A HYBRID DNN-HHO APPROACH FOR EVENT DETECTION IN BIG DATA

K. Swapnika[1*]

PhD Scholar, Computer Science and Engineering Department
Jawaharlal Nehru Technological University, Hyderabad, Telangana 500085, India.
swapnika.griet@gmail.com

D. Vasumathi[2]

Professor, Computer Science and Engineering,
Jawaharlal Nehru Technological University, Hyderabad, Telangana 500085, India.
vasukumar_devara@jntuh.ac.in

Abstract

Social media are digitally mediated technologies which are interactive, and enable people to build and share content, career interests, ideas, and other modes of expression through virtual networks. Twitter has gained popularity as a medium of social media in recent years. Twitter is being used by users to post on real-life activities. The prime objective of this paper is to use big data to detect certain events in Twitter. Furthermore, considering the large number of tweets, the event detection algorithm must be scalable. This paper attempts to tackle these challenges with the help of DNN and Harris Hawk optimization algorithm. The main steps include in the proposed methodology are, preprocessing of input data, extraction of useful features, feature selection, and classification. Here, the proposed methodology is implemented using PYTHON platform. The performance of the proposed method is analyzed in terms of statistical measures such as F1-score, precision, accuracy, and recall. The proposed method is utilized for detecting five different events such as education, transportation, environment, geospatial and water and the accuracy obtained for each event is 94%, 92%, 98%, 96% and 96.5% respectively. The overall result shows that, our proposed methodology gives better performance in event detection.

Keywords: **Tokenization, stemming, Bag of words, global redundancy minimization, moth flame optimization, TF-IDF**

1. Introduction

Today we've seen an increase in the amount of digital textual data accessible, which leads in new perspectives and opportunities across new platforms. In the rapidly emerging field of big data analytic techniques, text mining has gotten a lot of attention for a variety of applications. The big data age has arrived as a result of the abundance of data in virtually every area of our society (Daniel, 2019). It is a hot topic that affects many facets of our lives. The word "massive" is also used to describe data that has a lot of value, volume, variety, velocity, and veracity (5 Vs). As large quantities of data are produced in the digital world, the analysis of big data has created unparalleled opportunities in many fields (Tao, 2020). Big data is an unavoidable product of the digitalization trend from the standpoint of scientific and technological development. If this trend continues, we will soon be living in a world where everything is recorded and everything is digitized. Big Data generally indicates huge amount of data or a set of methodologies for collecting, processing, storing, managing, and analyzing it. Text mining among unstructured big data is an important unstructured data analysis technique that has recently been used in many industries (Wan, 2019). Due to the large number of features, the data extraction process using big data text mining and classification creates lowest reliability and high computational complexity.

Despite losing millions of lives and trillions of dollars per year, road transportation remains the foundation of global economies. Twitter is considered as a valuable source of transportation statistics, but there are significant problems in processing of big data. Furthermore, traffic congestion is one of the most serious issues in urban towns, including worries about the cost of congestion (Pandhare and Shah 2017 March). Congestion, collisions, and other public health issues are caused by different parameters such as poor weather, road works, and other uncertainty. These occurrences must be identified in order to facilitate prompt traffic preparation and procedures, as well as the avoidance of negative impacts on public services and health. Twitter is one of the most popular social media platforms, which offers large source of public intelligence.

Different organizations and many people send out tweets to share news, incidents, status updates, and other information, resulting in a massive amount of real-time data on a variety of subjects, including transportation. Twitter is proving to be a useful sensor for detecting incidents, predicting congestion, forecasting flow, and traffic

tracking. However, despite its enormous promise, many significant obstacles must be addressed before it can be widely adopted in transportation. Many researchers are using social media platforms to detect events in different languages such as Chinese, English, Italian and Thai, among other languages (Salas, 2017, July). Second, Twitter generates large amounts of data, which is difficult to manage for event identification because of its characteristics (5V's). There is no methods exists to combine the two problems such as big data and natural language processing in the identification of transportation-related events from Twitter data. Neural networks are used for the classification process, thereby reducing the drawbacks of twitter data.

Motivation

People utilize social media to express their views on socioeconomic issues, post personal events, and connect with one another. Twitter is one such platform, where millions of people used on daily basis to share their personal as well as official matters. Different kinds of tweets are transferring through this social media. Therefore, it is very important to identify and classify these different types of tweets. The main motivation of this research work is to detect and classify events in twitter platform. To achieve better performance also utilizes big data. Thereby, obtaining a better classification results in event detection using twitter.

Objectives

- To design and implement a novel hybrid technique for efficient feature selection process.
- To improve the accuracy of event detection in big data than the existing approaches.
- To modify the performance of classification process and minimize the overall time complexity in event detection process using big data.
- To identify and analyze the events using twitter platform

Organization: The paper is organized as follows. In section 1, a brief introduction about the topic is given and in section 2 provides the works related to the proposed algorithm and section 3 gives the proposed methodology and section 4 provides the result of the work. The entire paper is concluded in fifth section.

2. Literature Review

For big data identification, a multi-feature extraction scheme was proposed by J Wan et al. (Wan, 2019). This work utilizes AI driven MFE scheme using big data. T Wang et al. (Wang, 2018) identified big data reduction for critical infrastructure health monitoring in a smart city. A cloud-based health monitoring application with an IoT framework was proposed as the tool. This reduced the burden of big data analysis at the BS while also improving the efficiency of event detection. Two novel health monitoring schemes were given in this process. One is for big data reduction and the other is for decision-making. In addition, the proposed approach performed admirably in terms of energy cost reduction, data reduction, and monitoring efficiency. The proposed approach shows poor performance in different parameters.

An event detection survey on twitter was proposed by Z Saeed et al. (Saeed, 2019). The efficiency of different Twitter based and other features were described and compared, as well as different evaluation parameters and methodologies. Analysis of big data and ML in intensive care units was proposed by A Nunez et al. (Reiz, 2019). The proposed method was used to improve the decision making in clinical field. The foundations of BDA and ML were reviewed in this paper and suggested some potential strategies used to optimize new technologies and described some hybrid health care data. The emerging problems of feature selection methods for big data analysis were introduced by V Bolon et al. (Bolón-Canedo, 2015). Feature selection is crucial in reducing high-dimensions in ML problems. The major significance of feature selection process was explored in this paper. This big data approach creates both opportunities and challenges for researchers. A Big Data Road-Traffic Event Detection Tool was proposed by E Alomari et al. (Alomari, 2020). The tool was created using spark ML technique and Twitter.

Problem statement

Big data's popularity poses some difficulties for the traditional feature selection task. The challenges on feature selection for big data analytics are scalability and stability. There is minimum works exists for event detection using twitter. The techniques used for event detection is to be precise. The extracted information is to be valid to make better decisions. Big data methodologies are considered as very important because they enable the scalability of the system for data analytics and management. The classification techniques used in the previous research papers are not accurate in nature. A new method is proposed to increase the performance parameters; also preprocessing step includes number of techniques to increase the effect of preprocessing.

3. Proposed Methodology

Initially the data collection step includes the collection of big data from the social media website (twitter) and the duplicates were removed from the dataset. Then preprocessing step is applied to eliminate noise and is prepared for classification. The preprocessing can be done by different methods such as tokenization, lowercase filter, stop word filter, and stemming. A list of cleaned tokens are the output of this component. After that the features from the clear data is extracted using two different techniques such as Bag of Words, and TF-IDF. Feature selection is achieved by using hybrid method of Global Redundancy Minimization and moth flame optimization (GRM-

MFO). Then the selected data is given to the classifier. To improve the classification accuracy, a meta-heuristic optimization algorithm (Harris Hawk Optimization) (DNN-HHO) is hybrid along with the proposed neural network. The trained classifiers are used for event detection. The following figure 1 represents the architecture of proposed methodology.

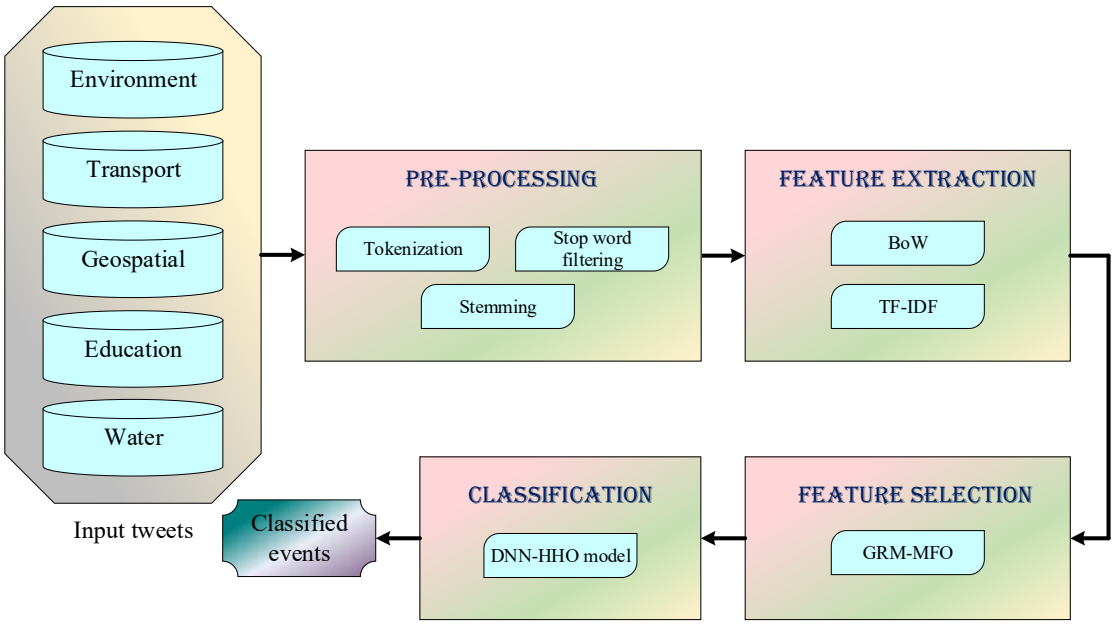

Figure1: Architecture of proposed methodology

3.1 Pre-processing

Pre-processing is considered as the initial step of event detection. The collected data contain number of redundant information as well as noise. To achieve better classification results, preprocessing is very necessary. Here we are used four methods such as tokenization, lowercase filter, stop word filter and stemming.

Tokenization is (Aliwy, 2012) a process which splits the sentence, paragraph or text into its smaller units. Stemming is a process used to reduce words into its stem by chopping off the ends of words and often by removing derivational affixes. Stop words represents the words such as 'the', 'an', 'a', 'in' which need to be ignored while processing. Stop words are removed using a list of words that are already considered as stop words. Natural Language Toolkit (NLT)has a stop word list which consists of nearly 16 different languages. Filtering is another important step in the data cleansing stage.

3.2 Feature extraction

Feature extraction can be defined as the process of choosing or combining variables into features (Xu and Overbye 2015, November). For the text analysis to be perfect, feature extraction need to be done, which makes the process easier. It helps to deduct the data from the vast dataset accurately without losing the important information. The speed of the processing technique is increased due to the feature extraction stage. The techniques used for the extraction of data from the clear data are BoW and TF-IDF. BoW is a method used for the extraction of features from a text document (Zhang, 2010). It tells about the word's occurrence in a document. BoW are also known as n-gram as it extract n words that are contiguous, from the data. The number of occurrence of word is mainly considered in this technique. Each words extracted from the text are assigned by weights in order to represent their importance in the entire text.

TF-IDF is a method which evaluates the relevancy of word in a text document (Ramos, 2003, December). This technique is performed through the multiplication of IDF of word in a document set and number of times a word appears in a document. TF-IDF helps in searching a document or for any information retrieval. The common words like 'this', 'what', 'if' etc have less importance even though they appear more often in the document. This is because, these words have less importance in the document. But, if the word 'bug' appears more often in the document, it is considered as an important word since it has relevancy. The TF-IDF calculation is done by multiplying two metrics such as:

Term frequency: For calculating the frequency of the word, a raw count of the appearance of word can be taken. The frequency can be adjusted using the document length or raw frequency of the word which appears the most.

Inverse document frequency: This means, how rare a word appears or how common a word appears. If the word appears more often, it is considered as 0. This metric is calculated as the ratio of total number of documents to number of documents that contain a word. The value closer to 0 means the words appears more often.

Using these metrics the multiplication of TF-IDF is done. The value obtained is the TF-IDF score which shows whether the word is relevant or not. If the TF-IDF score is high, it means that the word is more relevant. The mathematical calculation for the TF-IDF score is given below. A document set Q which consists of q documents is considered for the calculation.

$$tf\,idf(t,q,Q) = tf(t,q).idf(t,Q) \tag{1}$$

Where,

$$tf\ (t,q) = \log(1 + freq(t.q)) \tag{2}$$

$$idf\ (t,Q) = \log\left(\frac{N}{count(q \in Q : t \in q)}\right) \tag{3}$$

3.3 Feature selection

Feature selection is one of the important stage of every classification and detection process. In our proposed method, the feature selection process can be achieved with the help of a hybrid method known as Global redundancy minimization using moth flame optimization (GRM-MFO).

Global redundancy minimization: This feature selection method is considered as a global redundancy minimization technique in which takes every form of feature ranking score into account (Wang, 2015) (Nie, 2018). The GRM model will minimize global feature redundancy and maintain ranking consistency given a collection of feature ranking scores. This process can be achieved using the following equation.

$$\min_{a^x \geq 0} = \frac{a^x\ y\,z}{p} \tag{4}$$

Where a^x represents the new feature matrix obtained y represents the redundancy matrix and p represents the input feature matrix.

To achieve more accurate results in the feature selection process can be optimized using a moth flame optimization algorithm (Mirjalili, 2015). The moth updates its position with flame according to the following equation.

$$P(x,y) = B.e^{ct}\cos(2\pi t) + F_i \tag{5}$$

Where B represents the euclidian distance c indicates a constant, *x* represents the ith moth F_i represents the ith flame and t represents a number in between [-1,1].

According to the spiral equation 2, moths in MFO tries to shift their locations to every point in the search space. The parameter *t* in the equation is chosen at random and determines a moth's next location. The key component of the algorithm is spiral movement, which determines the position of moths around flames. The moths are flying around the flame using the equation. The new position of *x* can also be expressed using change in the position of F_i and vector change, and it is represented in equation.

$$\alpha_{x,F_i,d} = B.e^{ct}\cos(2\pi t) \tag{6}$$

This equation can be used to optimize the global redundancy minimization process to achieve better feature selection results.

3.4. Feature classification

The classification of event detection can be achieved with the help of a DNN architecture. The architecture of DNN is represented by figure 3. As the figure depicts, the DNN mainly consist of three layers such as input layer (IL), output layer (OL) with multiple hidden layers (HL) (Phan, 2017). In DNN the total hidden layers are more. The data flow in DNN move from input to output layer without twisting back and so it is termed as feed-forward network (Takahashi, 2016). Initially, a map of a virtual neuron is created by DNN model and assigns weights to connect the layers. Thus, the output value is obtained by multiplying the inputs and weights.

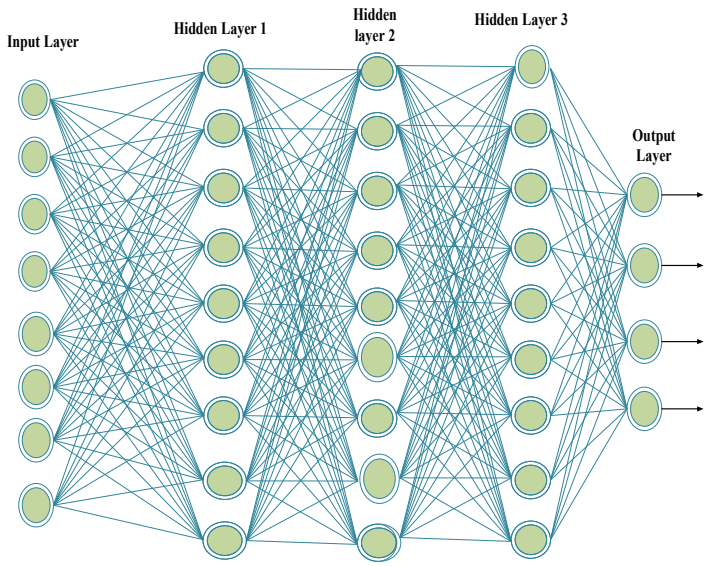

Figure 2: Architecture of DNN model

The inputs applied on the input layer is passed to the HL by a multiplication process using the weight with each attribute. The nodes in HL were designed to compute the weighted sum with a bias value represented as:

$$I_j = \sum_{i}^{m} i_i * b_{ij} + \beta_j \tag{7}$$

Where, i_i signifies the input data, β indicates bias function and b_{ij} represents the weighted link among the nodes. Therefore, I_j is transformed with the help of a sigmoid transfer function:

$$F\left(I_j\right) = \frac{1}{1 + e^{-I_j}} \tag{8}$$

During network training, weights are adjusted according to the following equation.

$$\Delta w_{ij} = -n \frac{\beta E}{\partial w_{ij}} \tag{9}$$

Where, E denotes the error and *n* is the learning rate. Finally, the weight in DNN is optimized with the help of an algorithm known as Harris hawk optimization (HHO).

HHO is considered to be a gradient-free optimization approach. With the right formulation, it is able to solve any optimization problem (Heidari, 2019). The Harris hawks are the candidate solutions in HHO, and the best solution of candidate is referred to as intended prey. They are perch at random areas in HHO and use two strategies to wait for prey to appear (Abdel-Basset, 2021). They perch depending on the locations of other family members and the rabbit, as modeled in Eq. (1) for the condition of $q < 0.5$, or perch on random tall trees, as modeled in Eq. (1) for the condition of $q \geq 0.5$.

$$g(t+1) = \begin{cases} g_{rabbit}(t) - g_m(t) - r_3(AB + r_4(MB - NB) & q < 0.5 \\ g_{rand}(t) - r_1 \left| g_{rand} - 2r_2\, g(t) \right| & q \geq 0.5 \end{cases} \tag{10}$$

This equation can be used to optimize the output of DNN. The following figure represents the flowchart of the proposed methodology.

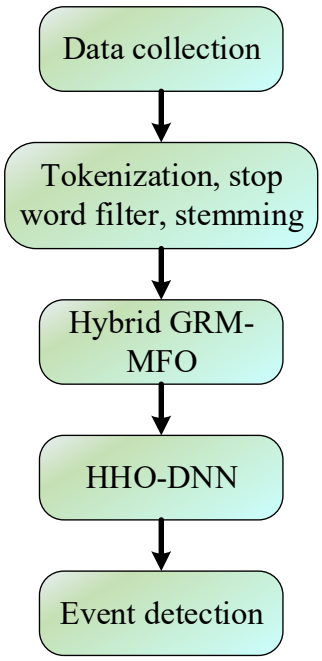

Figure 3: Flowchart of proposed methodology

The flow chart of the proposed event detection technique using big data is as shown in the figure 3.as the figure represents, data collection is the initial step. After that, the given data is forwarded to the pre-processing stage. The pre-processing is carried out with the help of three main processes such as tokenization, stop word filter and stemming. The next step is the feature extraction. It is done with the help of a hybrid technique termed as GRM-MFO. The next stage is the feature classification stage. This step is done with the help of HHO-DNN method. And the last stage is the event detection stage.

4. Result and discussion

The result and discussion of the proposed methodology is explained in this section along with dataset description and comparison results.

4.1 Dataset description

Here the proposed method is analyzed using big data for event detection. Here we are combining five dataset for the purpose. The detailed description about the datasets used here is explained in this section.
Education – (https://data.world/cityofchicago/performance-metrics-city-colleges-of-chicago-course-s): The dataset contains success metrics in the college of Chicago city. The percentage of students who received A, C or fail is referred to as course success rate.
Transportation – (https://data.world/buffalony/y93c-u65y): The GBNRTC and its member agencies provide traffic count data as a public service in this dataset. Manual and mechanical methods are used to capture traffic data, which is then stored at a separate location.
Environment –(https://data.world/dublin-city): This dataset contains the results of annual traffic counts conducted at 33 locations across the city cordon created by the Royal and Grand Canals from 2008 to 2012 at 15 minute intervals. Every year, during the month of November, the counts are carried out. Counts are normally done twice on two different days at each venue, with the average of the two counts used
Geospatial – (https://data.world/dcopendata/03fcb333c5a3441ab9823116b4359c4a-4): This dataset shows 311 Service Requests in the last 30 days for illegal dumping on public property.
Water – (https://data.world/city-of-ny/bkwf-xfky): The data tables summarize the turbidity values, coli form, fluoride and chlorine found at sides in distribution.

Here we combine these five dataset for the evaluation of proposed event detection process. Here we use the concatenation method for combining the five datasets. It is a method by which combining two or more datasets side by side, and treated as a single dataset. The total data consider here for the evaluation is 74270. From this, 80% data is used for training and 20% data is used for testing.

4.2 Evaluation metrics

The performance of the proposed method is measured in terms of statistical measures such as classification accuracy, precision, F1 score and recall. The following equations represents the expressions of the accuracy,

precision and recall (Mesaros, 2010, August). Accuracy is defined as the value close to the true value which is given by the following expression.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

Precision can be defined as the ratio of total number of positive samples that is classified in to total number of samples. It is computed using the following expression.

$$precision = \frac{TP}{TP+FP} \tag{12}$$

Recall can be defined as the ratio of number of positive samples classified as positive to total number of positive samples and can be computed using the following expression.

$$recall = \frac{TP}{TP+FN} \tag{13}$$

F-Measure generates a single score that accounts for both accuracy and recall concerns in a single number. It is also known as F1-score. The equation is used to calculate the value.

$$F\ measure = 2 * \left[\frac{precision * recall}{precision + recall} \right] \tag{14}$$

In these equations TN and TP indicates the number of true negative and true positive values. The FN and FP represents the false negative and false positive values

4.3 Experimental results

In this section, the results of the proposed method and also the comparison results with the existing method is described. Here we are analyzing and detecting the events with the help of big data and the performance is evaluated in statistical parameters such as accuracy, precision, recall and f-score. The following table represents the outcomes of the proposed event detection approach.

Table 1: Performance analysis of proposed technique

Parameters	**Events**				
	Education	**Transportation**	**Environment**	**Geospatial**	**Water**
Accuracy (%)	94	92	98	96	96.5
Precision (%)	93	91.5	97	95.5	95.5
Recall (%)	95	93	99.9	96	98.9
F1 score (%)	93.5	92.5	97.8	96.5	97

Table represents the performance evaluation of the proposed method. As the table represents, the proposed method shows better performance on the five dataset such as education, transportation, environment, geospatial and water. The parameters taken here for the evaluation process are precision, accuracy, recall and F1 score. The graphical representation of proposed method is shown in the following figure.

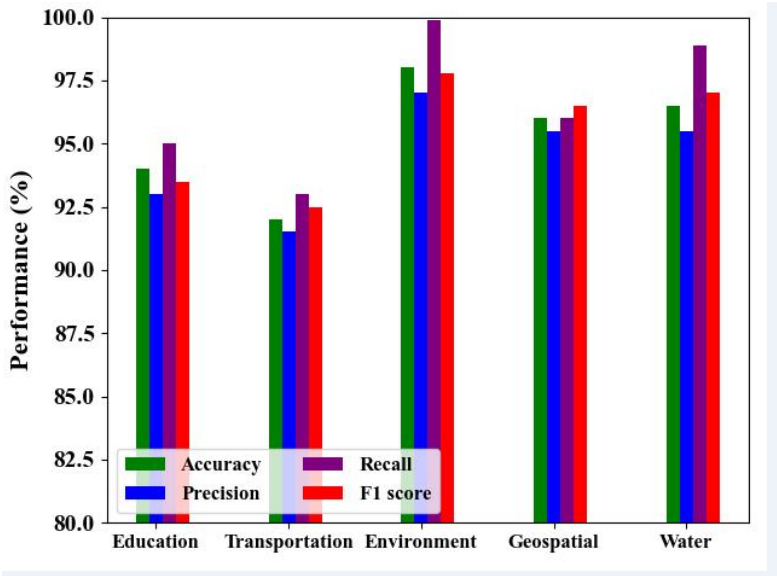

Figure 4: Performance of the proposed method

The figure depicts the outputs of the proposed method in terms of accuracy, precision, recall and F1 score. As the figure depicts, the proposed method detects five events such as education, transportation, environment, geospatial and water, which are indicated by the X axis. The corresponding performance is indicated by the Y axis. The figure is depicts based on table 1. The following figure 3 represents the accuracy comparison graph.

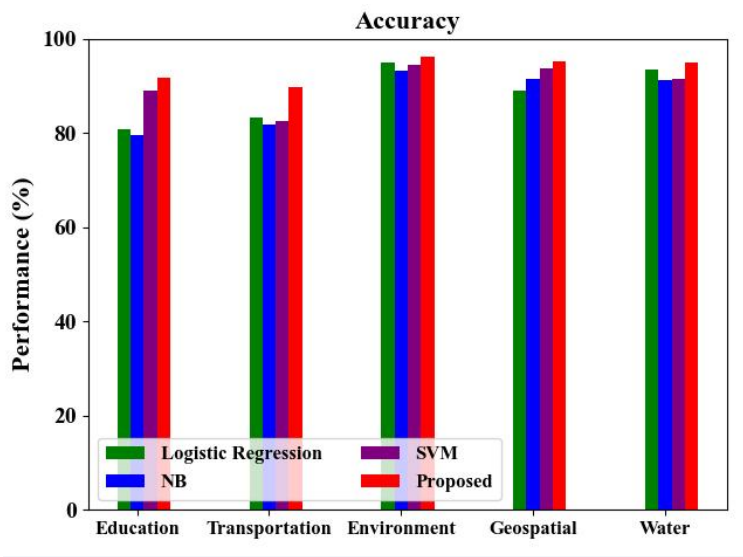

Figure 5: Accuracy comparison

Figure 5 represents the accuracy comparison of proposed method with some of the existing methods such as logistic regression, Naïve Bayes and SVM (Pandhare and Shah 2017, March). As the figure depicts, the outcomes of the suggested method is better for every dataset than the existing methods. In the figure, X axis represents the events and Y axis represents the Accuracy. The following figure 6 represents the precision comparison graph.

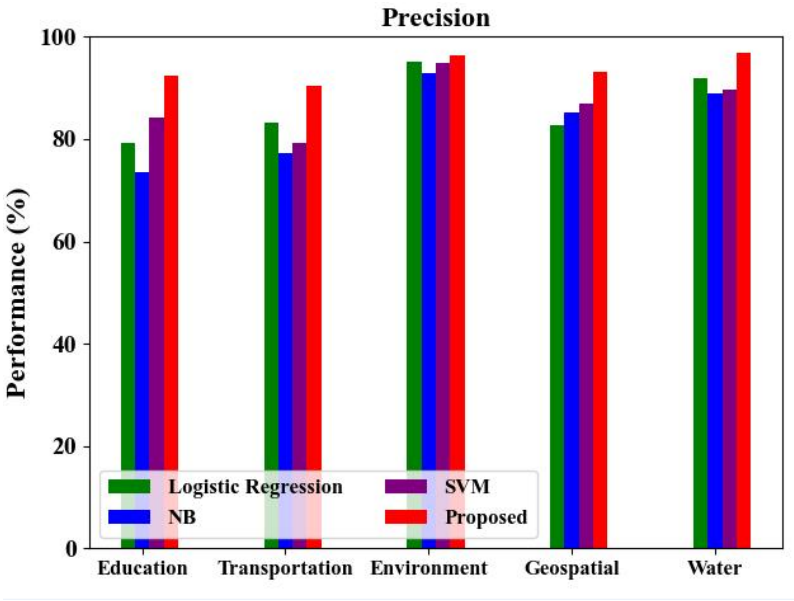

Figure 6: Precision comparison

Figure 6 represents the precision comparison graph of the proposed method with some existing method. It is clear from the figure that, the proposed method shows better performance for every dataset. The precision value of proposed method for the event education is 93%, transportation is 91.5%, environment is 97%, geospatial is 95.5% and for water is 95.5%. It is very much better than the all other existing methods. The following figure represents the recall comparison graph.

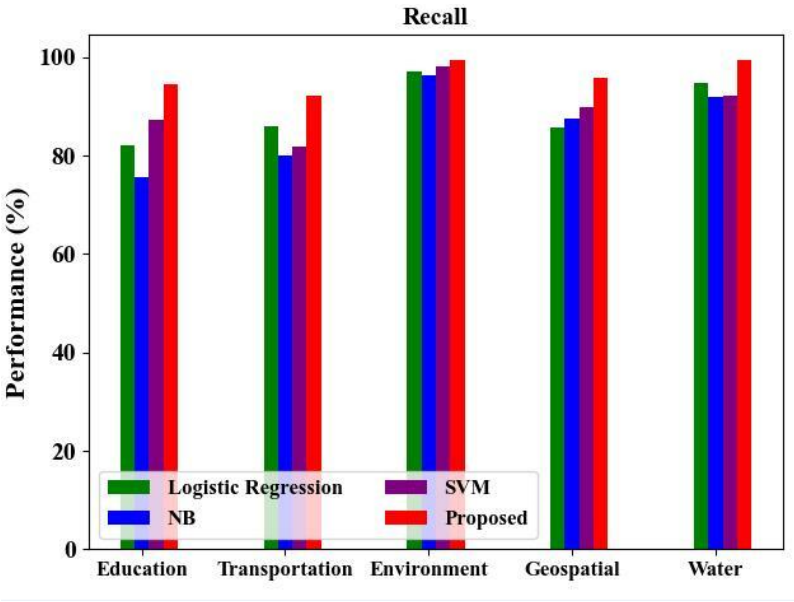

Figure 7: Recall comparison

Figure 7 represents the recall comparison graph of the proposed method with existing methods. As the figure depicts the recall values for the education, transportation, environment, geospatial, and water dataset are 95%, 93%, 99.9%, 96% and 98.9%. This means, the recall value of the suggested approach shows very good performance than the existing methods. The following figure represents the F1 score comparison graph.

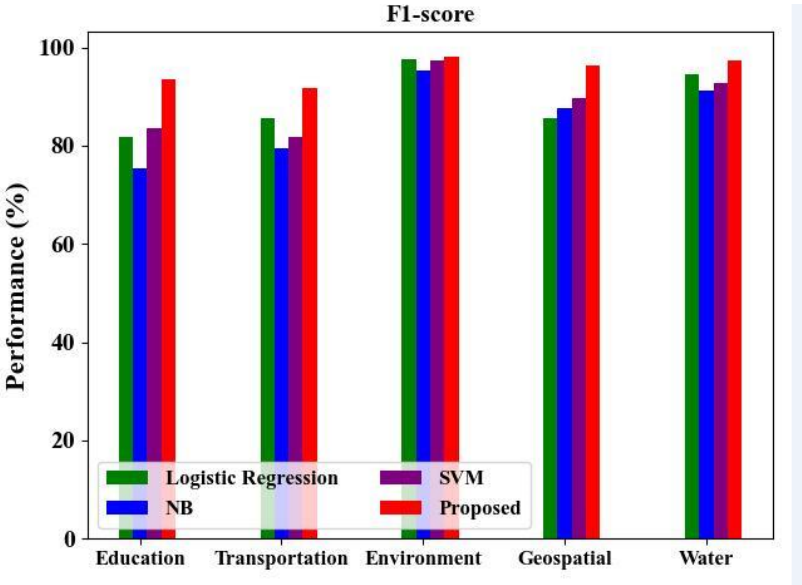

Figure 8: F-measure comparison

Figure 8 represents the F-measure comparison graph. In the figure, X axis represents the different events and the methods and the Y axis represents the F1 score performance. The F1 score of education dataset is 93.5%, transportation is 92.5%, environment is 97.8%, geospatial is 96.5% and for water is 97%. Thus, it can be observed that the proposed event detection technique shows better performance in F1 score also.

5. Conclusion

For the event detection in twitter, various methods have been analyzed in this paper. Twitter is an online platform used by millions of users for different communication purpose. Therefore, the event detection process is carried out here for the twitter data. Many research works had been done in the field of crime detection due to the benefits of application of social media platform. The conventional methods, that are used for the prediction of crimes based on twitter data has some limitations. Some methods for crime detection do not provide enough security and privacy to the users. Hence, to overcome those limitations, DNN is incorporated with HHO algorithm for the classification of events. The accuracy of the proposed technique for different dataset such as education, transportation, environment, geospatial and water is 94%, 92%, 98%, 96% and 96.5% respectively. Also, the proposed method was compared with some of the existing algorithms. The overall results obtained showed that the proposed approach is far better than other existing or conventional algorithms. In future, we extend this work using more efficient dataset as well as efficient optimization method to achieve further robustness of event detection process.

Acknowledgments
None

Conflicts of interest
The authors have no conflicts of interest to declare.

Funding
No funders were contributed for the preparation of the manuscript.

References

[1] Abdel-Basset, M.; Ding, W.; and El-Shahat, D. (2021). A hybrid Harris Hawks optimization algorithm with simulated annealing for feature selection. Artificial Intelligence Review, **54**(1), pp. 593-637.
[2] Aliwy, A.H. (2012). Tokenization as preprocessing for arabic tagging system. International Journal of Information and Education Technology, **2**(4), pp. 348.
[3] Alomari, E.; Katib, I.; and Mehmood, R. (2020). Iktishaf: A big data road-traffic event detection tool using Twitter and spark machine learning. Mobile Networks and Applications, pp. 1-16.
[4] Bolón-Canedo, V.; Sánchez-Maroño, N.; and Alonso-Betanzos, A. (2015). Recent advances and emerging challenges of feature selection in the context of big data. Knowledge-Based Systems, **86**, pp. 33-45.
[5] Daniel, B.K. (2019). Big Data and data science: A critical review of issues for educational research. British Journal of Educational Technology, **50**(1), pp. 101-113.
[6] Heidari, A.A.; Mirjalili, S.; Faris, H.; Aljarah, I.; Mafarja, M.; and Chen, H. (2019). Harris hawks optimization: Algorithm and applications. Future generation computer systems, **97**, pp. 849-872.
[7] Mesaros, A.; Heittola, T.; Eronen, A.; and Virtanen, T. (2010, August). Acoustic event detection in real life recordings. In 2010 18th European Signal Processing Conference, IEEE, pp. 1267-1271.

[8] Mirjalili, S. (2015). Moth-flame optimization algorithm: A novel nature-inspired heuristic paradigm. Knowledge-based systems, **89**, pp. 228-249.
[9] Nie, F.; Yang, S.; Zhang, R.; and Li, X. (2018). A general framework for auto-weighted feature selection via global redundancy minimization. IEEE Transactions on Image Processing, **28**(5), pp. 2428-2438.
[10] Pandhare, K.R.; and Shah, M.A. (2017 March). Real time road traffic event detection using Twitter and spark. In 2017 International conference on inventive communication and computational technologies (ICICCT), IEEE, pp. 445-449.
[11] Pandhare, K.R.; and Shah, M.A. (2017, March). Real time road traffic event detection using Twitter and spark. In 2017 International conference on inventive communication and computational technologies (ICICCT), IEEE, pp. 445-449.
[12] Phan, H.; Krawczyk-Becker, M.; Gerkmann, T.; and Mertins, A. (2017). DNN and CNN with weighted and multi-task loss functions for audio event detection. arXiv preprint arXiv:1708.03211.
[13] Ramos, J. (2003, December). Using tf-idf to determine word relevance in document queries. In Proceedings of the first instructional conference on machine learning, **242**(1), pp. 29-48.
[14] Reiz, A.N.; de la Hoz, M.A.; and García, M.S. (2019). Big data analysis and machine learning in intensive care units. Medicina Intensiva (English Edition), **43**(7), pp. 416-426.
[15] Saeed, Z.; Abbasi, R.A.; Maqbool, O.; Sadaf, A.; Razzak, I.; Daud, A.; Aljohani, N.R.; and Xu, G. (2019). What's happening around the world? a survey and framework on event detection techniques on twitter. Journal of Grid Computing, **17**(2), pp. 279-312.
[16] Salas, A.; Georgakis, P.; Nwagboso, C.; Ammari, A.; and Petalas, I. (2017, July). Traffic event detection framework using social media. In 2017 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC), IEEE, pp. 303-307.
[17] Takahashi, N.; Gygli, M.; Pfister, B.; and Van Gool, L. (2016). Deep convolutional neural networks and data augmentation for acoustic event detection. arXiv preprint arXiv:1604.07160.
[18] Tao, D.; Yang, P., and Feng, H. (2020). Utilization of text mining as a big data analysis tool for food science and nutrition. Comprehensive Reviews in Food Science and Food Safety, **19**(2), pp. 875-894.
[19] Wan, J.; Zheng, P.; Si, H.; Xiong, N.N.; Zhang, W.; and Vasilakos, A.V. (2019). An artificial intelligence driven multi-feature extraction scheme for big data detection. IEEE Access, **7**, pp. 80122-80132.
[20] Wang, D.; Nie, F.; and Huang, H. (2015). Feature selection via global redundancy minimization. IEEE transactions on Knowledge and data engineering, **27**(10), pp. 2743-2755.
[21] Wang, T.; Bhuiyan, M.Z.A.; Wang, G.; Rahman, M.A.; Wu, J.; and Cao, J. (2018). Big data reduction for a smart city's critical infrastructural health monitoring. IEEE Communications Magazine, **56**(3), pp. 128-133.
[22] Xu, T.; and Overbye, T. (2015, November). Real-time event detection and feature extraction using PMU measurement data. In 2015 IEEE International Conference on Smart Grid Communications (SmartGridComm), pp. 265-270.
[23] Zhang, Y.; Jin, R.; and Zhou, Z.H. (2010). Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics, **1**(1-4), pp. 43-52.

Authors Profile

K Swapnika pursuing Ph. D in Data Mining and Information Retrieval Systems stream at Jawaharlal Nehru Technological University, Hyderabad Hyderabad, she completed M. Tech in Software Engineering from Jawaharlal Nehru Technological University Hyderabad and has 8 years of academic experience. Her Research Interest includes Information Retrieval Systems and Bigdata Analytics.

Dr. D. Vasumathi is a Professor and HOD of Computer Science and Engineering Department in JNTUH College of Engineering, J. N. T. University, and Hyderabad. She has completed her PhD at J. N. T. University, Hyderabad in 2011 and has more than 20 years of experience in Teaching and Research. She is guiding 09 PhD scholars in Computer Science and Engineering, and she is also Vice-President of National ST Employees and Officers Welfare Association (NST & OWA), and General Secretary for Teaching Association (NECTA), in JNTUH college of Engineering. She is Finance Secretary for both TS & AP States Tribal Development Association, TDA- Hyderabad.

a