th of April, 2020, there are 18,53,155 people impacted in 213 nations; 1,14,247 fatalities; and just 4,23,625 people who have been rescued over the globe. We were only able to contact via social media [4]. It assists us in informing the public about the need for a lockdown. As a result, there was a great deal of encouragement for individuals to learn about the realities of the COVID-19 pandemic scenario. In addition, we may use this knowledge to produce appropriate system analyses and system solutions to contemporary situations [5], [6].

Nevertheless, when it comes to a social media solution, the first question that comes to mind is the data privacy system [7]. As we all know, individuals are quite interested in keeping their data safe from the outside world [8]. As a result, if we look at social media data, we can see that there has been the potential for gathering social data that has not been permitted for public disclosure. However, we primarily collect public data that has been granted permission to be shared publicly. In addition to this, we will employ several predetermined legitimate ways to obtain user private data to maintain appropriate privacy. To collect all forms of system data, we will save all of the system data in big-data format. Our machine learning methods will assist in determining which sorts of data should be processed in which directory system, and after that, we grouped the data into three variants for further analysis. Clusters are classified as follows: 1) structured, 2) non-structured, and 3) pre-structured.

Another concern will arise in the context of social media data harvesting. And it is the integrity of the data [9]–[11]. As we all know, most social data is not entirely trustworthy, so we must rely on a machine-learning algorithm to verify the data's validity [8]. We can identify social network user data-sharing variations depending on age and education to aid in the development of a suitable algorithm. Based on summarizing the resulting values using social data from users, the system can automatically identify the data that is legitimate from

979-8-3503-1965-1/23/$31.00

different data sources.

In this project, we will collect 10,000 social media opinions on a variety of timely topics. As for user opinions, we mostly consider user-generated polls, forum topics, Twitter trends, and Instagram polls. After that, they validated the legitimacy of this data based on their age, their taste in content, their history of content sharing, and their level of education. After Having verified the data, they stored it in the big data system and used it as a legitimate search result. Whenever a user does a search based on the system's training records for the search pole, we use the previously recorded data as training data. Then, a machine learning model will be developed using linear regression analysis. Then, it can view the most reliable information from the search results.

II. Related Work

Previously, so much work has been done on this topic. Everyone reflects proper logic and methods in their research. S. Gole et al. proposed a Map-Reduce Framework using Clust Big FIM. They use the Clust Big FIM algorithm. Their outcome is to perform associations, emerging patterns, sequential patterns, correlations, and other significant data mining tasks [12]. Another author, A. Akay et al., proposed a Network-Based Modeling and Intelligent Data Mining of Social Media for Improving Care [13]. P.Liang et al. also proposed a social media opinion-based data mining system. They analyze user reactions and detect positive as well as negative reactions [14]. Flesca et al. published a research work titled "How Data Mining and Machine Learning Evolved from Relational Databases to Data Science". They use the usage of a relational database for data mining and a brief discussion of the data mining scopes using machine learning [15]. E. Cambria and colleagues use cross-domain social analytic tools for big-data analyses. Their research returns the analytical representation for Big Social Data [16]. L. Oneto et al. use mass media and social network opinions to generate sentimental analyses. Their work is based on statistical learning theory to develop an ELM(extreme learning machine) [17]

III. Problem Statements

Mining big social data is not an easy task. It takes proper algorithms to fetch only the relevant information. Only a proper data mining process can help us to develop efficient big-data analyses. The three main issues that cause problems during the data filtering process are 1) fake information filtering, and 2) inappropriate data filtering. 3) Hypothetical data returns. Major search engines don't trust social network information, but there is more information on social network platforms than on any other search engine.

Social media content is not well structured. Sometimes we find authentic information from social posts or shared posts. Sometimes we have to look through social post comments or replied comments to find recent usable information. We know people like to talk about current content on social media. That's why Twitter brought an option named Twitter Trends. And so, for first and reasonable data mining, social media is the best option for recent trends.

To utilize their data, we need multiple algorithms for finding accurate information from social networks. Having that in mind, we designed a system that will take social network data and set an analysis based on information validity.

IV. Methodology

In this section describes the methodology of the entire research project. Several techniques were used here for data insertion and data analysis from Big Social Data using Machine Learning. The flowchart of the whole system is illustrated in fig 1

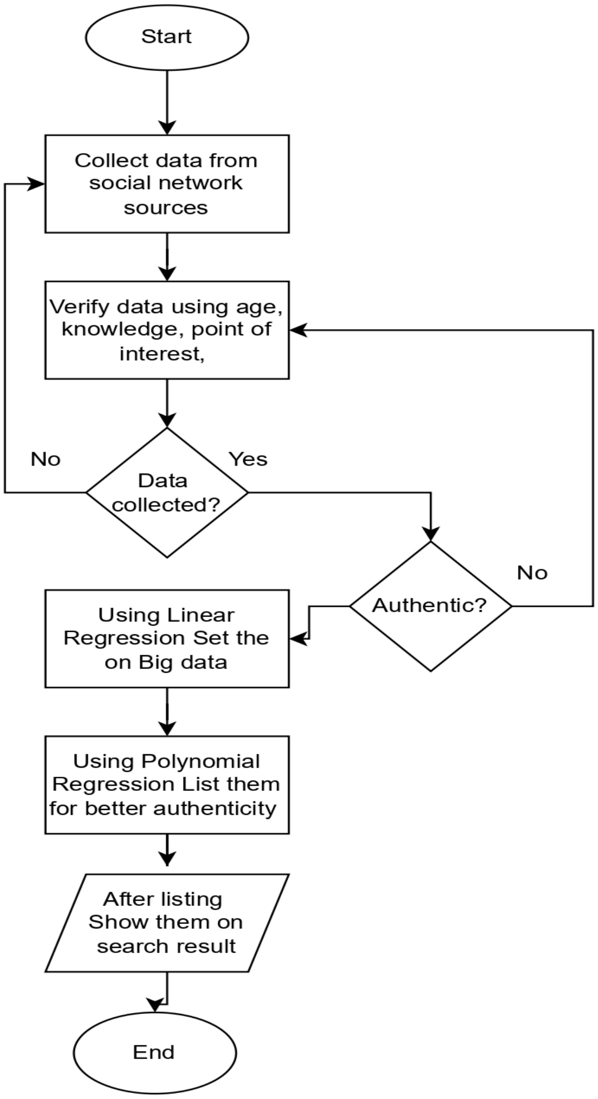
Start
Collect data from social network sources
Verify data using age, knowledge, point of interest,
No
Yes
Data collected?
No
Authentic?
Using Linear Regression Set the on Big data
Using Polynomial Regression List them for better authenticity
After listing Show them on search result
End

Fig. 1. Flowchart of the whole system

V. System Design

This section describes the techniques used for data insertion and data analysis from Big Social Data using Machine Learning.

A. Fake News Detection

The Fake Detector addresses two main components: representation feature learning and credibility label inference, which together will compose the deep diffusive network model Fake Detector [18]. Fake information returns more negative opinions. So, by analysis, the comments can often detect fake news. On the other hand, the author's data-sharing possibility also returns a fake news probability. All the following procedures are visualized in fig. 2

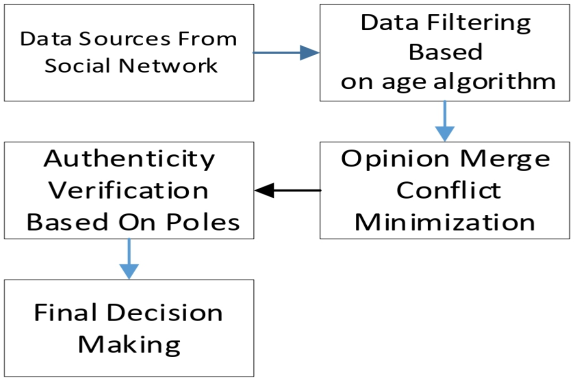
Data Sources From Social Network
Data Filtering Based on age algorithm
Opinion Merge Conflict Minimization
Authenticity Verification Based On Poles
Final Decision Making

Fig. 2. Fake News Detection Process

B. Data filtering

This section describes the process of collecting data into various classes. People of younger ages have less experience, so the information they are sharing will have a low value of acceptance [19]. People of older age will have more acceptance because of their experience. We used Euclidean measurement for generating data processing values.

$$S_{\epsilon} = \int_{N_{\iota}}^{N_{\mu}} LR + C \qquad (1)$$

In equation (1), S$_\epsilon$ variable represents for social data accuracy report using machine learning. Here N$_\mu$ shows the higher age limit and N$_\iota$ shows the lower age limits from the system. L represents the number of shared posts from an experimental user, and R represents post density per unit time value. C is the constant value for a unique platform. If we count data validation from Twitter posts, then our constant will be based on the total trending topics on Twitter. If data validation is from a Facebook profile, then the constant value will be the total time length of that profile. The whole process is shown in Fig. 3.

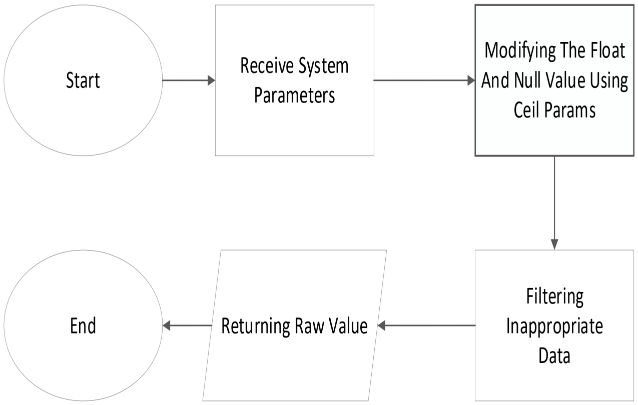
Start
Receive System Parameters
Modifying The Float And Null Value Using Ceil Params
Filtering Inappropriate Data
Returning Raw Value
End

Fig. 3. Data filtering based on multiple source process

C. Data Validity Based on User Choice

Sometimes user-defined logic does not work. In that case, here we use linear regression analyses based on hypothetical logical returns from user opinions. We took 1000 users' data for the solution to socioeconomic condition change. Every sample gave their own opinion on the issues [20] [21] [22] [23]. The process of the find data validation is shown in fig 4

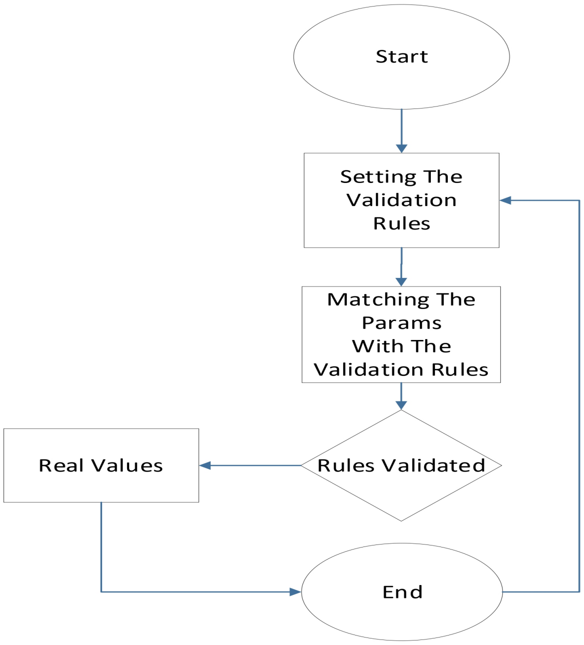
Start
Setting The Validation Rules
Matching The Params With The Validation Rules
Rules Validated
Real Values
End

Fig. 4. Data validity based on user choice process

VI. Result And Analysis

In this section, we describe the result and the accuracy of the system.

A. System Calculation

Using the Softmax function for the system's probability generation, here we calculate the posterior probability of the nth class [24]. It helps to collect the proper data format from the big-data system. After data mining, one major task is to use them in the proper sector. Here we need actual information in a specific place in an actual format. Without proper classification, it is quite impossible.

$$P(y=j|h)=\frac{e^{x^T w_j}}{\sum_{k=1}^{K} e^{x^T w_k}} \tag{2}$$

Equation (3), is about system categorization techniques for data collection from the system. where K is the number of different functions and w is the weighting vector. The pre-trained BERT model is then fine-tuned on individual datasets. To measure the effectiveness of the fine-tuned BERT model, various baseline models are selected to undertake the identical classification tasks [25]. The basic models are Bidirectional GRU (BiGRU), Bidirectional LSTM (BiLSTM), Hybrid CNN and GRU, Hybrid CNN and LSTM, Deep Pyramid CNN (DPCNN), CNN with K-Max pooling (KMax-CNN), Region-based CNN (R-CNN) [26]. The accuracies for validation and testing results are calculated to evaluate the performance of different models:

$$Accuracy=\frac{TruePositive+TrueNegative}{TotalNumberofPost} \tag{3}$$

In Equation (3) are the algorithms for gaining data accuracy. Here, label each post with a humanitarian category so that information about the situation can be sorted reliably without any manual work.

B. Result

1) Linear Regression: Here, 1,000 user data points were collected to address changing socioeconomic conditions. Each sample provided its viewpoint on the matter. The main issue was the number of losses during the COVID-19 pandemic's effects on both micro-business systems and large enterprise systems. output statistical review is given below. Now, based on the user opinions, we generate machine-learning linear regression analyses to get the actual data value from the system [27]. It is seen that most of the data validation counts come from a negative review because negative reviews contain proper fact analyses [28].

TABLE I. Data Statistics Based On User Opinions

User Opinion	Percentage of people agreed opinions
People who think small business causes much damage on pandamic	73% people of total Analysis
People who thinks both small and large business causes damage	8% people of total analysis
People who think only large business causes much damage on pandemic	19% people of total analysis

Fig. 5 shows the linear regression analysis based on social media opinion. On the x-axis, we have counted the total number of people in parentage given opinions, and the y label shows the opinion type is Boolean. If opinions are positive, then it returns 1, and if opinions are negative, then it returns 0. So, using that fact analysis, we can determine the actual solution.

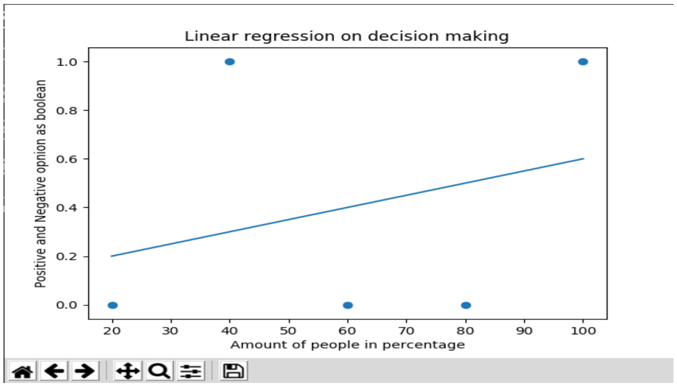

Fig. 5. Linear regression analyses based on social media opinions

2) Experimental Result: To conduct a preliminary evaluation of the proposed approach, we analyzed some social media facts and compared them to random search engine results. Based on that analysis, we discovered some accuracy. The overall performance is given in Table II.

Here we can see some variations in results and accuracy.

TABLE II. Performance Evaluation Table

Topic	**Social Media Result**	**Google Result**	**Bing Result**	**Accuracy On our Search**
Covid-19 medicine efficiency	73% of people think they are effective	50% effective based on research results	52% effective based on search result	73.9%
Global warming facts	Most major analyses on social media focused on greenhouse effects and industrialization	Among major issues, google focused on greenhouse effects and volcanic activity	Among major issues, Bing focused on greenhouse effects	78%
Financial investment in a global stock exchange	56% of people think that investing in a global stock is worthy	google highly recommends investment	Bing returns to authentic results	50%

And so, based on those results, we can assume that without the consideration of social media for search engine algorithms, it is not a good decision.

VII. Conclusion

In this research, our main motive was to develop a search engine machine with more convenient information. Typical search engines are used to cover up social media content listings from trends and hashtags. However our analysis focuses more on a public opinion basis. So our analytical solutions demand a better search engine platform. However, fake and unauthenticated data filter the most out of it. But whatever remains will create a huge data center for us. The most popular search engines are still upgrading their listings and policies day by day. But our social networking community is growing more and more over the years. To compete in the most popular search engine race, the social media content mining process will be the best shortcut for us. In the future, our research will also be focused on the basis of video and personal blog basics. But if we want to add a major contribution from personal blogs and profiles, we need the revolution of Web 3.0.

References

[1] L. Liang, Y. Deng, Q. Huang, H. Rui, and R. Zhan, "Comparison of total rna extraction methods from nervilia fordii (hance) schltr. leaves," *Northern Horticulture*, vol. 2, pp. 91–94, 2013.

[2] M. Gastaldi, "Integration of mobile, big data, sensors, and social media: impact on daily life and business," in *2014 IST-Africa Conference Proceedings*, pp. 1–10, IEEE, 2014.

[3] S. X. Mashal and K. Asnani, "Emotion intensity detection for social media data," in *2017 International Conference on Computing Methodologies and Communication (ICCMC)*, pp. 155–158, IEEE, 2017.

[4] R. Sathish, R. Manikandan, S. S. Priscila, B. V. Sara, and R. Mahaveerakannan, "A report on the impact of information technology and social media on covid–19," in *2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)*, pp. 224–230, IEEE, 2020.

[5] A. I. Ikram, M. S.-U. Islam, M. A. B. Zafar, M. K. R. Dept, A. Rahman, *et al.*, "Techno-economic optimization of grid-integrated hybrid storage system using ga," in *2023 1st International Conference on Innovations in High Speed Communication and Signal Processing (IHCSP)*, pp. 300–305, IEEE, 2023.

[6] X. Cao, D. R. Vogel, X. Guo, H. Liu, and J. Gu, "Understanding the influence of social media in the workplace: An integration of media synchronicity and social capital theories," in *2012 45th Hawaii International Conference on System Sciences*, pp. 3938–3947, IEEE, 2012.

[7] M. T. Ahvanooey, Q. Li, J. Hou, H. D. Mazraeh, and J. Zhang, "Aitsteg: An innovative text steganography technique for hidden transmission of text message via social media," *IEEE Access*, vol. 6, pp. 65981–65995, 2018.

[8] C. Hutto and C. Bell, "Social media gerontology: Understanding social media usage among a unique and expanding community of users," in *2014 47th Hawaii International Conference on System Sciences*, pp. 1755–1764, IEEE, 2014.

[9] W. M. Al-Rahmi, N. Yahaya, M. M. Alamri, N. A. Aljarboa, Y. B. Kamin, and F. A. Moafa, "A model of factors affecting cyber bullying behaviors among university students," *Ieee Access*, vol. 7, pp. 2978–2985, 2018.

[10] S. E. A. Himu, S. Sultana, M. S. H. Chowdhury, A. I. Ikram, H. R. Saium, and M. M. Hossain, "Modification of dynamic logic circuit design technique for minimizing leakage current and propagation delay," in *2022 4th International Conference on Sustainable Technologies for Industry 4.0 (STI)*, pp. 1–5, IEEE, 2022.

[11] A. I. Ikram, A. Ullah, D. Datta, A. Islam, and T. Ahmed, "Optimizing energy consumption in smart homes: Load scheduling approaches," *IET Power Electronics*, 1-13 2024.

[12] S. Gole and B. Tidke, "Frequent itemset mining for big data in social media using clustbigfim algorithm," in *2015 International Conference on Pervasive Computing (ICPC)*, pp. 1–6, IEEE, 2015.

[13] A. Akay, A. Dragomir, and B.-E. Erlandsson, "Network-based modeling and intelligent data mining of social media for improving care," *IEEE journal of biomedical and health informatics*, vol. 19, no. 1, pp. 210–218, 2014.

[14] P.-W. Liang and B.-R. Dai, "Opinion mining on social media data," in *2013 IEEE 14th international conference on mobile data management*, vol. 2, pp. 91–96, IEEE, 2013.

[15] S. Flesca, S. Greco, E. Masciari, and D. Saccà, *A comprehensive guide through the italian database research over the last 25 years*, vol. 31. Springer, 2018.

[16] E. Cambria, N. Howard, Y. Xia, and T.-S. Chua, "Computational intelligence for big social data analysis [guest editorial]," *IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 8–9, 2016.

[17] B. Oneto, "Oneto l., bisio f., cambria e., anguita d," *Statistical learning theory and ELM for big social data analysis, IEEE Computational Intelligence Magazine*, vol. 11, no. 3, pp. 45–55, 2016.

[18] S. I. Manzoor, J. Singla, *et al.*, "Fake news detection using machine learning approaches: A systematic review," in *2019 3rd international conference on trends in electronics and informatics (ICOEI)*, pp. 230–234, IEEE, 2019.

[19] S. Athmaja, M. Hanumanthappa, and V. Kavitha, "A survey of machine learning algorithms for big data analytics," in *2017 International conference on innovations in information, embedded and communication systems (ICIIECS)*, pp. 1–4, IEEE, 2017.

[20] N. Benchettara, R. Kanawati, and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks," in *2010 international conference on advances in social networks analysis and mining*, pp. 326–330, IEEE, 2010.

[21] A. Porshnev, I. Redkin, and A. Shevchenko, "Machine learning in prediction of stock market indicators based on historical data and data from twitter sentiment analysis," in *2013 IEEE 13th International Conference on Data Mining Workshops*, pp. 440–444, IEEE, 2013.

[22] C. Stanik, M. Haering, and W. Maalej, "Classifying multilingual user feedback using traditional machine learning and deep learning," in *2019 IEEE 27th international requirements engineering conference workshops (REW)*, pp. 220–226, IEEE, 2019.

[23] M. M. Ahsan and Z. Siddique, "Machine learning-based disease diagnosis: A bibliometric analysis," *arXiv preprint arXiv:2201.02755*, 2022.

[24] C. Fan, F. Wu, and A. Mostafavi, "A hybrid machine learning pipeline for automated mapping of events and locations from social media in disasters," *IEEE Access*, vol. 8, pp. 10478–10490, 2020.

[25] A. U. Hassan, J. Hussain, M. Hussain, M. Sadiq, and S. Lee, "Sentiment analysis of social networking sites (sns) data using machine learning approach for the measurement of depression," in *2017 international conference on information and communication technology convergence (ICTC)*, pp. 138–140, IEEE, 2017.

[26] M. G. Kibria, K. Nguyen, G. P. Villardi, O. Zhao, K. Ishizu, and F. Kojima, "Big data analytics, machine learning, and artificial intelligence in next-generation wireless networks," *IEEE access*, vol. 6, pp. 32328–32338, 2018.

[27] M. Mohammadi and A. Al-Fuqaha, "Enabling cognitive smart cities using big data and machine learning: Approaches and challenges," *IEEE Communications Magazine*, vol. 56, no. 2, pp. 94–101, 2018.

[28] D. Nallaperuma, R. Nawaratne, T. Bandaragoda, A. Adikari, S. Nguyen, T. Kempitiya, D. De Silva, D. Alahakoon, and D. Pothuhera, "Online incremental machine learning platform for big data-driven smart traffic management," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 12, pp. 4679–4690, 2019.