An Ensemble Framework to analyze the Crimes in social media by using Deep Learning Algorithm

Swathi.Alluri[1], Mr.M.Krishna Satya Varma[2]
[1]*Student of IT Department, Sagi Rama Krishna Raju Engineering College, Bhimavaram, AP, India*
[2]*Assistant Professor, Department of Information Technology, SRKR College, Bhimavaram, AP, India*

Abstract **- Nowadays crime is seeing everywhere. Social networks are one of the platforms for the crimes that are occurring everywhere. In social networks huge data is generating based on the daily news, personal, entertainment etc? Various social media crimes can be seen such as Cyber stalking, Cyber bullying, Cyber Hacking, Cyber Harassment, and Cyber Scam using the data obtained from social media website. In this proposed framework we mainly focus on finding the social media crimes on twitter data. This process consists of three stages such as data (tweet) pre-processing, classifying model builder and prediction. Various existing systems are already implemented on twitter datasets and predicted the results. Here, An Ensemble Framework to Analyze the Crimes in social media by Using Deep Learning Algorithm is introduced to increase the performance.**

Index Terms **- Cyber stalking, Cyber bullying, Cyber Hacking, Cyber Harassment.**

1.INTRODUCTION

Nowadays social media plays the major role in communicating with the people. Many people are using social networking sites (SNS) such as Facebook, twitter because of their popularity. Everyday multiple users are registering in these websites to communicate with new friends. People are showing more interest on spending time on social media. Many companies collect the data from SNS to analyze the user's data to predict their interests. These predicted data is used to start new companies such as start-ups. On the other side these data is misused by the many of the hackers by doing illegal activities. Exchanging several types of data between the users such as opinions, news, personal information such as profile pictures, videos, and text can be shared by using this online platform. These networks will help us to understand the opinions on various public talk and also personal. SNS shows the impact on several opinions and also this shows the huge impact on elections in all over the world. Nowadays these platforms used for the terrorists to expand their activities to all over the world.

By using the previous data from the SNS, these reports different types of crimes. For example, several fake persons creating fake profiles with the name of popular persons and demanding money for the different purposes. Some people are using this platforms as a medium for exchanging their messages with code language. Some people share pictures of orphans and demanding money to solve the problems of orphans. All these are belongs to the crimes in the SNS. These types of crimes are reporting regularly in the twitter and facebook.

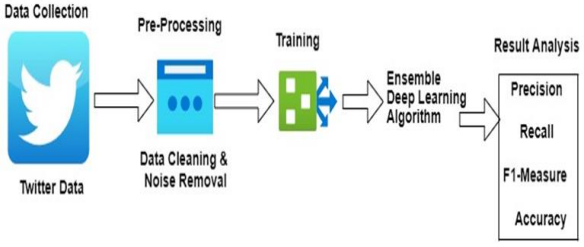

Figure 1: System Architecture

From the past many years, SNS spreading hate messages. These hate messages are mainly belongs to religion, cast, personal and political parties. Hate messages and speeches mainly degrades the person in society. From the United Nations the Rabat plan of action [1] defines the rules to differentiate among normal speech and hate speech. In this paper, an ensemble deep learning algorithm is introduced to classify the different crimes that occur within the given datasets. This algorithm is based on three steps such as data (tweet) pre-processing, classifying model builder and prediction.

2.LITERATURE SURVEY

The author Wei et al. [2] proposed the ML based algorithm that finds the huge conversations on twitter. Several features are observed for finding the abnormal behavior on twitter based on tweets given by the users by utilizing the KNN classifier. The author [3] conducted experiments on detecting the abnormal behavior using ML methods such as Naive Bayes algorithm (NBA). This one of the better algorithm that shows the best results among other ML classifiers. Based on the classical features several abnormal behaviors is detected by using traditional algorithms. The classification is developed by using positive and negative sentiments by showing the abnormal behavior with external groups. The major drawback of this approach is to take all the dependencies belongs to a words present in document. Hence it is observed that ML algorithms are does not provide an efficient way for classification of normal and abnormal.

Hartung et al. [4] proposed an AI manner for figuring out radical posts in German Twitter accounts. Various highlights are tested, like feelings, etymological examples, and published hints. The framework yielded progressed effects over the modern works. Studies on grouping radical affiliations almost about on-line media content material are likewise perceptible in illicit medicine utilization. as an example, of their paintings on Maryjane-associated microblogs, Nguyen et al. [5] amassed in greater than 30,000 tweets regarding weed all through 2016. The content material mining technique offers a few precious stories to the acquired statistics, for instance, (I) consumer disposition are regularly taken care of as certain or bad, (ii) over 65% of tweets are commenced from mobileular phones, and (iii) recurrence of tweets at the give up of the week is above specific days.

Ryan et al. [6], through presenting a totally precise technique hooked into grammatical function labeling and evaluation-pushed identity of enthusiast reporters from internet discussions. The exam relied on round 1 million posts from greater than 25,000 unmistakable customers browsing on 4 radical discussions. The proposed approach relied on the consumer's evaluation rating, registered through collecting the rating of no. of bad posts, a span of bad posts, and seriousness of bad posts. The framework is adaptable to distinguish on-line doubtful physical games on radical customers.

In 2012, Chalothorn and Ellman [7] proposed an evaluation exam version to break down on-line innovative posts utilising various lexical assets, for instance, SentiWordNet, WordNet, and NLTK toolbox. the feeling magnificence and strength of the content material are registered. At first, published statistics changed into received from diverse internet gatherings like Montada and Qawem, and after acting essential pre-managing assignments, various component-pushed measures have been implemented to identify and manage strict and radicals content material. Test effects display that the Montada amassing has greater sure posting than the Qawem dialogue. it have been reasoned that Qawem's dialogue is skilled greater excessive postings.

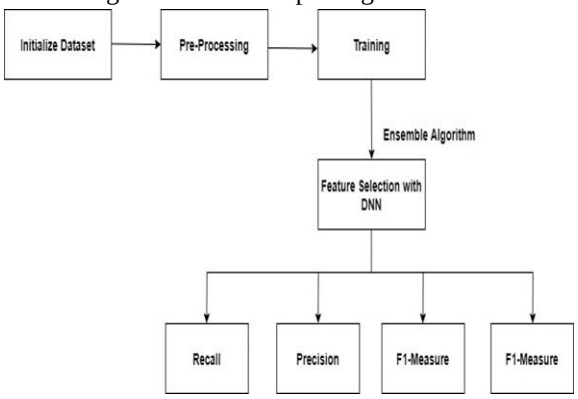

Figure 2: Steps for applying ensemble deep learning algorithm

3.STESPS FOR ENSEMBLE ALGORITHM

3.1 Initialize Dataset

To create a deep learning model, the first thing we required is a dataset as a deep learning model completely works on data. The collected data for a particular problem in a proper format is known as the dataset. Dataset may be of different formats for different purposes, such as, if we want to create a deep learning model for business purpose, then dataset will be different with the dataset required for a social networking dataset.

3.1.1 Data Pre-processing

This step plays major role in analyzing the data present in the tweeter. To process this data for experiments several steps have to be conducted to remove the noise from the collected data. This shows the impact on output. The data which is collected for experiments will have the unstructured data and also the noise data. This step removes all such type of data to increase the data.

3.1.2 Data Cleaning: Data cleaning is one the important step in cleaning data. This removes the corrupt and irrelavent data from the records from the tables or dataset columns. This database refers to finding the incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data.

3.1.3 Noise Removing:
Noisy data are data that is corrupted, or distorted, or has a signal-to-noise ratio. Improper procedures (or improperly-documented procedures) to subtract out the noise in data can lead to a false sense of accuracy or false conclusions.
Data = true signal + noise.

3.1.4 Training
Training in the ML plays the significant role to learn the crime or hate messages. Analyzing crime messages are entirely different by analyzing normal messages. Algorithms such as supervised learning algorithms, every analysis of data consists of output variables and multiple numbers of input variables.

3.1.5 Dataset description
The dataset collected from kaggle website and it is collection of several tweets that are belongs to the users. This consists of 2000 tweets belongs to various crimes. Analyzing few crimes that occur within the given text which are in the form of tweets.

3.1.6 Algorithm Steps:
Ensemble Deep Learning Algorithm The algorithm is combination of deep neural networks with feature selection.
Step: 1 the social networking dataset is initialized.
Step: 2 Pre-processing is done and noise filters are ready.
Step: 3 Training on raw data.
Step: 4 Apply algorithms, Ensemble deep Learning Algorithm.
Step: 5 detect the crime analyzed messages.
Step: 6 Show results for the given input.
Step: 7 Parameters such as precision, recall and accuracy are shown.

4.PERFORMANCE METRICS

The evaluation metrics include accuracy, F1 score, precision, recall.
Accuracy: This is very important parameter that initializes the total number of exactly classified data instances over the total number of data instances.

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

Precision: When the FP is high the precision helps. If the precision is low, then the patients will be told that they are affected with lung infection; this may show some mistakes within the tests.

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall: Recall is calculated when the false negatives are high.

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1 Measure} = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$$

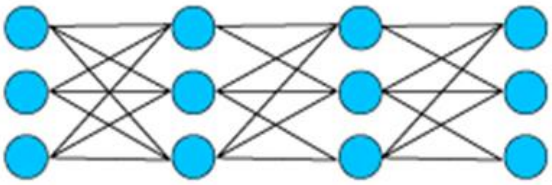

Figure 3: Combination of DNN and Feature Selection

Algorithms	KNN	SVM	EDLA
Precision	87.6	88.2	98.3
Recall	88.1	89.3	97.2
F1-Score	86.2	90.1	96.2
Accuracy	88.32	92.1	97.8

Table 1: comparative results

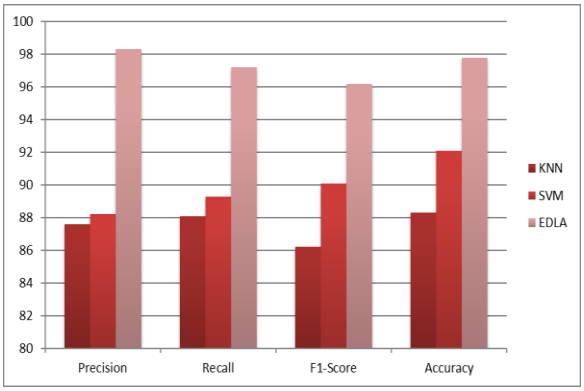

Figure 4: graph representation of comparative algorithms.

5.CONCLUSION

Detecting the crime tweets from the twitter related datasets is most widely used in many applications. The entire world analyzing the abnormal behavior with tweets is more tedious because large data is generating every day. Crimes are doing in different ways, by doing the crimes using twitter is most widely following by the several users. Analyzing crime tweets manually is more complex, to overcome this, the ensemble algorithm is applied on twitter dataset by analyzing the patterns of the text given in the twitter dataset. The algorithm is the combination of deep neural networks (DNN) and Feature Selection which increases the accuracy and performance.

REFERENCES

[1] Office of the United Nations High Commissioner for Human Rights. Report of the United Nations High Commissioner for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred. Office of the United Nations High Commissioner for Human Rights; Geneva, Switzerland: 2013.

[2] Wei Y, Singh L, Marti S (2016) Identification of extremism on Twitter. Proceedings of the IEEE/ACM international conference on advances in social networks analysis and mining. IEEE, New Jersey, pp 1251–1255.

[3] Azizan SA, Aziz IA (2017) Terrorism detection based on sentiment analysis using machine learning. J Eng Appl Sci 12(3):691–698.

[4] Hartung M, Klinger R, Schmidtke F, Vogel L (2017) Identifying right-wing extremism in german Twitter profiles: a classification approach. International conference on applications of natural language to information systems. Springer, Cham, pp 320–325.

[5] Nguyen A, Hoang Q, Nguyen H, Nguyen D, Tran T (2017) Evaluating marijuana-related tweets on Twitter. IEEE 7th annual computing and communication workshop and conference (CCWC). IEEE, New Jersey, pp 1–7.

[6] Ryan S, Garth D, Richard F (2018) Searching for signs of extremism on the web: an introduction to sentiment-based identification of radical authors. Behav Sci Terror Pol Aggres 10:39–59.

[7] Chalothorn T, Ellman J (2012) Using SentiWordNet and sentiment analysis for detecting radical content on web forums.