Computer Science and Information Technologies
Vol. 4, No. 2, July 2023, pp. 149~159
ISSN: 2722-3221, DOI: 10.11591/csit.v4i2.p149-159

An ensemble approach for the identification and classification of crime tweets in the English language

Tooba Siddiqui[1], Saman Hina[1], Raheela Asif[1], Saad Ahmed[2], Munad Ahmed[3]
[1]Department of Computer Science, N.E.D. University Karachi, Karachi, Pakistan
[2]Department of Computer Science, IQRA University Karachi, Karachi, Pakistan
[3]Research Department, MSN360.pk, Karachi. Pakistan

Article Info

Article history:

Received Nov 22, 2022
Revised May 15, 2023
Accepted Jun 10, 2023

Keywords:

Classification
Crime tweets
Ensemble approach
Natural language processing
Twitter

ABSTRACT

Twitter is a famous social media platform, which supports short posts limited to 280 characters. Users tweet about many topics like movie reviews, customer service, meals they just ate, and awareness posts. Tweets carrying information about some crime scenes are crime tweets. Crime tweets are crucial and informative and separate classification is required. Identification and classification of crime tweets is a challenging task and has been the researcher's latest interest. The researchers used different approaches to identify and classify crime tweets. This research has used an ensemble approach for the identification and classification of crime tweets. Tweepy and Twint libraries were used to collect datasets from Twitter. Both libraries use contrasting methods for extracting tweets from Twitter. This research has applied many ensemble approaches for the identification and classification of crime tweets. Logistic regression (LR), support vector machine (SVM), k-nearest neighbor (KNN), decision tree (DT), and random forest (RF) Classifier assigned with the weights of 1,2,1,1 and 1 respectively ensemble together by a soft weighted Voting classifier along with term frequency – inverse document frequency (TF-IDF) vectorizer gives the best performance with an accuracy of 96.2% on the testing dataset.

Corresponding Author:

Saad Ahmed
Department of Computer Science, IQRA University Karachi
Karachi, Pakistan
Email: saadahmed@iqra.edu.pk

1. INTRODUCTION

Social media is playing a central role in modern life. Online social media, such as Twitter, Facebook, and many enterprises social media, have become very popular in the last few years. People spend a huge amount of time on social media to interact with people. The number of people who use social media is increasing day by day. Twitter has millions of users and it is one of the biggest platforms for users to share their thoughts, feelings, opinion, and ideas. Text available on Twitter is expanding drastically. Unlike other social media platforms, almost all user tweets are public and extractable. If you're trying to get a large amount of information to perform analytic tasks, then Twitter is the best option. Each tweet on Twitter is about some specific topic. Twitter's application programming interface (API) allows you to create complex questions and analyze them like what are the trending topics on social media by extracting the latest tweets, or customer reviews about some XYZ Company by collecting many tweets that talk about your company and applying an analysis algorithm to it. Crime is an act that is prohibited and punishable by law. It is dangerous not only to the victim but also to the whole community. Every tweet is about specific topics, like movie reviews, customer service, awareness posts, and more. Few tweets are about robbery, murder, abduction, and

other criminal activities. Such tweets are called Crime Tweets. Crime tweets are crucial and informative and should be separately available for all users to view. And it is also helpful for police and other civil authorities. They can find recent criminal activities. Police can also figure out sensitive cities and areas with the help of these tweets. Also, these tweets bring awareness to all Twitter users. On average, tweets posted on Twitter are 500 million per day. Manually extracting crime tweets from these bulk tweets is quite tiresome. Therefore, a separate classifier for the identification and classification of crime tweets is required. In previous research, authors have used different approaches to identify and classify crime tweets including machine learning [1]–[3], multiclass multilevel classifiers [4], and artificial neural networks [5]. The target of this research is to find and build a model that can identify and classify tweets into two categories: crime tweets and non-crime tweets. To make this classifier, this research has used ensemble approaches for the identification and classification of crime tweets. The ensemble method is a machine learning technique that combines several ML models to produce one optimal predictive model. Initially, the researchers evaluated and selected a machine-learning algorithm for the ensemble approach. This research has applied multiple ensemble approaches, including a voting classifier, overall local accuracy (OLA) classifier, adaptive boosting (AdaBoost), extra tree classifier, bagging, light gradient boosted machine (LGBM) classifier, category boosting (CatBoost) classifier, and extreme gradient boosting (XGBoost) classifier to build the best ensemble classifier for the identification and classification of crime tweets.

Crime Tweets carry information about some crime scenes like robbery, abduction, and murder. Crime tweets are very crucial and separate classification is required. In recent years, researchers worked on the identification and classification of crime tweets classification. Lal *et al.* [1] applied several machine learning algorithms for the identification and classification of crime tweets. Researchers collected 500 tweets manually, comprising 230 non-crime tweets and 270 crime tweets posted on a particular Twitter account. Ahmed *et al.* [6] is another research that has used the TF-IDF vectorizer for sentiment analysis. For classification, this research has applied many machines learning algorithms, including Naive Bayesian, random forest, J48, and ZeroR. Random forest outperforms other machine learning algorithms with 98.1% accuracy. Vomfell *et al.* [2] focused on improving the forecasting of crime count with the help of tweets and taxi datasets. For the big dataset, Naıve Bayes outperforms with the highest accuracy of 94.82%. Shoeibi *et al.* [3] research related to tweets categorization into crime-related tweets and not crime-related tweets. This research has extracted 3,200 tweets. In this research, tweets went through two major steps: topic classification and aspect-based sentiment analysis. The support vector machine (SVM) model with TF-IDF vectorizer performs better with 88.89% accuracy. Santhiya *et al.* [4] focused on finding crime geological predictions on the basis of tweets. They used Twitter Search API for the extraction of tweets. The total dataset comprised 1,48,707 tweets, which were categorized into different categories, including sexual harassment, rape, dowry death, kidnapping, abduction, stalking, groping, and suicide. This research used a multiclass, multi-level Naïve Bayes (NB) classifier and gained 82% of accuracy in identifying the location. [5] categorizes crime tweets into assault, burglary, drugs violations, homicide, and sex offences using the artificial neural network (ANN) approach. 100,000 tweets were collected to conduct this research. The neural network approach outperformed with 90.33% accuracy.

Numerous types of research have been carried out on the Twitter dataset. But unfortunately, enough researches are not available to review the identification and classification of crime tweets. Therefore, other research based on Twitter dataset classification having a close resemblance to the identification and classification of crime tweets is also being reviewed. In [7] and [8] detected Malicious accounts and suspicious messages on Twitter. Pakaya *et al.* [7] uses machine learning to detect malicious accounts based on Tweet account features. This classification assumes spam bots and fake followers fall into a greater classification of malicious accounts. In this research, the best model with 95.55% accuracy for the binary classification scheme is on XGBoost with TF-IDF features. AlGhamdi and Khan [8] analyzed Arabic tweets to detect suspicious messages. The Dataset comprises 1,555 tweets, out of which 826 tweets are suspicious, and 729 are not suspicious. SVM outperforms which yields 86.72% mean accuracy.

In [9] detected abusive text on the basis of abusive and non-abusive words. This research used unsupervised learning and achieved 94.15% accuracy. In [10], [11], and [12] categorized news-related datasets into fake news and real news. Hakak *et al.* [10] used the decision tree (DT) classifier, random forest algorithm, and extra tree (ET) ensemble together for fake news classification and give 99.6% accuracy on the training dataset and 44.15% accuracy on the testing dataset. Malla and Alphonse [11] created a new model which detects fake COVID-19 tweets with an accuracy of 98.88%. Ahmad *et al.* [12] uses 4 different datasets and detected fake news among those datasets. In DS1, a random forest algorithm achieved an accuracy of 99%. On DS2, the bagging classifier (decision trees) and boosting classifier (XGBoost) are the best-performing algorithms, achieving an accuracy of 94%. In DS3, the benchmark algorithm (Perez-LSVM) achieved an accuracy of 93.5. In DS4, the best-performing algorithm is random forest (91% accuracy). Sembodo *et al.* [13] classified news tweets into 11 categories, namely religion, business, entertainment, law,

health, motivation, sports, government, education, politics, and technology. This research has collected and labeled 4,230 tweets. It also applied many machine learning algorithms, whereas Naive Bayes multinomial gives the highest accuracy of 77.47%. In [14]–[16], and [17] worked for Hate and offensive speech classification. Taradhita and Putra [14] uses convolutional neural network (CNN) Classifier for hate speech classification in Indonesian language Tweets. CNN with 100 epochs gives the best accuracy of 88.34%. Swamy *et al.* [15] uses an ensemble approach to identify hate speech. L1-regularised logistic regression, L2 - regularized logistic regression, linear support vector classifier (SVC), stochastic gradient descent (SGD), and passive-aggressive (PA) ensemble together by voting classifier gives the best performance. Fauzi and Yuniarti [16] used an ensemble approach for Indonesian hate tweets. Where, the voting classifier, an ensemble of the three best classifiers outperforms (Naıve Bayes, support vector machine, and random forest) with the F1 measure of 79.8%. Febriany and Utama [17] focused on identifying negative posts on social media using machine learning algorithms. K-nearest neighbors (K-NN) gives the highest accuracy of 99.85%. In [18] and [19] worked on spam detection using an ensemble approach. Ahraminezhad *et al.* [18] and other authors proposed a new algorithm for the detection of spam which outperforms with an accuracy of 91.77%. Saeed *et al.* [19] detects spam in Arabic opinion text. The stacking ensemble classifier achieves maximum accuracy values of 95.25% by integrating the outputs of the rule-based classifier with the K-means classifier. Ansari *et al.* [20] analyzed the political sentiments on Twitter. The random forest with TF-IDF uni-gram exhibits the highest precision of 77%. Research (including [21], [22], and [23] and inter alia) shows that the ensemble approach has a high tendency to outperform the machine learning algorithms for the identification and classification of tweets. In contrast to the research work, this research has applied different settings of the ensemble approach to the dataset for the identification and classification of crime tweets.

2. DATA COLLECTION AND PRE-PROCESSING

In this research, data collection and Pre-processing were completed in two different stages. The researcher used Python language to conduct this research. The reason behind choosing Python language over other languages is that Python is a better choice for machine learning and large-scale applications, especially for data analysis within web applications. In the first step, this research extracted the dataset that was used for the identification and classification of crime tweets. After that, it cleaned the dataset before applying word embedding techniques. Thus, data collection and pre-processing are sub-divided into the following two tasks:

- Dataset extraction and annotation
- Dataset cleaning

2.1. Data extraction and annotation

In natural language processing (NLP) research, data collection is a crucial task. To extract tweets from Twitter, different researcher uses different methods. Many researchers used the Twitter dataset (tweets) available on open sources (including [11], [24], [25], and [26]. Whereas some researchers extracted tweets manually (including [1], [27]) to proceed with their research. To research the identification and classification of crime tweets, no open-source Twitter dataset is available specific to crime tweets. As an alternative to manual extraction, automatic extraction techniques are also available to extract tweets from Twitter, including Twitter application programming interface (API) and BeautifulSoup6. BeautifulSoup is used to extract datasets from many microblogging platforms like Twitter, and Facebook, and some researchers (including [28]) also used Beautiful Soup to extract tweets from Twitter. Whereas in lots of research, authors used Twitter API for tweets extraction to conduct their research (including [29]–[31], and [32]). Many services that rely on data, including Facebook and Twitter, have APIs, which help the developers and researchers to interact with their dataset and also allow them to extract information including posts (tweets) from their database. Setty *et al.* [33] also allows them to create and post new information into their database. Facebook has an API (Rest FB Java APIs) that many researchers (including [33]) have used to extract the Facebook dataset. and they are usually pretty easy to work with and pretty straightforward than other services like BeautifulSoup. Initially, this research has Twitter Figure 1 heat map showing a correlation between different feature APIs to extract tweets from Twitter. To access the dataset from Twitter API, a developer account is mandatory. Therefore, the researchers applied for the developer account, and, after its approval, credentials from API were available to use. Later, to access the tweets, a library was required to connect to Twitter API to proceed to the extraction of tweets. There are a couple of different libraries that this research has used to access the Twitter API, but because of its easy-to-use environment, Tweepy was used to extract tweets from Twitter. Tweepy has many methods to extract tweets from Twitter. Although during this research, only two methods were used to extract tweets using tweepy,

- Tweets extraction based on a specific hashtag
- Tweets extracted from an individual user

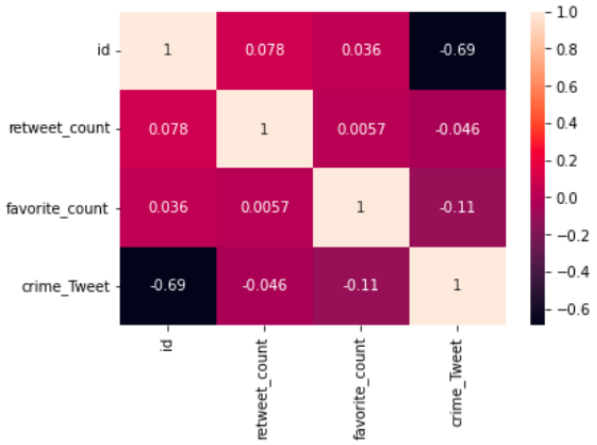

Figure 1. Heat map showing a correlation between different feature

The dataset extracted during the data collection stage comprises 6,483 tweets including 3,186 crime tweets and 3,297 non-crime tweets. Initially, this dataset contains some duplicate tweets. This research used usernames and keywords, both methods for the extraction of tweets. During the extraction of tweets by the keyword method, researchers used keywords like robbed, dead, captured, arrested, abducted, kill, police, suspect, steal, and charge. Whereas during extracting tweets using a specific username method, researchers collected tweets from different users that contains tweets related to crime. During the extraction of the dataset, this research also extracted some duplicate tweets from Twitter, which were removed later in the cleaning stage. After extracting all tweets, the authors manually labeled all the extracted tweets into crime tweets and non-crime tweets.

2.2. Data cleaning

It is necessary to remove unnecessary information from the Twitter extracted data, before training the classification model. This research has used Tweepy and Twint libraries to extract datasets from Twitter and information extracted by these two libraries contains some unnecessary information that has to be removed before training the dataset. The structure of the dataset extracted from these two libraries was different, so first unnecessary information existing in the dataset was analyzed by using data analysis techniques like heat maps as shown in Figure 1. By using these data analysis techniques, all unnecessary information that was useless for the identification and classification of crime tweet were removed. This research is focused on tweets in the English language, so all the tweets in other languages existing are removed from the dataset. Afterward, this research proceeded to clean the tweets in the dataset. These extracted tweets contain noise inside them, which includes web links, hashtags, non-English tweets, stop words, duplicate tweets, audio/video tags, and much more. For cleaning the dataset used for the identification and classification of crime tweets, given normalization steps were taken:

- Irrelevant features were removed from the dataset.
- All tweets in a language other than English were removed from the dataset.
- Duplicate entries of the tweets in the dataset were removed.
- The text of all tweets in the dataset was converted into lowercase.
- web links, retweets, @user information, hashtags, and AUDIO/VIDEO tags were removed using the Python libraries from the dataset.
- punctuation, double space, and numbers were also removed from the dataset.
- Tokenization was performed to get the tokenized representation of words in the dataset
- Stop words were removed from the dataset.
- An empty string in the dataset was removed.
- WordNet Lemmatization was applied on each token to retrieve the meaning of the text and these tokens were saved in the dataset.

At the end of the preprocessing, information containing meaningful tokens for each tweet was stored in the dataset. With the help of these tokens, the dataset was further analyzed using some data analysis techniques like finding the most occurring words in the dataset for crime tweets using word cloud and bar plots as shown in Figures 2 and 3. During cleaning and preprocessing, all duplicate and unnecessary information was removed and the resultant dataset comprise 6,457 unique tweets including 3,177 crime

tweets and 3,280 non-crime tweets as shown in Figure 4. Later, word embeddings were applied to transform the dataset into numerical vectors.

Figure 2. Word cloud for crime tweets

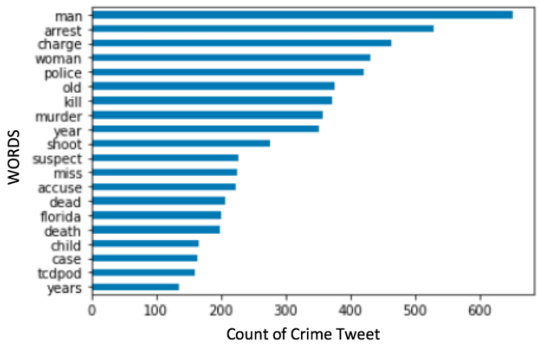

Figure 3. Bar plot between most common words and their count for crime tweets

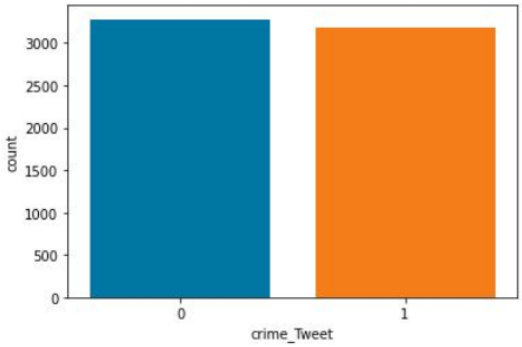

Figure 4. Count plot of crime and non-crime tweets

3. METHODOLOGY

After the cleaning of the dataset, the research has to go through some more stages to build the classifier for the identification and classification of crime tweets. From the research work, it was found that on the Twitter dataset TF-IDF and Hashing vectorizer, both word embeddings were used in different research to transform the text of the tweets into vectors. Thus, this research has applied these two techniques one by one on both machine learning and ensemble approaches to find out which one of these two-word embedding is better. Also, many machine learning algorithms are applied to find the best-performing algorithms suitable for ensemble approaches. Multiple ensemble approaches were applied to the dataset to find out the best ensemble approach that gives the highest accuracy for crime tweet classification.

3.1. Word embedding

Word embedding is a technique used for the representation of text using vectors. It helps to extract information from the patterns formed inside data. Many word embedding techniques can be used to get vector representation for their textual dataset including bag of words, TF-IDF, fast-text, and hashing vectorizer. Among these techniques, TF-IDF is the most commonly used word embedding technique applied on Twitter. Many researchers including [1], [14], [15], [20], and [32] used the TF-IDF technique on their Twitter data. Whereas in a few research (including [27]) hashing vectorizer is used as a word embedding technique on preprocessed tweets. This research has used these two different approaches for document representation, TF-IDF and hashing vectorizer, to figure out which is more suitable for the identification and classification of crime tweets.

3.1.1. TF-IDF vectorizer

Term frequency-inverse document frequency (TF-IDF) is used in machine learning and text mining as a weighting factor for features. The gist is that the weight increases as the word frequency in a document increases. This means the weight increases the more times a term occurs in the document, but that's offset by

the number of times the word appears in the entire dataset or corpus. This offset helps remove the importance from really common words like 'the' or 'a that appear frequently across all documents. To generate values for each word in TF-IDF, first, the term frequency is calculated, then inverse document frequency is calculated for each word and lastly, the product of these represents the value of the word inside the vector generated for a document d as shown in (1), (2) and (3) respectively:

$$\text{Tf} = \frac{Number\ of\ times\ a\ word\ occur\ in\ a\ document\ d}{2aTotal\ Number\ of\ words\ in\ a\ document\ d} \quad (1)$$

$$Idf = \log \frac{Total\ number\ of\ documents}{2aNumber\ of\ documents\ in\ which\ that\ word\ exist} \quad (2)$$

$$Tf.Idf = Tf * Idf \quad (3)$$

This research has applied a TF-IDF vectorizer on the clean dataset extracted for the identification and classification of crime tweets. Later, this dataset was split into training and testing datasets. Machine learning algorithms were applied to the training dataset and then its accuracy was evaluated on the testing dataset. Results describing the performance of each machine learning algorithm in identifying and classifying crime tweets with TF-IDF vectorizer are shown in Table 1.

Table 1. Performance of machine learning algorithms with TFIDF

Machine Learning Algorithm with TF-IDF Vectorizer	Accuracy	F-score
Logistic Regression	95.9%	96%
Support Vector Classifier	95.7%	96%
K- nearest neighbors	91.6%	92%
Decision Tree	90.7%	91%
Random Forest	92.2%	92%
Naive Bayes	87.0%	87%

In this research, six machine learning algorithms including logistic regression, support vector machine, random forest classifiers, K-nearest neighbors, Naïve Bayes, and decision tree classifier were applied with TF-IDF vectorizer on the dataset for the identification and classification of the crime tweets and performance of each classifier was evaluated. It was found that Logistic regression outperformed the rest of the machine learning algorithms and it gives the best performance with an accuracy of 95.9% on the testing dataset. Whereas, support vector machine (SVM) performed very well with an accuracy of 95.7% on the testing dataset. Whereas, Random Forest classifiers, K-nearest neighbors, and decision trees were also applied for the identification and classification of crime tweets and produced a model giving an accuracy of 92.2%, 91.6%, and 90.7% accuracy respectively. It was found that the Naive Bayes classifier with TF-IDF vectorizer gives the worst performance with the lowest accuracy of 87%.

3.1.2. Hashing vectorizer

Hashing vectorizer is another technique used for feature extraction of textual data. It is designed to generate a vectorizer for the text that is as memory efficient as possible. Instead of storing the tokens as strings, the vectorizer applies the hashing trick to encode them as numerical indexes. The downside of this method is that once text is vectorized, the words can no longer be retrieved. This research has applied Hashing vectorizer on the clean dataset extracted for the identification and classification of crime tweets. Later, this dataset was split into training and testing datasets. Machine learning algorithms were applied to the training dataset and then its accuracy was evaluated on the testing dataset. Results describing the performance of each machine learning algorithm in identifying and classifying crime tweets with hashing vectorizer are shown in Table 2.

In this research, six machine learning algorithms including logistic regression, support vector machine, random forest classifiers, K-nearest neighbors, Naïve Bayes, and decision tree classifiers were applied with hashing vectorizer on the dataset for the identification and classification of the crime tweets and performance of each classifier was evaluated. It was found that logistic regression outperformed the rest of the machine learning algorithms and it gives the best performance with an accuracy of 93.5% on the testing dataset. Whereas, support vector machine (SVM) performed very well with an accuracy of 92.6% on the testing dataset. Whereas, random forest classifiers, K-nearest neighbors, and decision trees were also applied for the identification and classification of crime tweets and produced a model giving an accuracy of 92.4%, 91%, and 89% accuracy respectively. It was found that the Naïve Bayes classifier with hashing vectorizer gives the worst performance with the lowest accuracy of 73.9%.

It is observed that except for random forest algorithm, TF-IDF helped machine learning algorithms in building a classifier with better accuracy than hashing vectorizer for the identification and classification of the crime tweets. Also, for random forest, both classifiers performed equally well. Similarly, among various ensemble approaches applied during this research, the Catboost classifier's accuracy obtained with both TFIDF and Hashing vectorizer was also equal. Whereas for the rest of the ensemble approaches applied during this research, the TF-IDF vectorizer outperformed hashing vectorizer and produced far better accuracy in each approach for the identification and classification of the crime tweets.

Table 2. Performance of machine learning algorithms with hashing vectorizer

Machine Learning Algorithm with TF-IDF Vectorizer	Accuracy	F-score
Logistic Regression	93.5%	93%
Support Vector Classifier	92.6%	93%
K- nearest neighbors	91.0%	91%
Decision Tree	89.0%	89%
Random Forest	92.4%	92%
Naive Bayes	73.9%	74%

3.2. Ensemble approaches

The Ensemble approach combines individual models to improve the stability and predictive power of the model. This approach permits higher predictive performance compared to a single model. The ensemble approach finds ways to combine multiple machine learning models into one predictive model to decrease variance, decrease bias, or improve predictions. This research has applied many ensemble approaches with TF-IDF vectorizer as well as hashing vectorizer on the clean dataset and their results describing their performance in identifying and classifying crime Tweets are shown in Table 3.

Table 3. Performance of ensemble approach with TF-IDF and hashing vectorizer on a preprocessed crime Tweets dataset

Ensemble Approach	ML Classifiers (If used)	TF-IDF Vectorizer		Hashing Vectorizer	
		Accuracy	F-score	Accuracy	F-score
Hard wgt. voting (2,2,2,1)	LR+SVM+KNN+DT	96.1%	96%	93.7%	94%
Soft wgt. voting (1,2,1,1,1)	LR+SVM+KNN+DT+RF	96.2%	96%	93.2%	93%
OLA Classifier	LR+SVM+KNN	95.8%	96%	94.6%	95%
AdaBoost	-	91.6%	92%	90.4%	90%
Bagging	-	92.56%	93%	91.4%	91%
ExtraTree	-	95.8%	96%	94.3%	94%
LightGBM	-	93.8%	94%	92.5%	93%
CatBoost	-	91.6%	92%	91.9%	92%
XGBosst	-	93.1%	93%	92.7%	93%

For the identification and classification of tweets, the ensemble approach has been used by many researchers ([10], [15], [18], [19], [34], and inter alia) and in most of this research, ensemble approach has given a better performance than machine learning algorithm. This research has applied multiple ensemble approaches, including a voting classifier (both hard and soft), overall local accuracy (OLA) classifier, adaptive boosting (AdaBoost), extra tree classifier, bagging, light gradient boosted machine (LGBM) classifier, category boosting (CatBoost) classifier and extreme gradient boosting (XGBoost) classifier to build the best ensemble classifier for the identification and classification of crime tweets. Voting classifier is a famous ensemble approach that combines various machine learning algorithms and makes predictions by evaluating the aggregate of the decision taken by each of the machine learning algorithms. In the case of the weighted voting classifier, various weights are assigned to each of these machine learning algorithms, and based on their weights, the decision varies. The voting classifier can take a biased decision if one algorithm has a big value of weight assigned to it. Different weights are assigned to each of the machine learning algorithms based on their performance on the dataset. Multiple researchers applied a voting classifier to the Twitter dataset. Swamy *et al.* [15] has applied it for the identification and categorization of offensive language. Saeed *et al.* [19] used it for spam detection and Fauzi and Yuniarti [16] used it for Hate speech detection. Whereas [27] and [32] applied it in sentiment analysis. In this research, voting classifier outperformed the rest of the ensemble approaches. Logistic regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), decision tree (DT), and random forest (RF) Classifier assigned with the weights of 1,2,1,1 and 1 respectively ensemble together by a soft weighted Voting classifier along with TF-IDF vectorizer gives the best performance with an accuracy of 96.2% on the testing dataset. Whereas, logistic regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), and decision tree (DT)

Classifier assigned with the weights of 2,2,2 and 1 respectively ensemble together by a hard weighted Voting classifier along with TF-IDF vectorizer performed very well with an accuracy of 96.1% on the testing dataset. The major difference between hard and soft voting classifier is that hard voting classifier takes the label and weight of each algorithm and evaluate their aggregate to predict the outcome whereas soft voting classifier takes probabilities instead of the label along with the weight and evaluate their aggregate to predict the outcome This research also applied overall local accuracy (OLA) Classifier for the identification and classification of the crime tweets. It evaluates the competence level of each Machine learning algorithm combined in the OLA Classifier and chooses one algorithm based on their competence level, to make the prediction. During this research, logistic regression (LR), support vector machine (SVM), and K-nearest neighbor (KNN) ensemble together by OLA classifier with TF-IDF vectorizer also give outstanding performance with an accuracy of 95.8% on the testing dataset.

ExtraTree Classifier combines random numbers of decision trees based on a training dataset and prediction is made by combining all the predictions taken from each decision tree. This research has applied ExtraTree Classifier for the identification and classification of the crime tweets and gives a good performance with accuracy with a TF-IDF vectorizer of 95.8% on the testing dataset. Light gradient boosted machine (LGBM) classifier is another ensemble approach that was applied for the identification and classification of crime tweets. It is also a tree-based approach. Light gradient boosted machine (LGBM) classifier with TF-IDF vectorizer gives a good performance with an accuracy of 93.8% on the testing dataset. extreme gradient boosting (XGBoost) classifier is another boosting algorithm applied in this research for the identification and classification of crime tweets. XGBoost classifier gives a good performance with TF-IDF vectorizer with an accuracy of 93.1% on the testing dataset. Bootstrap aggregation (bagging) classifier is another ensemble approach. Its focus is on minimizing the variance estimator by changing the settings of the machine learning algorithms, combined inside the bagging classifier. This research has applied bagging classifier for the identification and classification of the crime tweets and gives a good performance with TF-IDF vectorizer with an accuracy of 92.56% on the testing dataset. Category boosting (CatBoost) Classifier is a gradient boosting algorithm that combines oblivious decision trees. It gives the quickest predictions and gives a good performance in multiple categories of the classification problem. This research has applied CatBoost classifier for the identification and classification of the crime tweets and it gives a good performance with Hashing vectorizer with an accuracy of 91.9% on the testing dataset. Adaptive boosting (AdaBoost) classifier is a well-known iterative ensemble approach that tends to give good performance with weak learning classifiers. The researcher including [21] and [32] also used this approach for textual sentiment classification. This research has applied AdaBoost Classifier for the identification and classification of the crime tweets and it gives a good performance with the TF-IDF vectorizer with an accuracy of 91.6% on the testing dataset.

During this research, it was found that the AdaBoost classifier with hashing vectorizer gives the worst performance with the lowest accuracy of 90.4%. From Table 3, it was also observed that the CatBoost classifier's accuracy obtained with both TF-IDF and Hashing vectorizer is equal. Whereas for the rest of the ensemble approaches applied during this research, the TF-IDF vectorizer outperformed hashing vectorizer and produces good accuracies in the identification and classification of the crime tweets.

4. DISCUSSION AND RESULTS

This research was to build a classifier using an ensemble approach for the identification and classification of the crime tweets comprising four stages. Firstly, this research has used two different libraries for extraction of the dataset as there is no existing tweets dataset specific to crime tweets. Secondly, it used two different vectorizers, namely TF-IDF vectorizer and hashing vectorizer, for feature extraction. Thirdly, researchers applied many machine learning algorithms to find the best-performing algorithms suitable for ensemble approaches. Finally, the research has applied many ensemble approaches to find the best ensemble classifier for the identification and classification of crime tweets.

For data collection, this research has used twint and tweepy Python libraries. Both libraries use contrasting methods for extracting tweets from Twitter. Tweepy needs Twitter API and developer account credentials and it is capable of collecting comparatively a richer dataset. Whereas Twint does not need any API. Twitter developer account verification is a time-consuming process and sometimes gets rejected. In case of rejection, a developer can use twint to extract a dataset from Twitter, but twint is slower than tweepy. Although both methods have their pros and cons and are well-suited techniques for the extraction of tweets.

For feature extraction, this research has applied two different techniques and applied machine learning algorithms to them. Results for machine learning algorithm were mentioned in Tables 1 and 2. From these tables, it is clear that except for the random forest.

Classifier, TF-IDF performed better than hashing vectorizer in all machine learning algorithms. Also except for the CatBoost classifier, in the rest of the ensemble approach, TF-IDF gives far better performance than hashing vectorizer. Whereas in both, the random forest classifier and CatBoost classifier, TF-IDF and Hashing vectorizer has performed equally well.

Before applying ensemble approaches, it was required to find machine learning algorithms that give good performance on a clean dataset with both TF-IDF and hashing vectorizer. From Tables 1 and 2, it is clear that Machine learning algorithms that give good performance with both TF-IDF and hashing vectorizer were logistic regression, support vector classifier, random forest, K-nearest neighbors, and decision tree. From Tables 1 and 2, it is also clear that logistic regression and support vector classifier are overall best-performing algorithms for the identification and classification of the crime tweet.

For the identification and classification of the crime tweets, this research has applied many ensemble approaches including weighted voting classifiers (hard and soft voting), overall local accuracy (OLA) classifier, AdaBoost classifier, bagging classifier, extratree classifier, LightGBM Classifier, CatBoost classifier, and XGBoost classifier with TF-IDF as well as hashing vectorizer. The performance of these ensemble approaches for the identification and classification of the crime tweets are given in Table 3. The Soft weighted voting classifier has outperformed all classifiers with an accuracy of 96.2% on the testing dataset. Logistic regression (LR), support vector machine (SVM), K-nearest neighbor (KNN), decision tree (DT) and random forest (RF) Classifier assigned with the weights of 1,2,1,1 and 1 respectively ensemble together by a soft weighted Voting classifier along with TF-IDF vectorizer gives the best performance with an accuracy of 96.2% on the testing dataset. This classifier was also tested manually and correctly identified and classified each test tweet as a crime or non-crime tweet as shown in Figure 5.

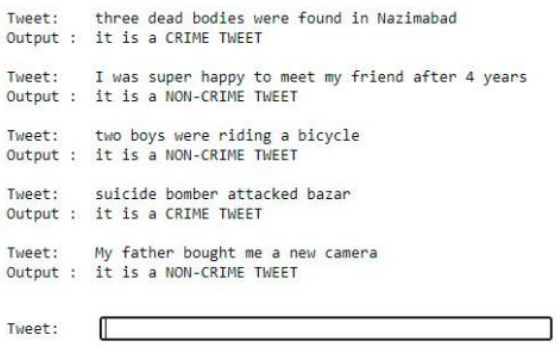

Figure 5. Demonstration of crime tweet classifier

5. CONCLUSION

Crime tweets are important, as it carries information about robbery, kidnapping, criminal escape, and other crime incidents. That is why a classifier for the identification and classification of crime tweets is needed that can easily help in identifying and classifying crime and non-crime tweets. Crime tweet has a variety of applications like they bring awareness among people. These tweets can also be used by civil authorities to take necessary actions. Many researchers have also used these crime tweets along with the text dataset for the prediction of crime counts. Several researchers worked on the identification and classification of the crime tweet using different machine learning techniques. In the study of numerous types of research work done earlier, it was found that ensemble approaches are more likely to give better performance for textual classification than any machine learning algorithms. Therefore, this research has applied an ensemble approach to the identification and classification of crime tweets. Out of all applied ensemble approaches, soft voting stands out as the best ensemble approach and it also outperforms all the tried machine learning algorithms with an accuracy of 96.2%. This classifier was also tested manually and correctly identified and classified each test tweet as a crime or non-crime tweet. This research can be extended by classifying these crime tweets using some deep learning technique or some unsupervised model. Also, different word embedding techniques like a bag of words can be applied to figure out any better feature-extracting techniques that give better results than the TF-IDF vectorizer. This research can also be extended by further classifying crime tweets into different categories of crime mentioned in each crime tweet. Also, the same research can be carried out in languages other than English, like Urdu tweets, for the identification and classification of the same crime and non-crime tweets.

REFERENCES

[1] S. Lal, L. Tiwari, R. Ranjan, A. Verma, N. Sardana, and R. Mourya, "Analysis and classification of crime tweets," *Procedia Computer Science*, vol. 167, pp. 1911–1919, 2020, doi: 10.1016/j.procs.2020.03.211.

[2] L. Vomfell, W. K. Härdle, and S. Lessmann, "Improving crime count forecasts using Twitter and taxi data," *Decision Support Systems*, vol. 113, pp. 73–85, 2018, doi: 10.1016/j.dss.2018.07.003.

[3] N. Shoeibi, N. Shoeibi, G. Hernández, P. Chamoso, and J. M. Corchado, "Ai-crime hunter: An AI mixture of experts for crime discovery on twitter," *Electronics (Switzerland)*, vol. 10, no. 24, 2021, doi: 10.3390/electronics10243081.

[4] K. Santhiya, V. Bhuvaneswari, and V. Murugesh, "Automated crime tweets classification and geo-location prediction using big data framework," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 14, pp. 2133–2152, 2021.

[5] S. P. C. . Sandagiri, B. T. G. . Kumara, and B. Kuhaneswaran, "Detecting crime related twitter posts using artificial neural networks based approach," in *2020 20th International Conference on Advances in ICT for Emerging Regions (ICTer)*, Nov. 2020, pp. 5–10, doi: 10.1109/ICTer51097.2020.9325485.

[6] S. Ahmed, S. Hina, E. Atwell, and F. Ahmed, "Aspect based sentiment analysis framework using data from social media network," *International Journal of Computer Science and Network Security*, vol. 17, no. 7, pp. 100–105, 2017.

[7] F. N. Pakaya, M. O. Ibrohim, and I. Budi, "Malicious account detection on twitter based on tweet account features using machine learning," *Proceedings of 2019 4th International Conference on Informatics and Computing, ICIC 2019*, 2019, doi: 10.1109/ICIC47613.2019.8985840.

[8] M. A. AlGhamdi and M. A. Khan, "Intelligent analysis of Arabic tweets for detection of suspicious messages," *Arabian Journal for Science and Engineering*, vol. 45, no. 8, pp. 6021–6032, 2020, doi: 10.1007/s13369-020-04447-0.

[9] H. S. Lee, H. R. Lee, J. U. Park, and Y. S. Han, "An abusive text detection system based on enhanced abusive and non-abusive word lists," *Decision Support Systems*, vol. 113, pp. 22–31, 2018, doi: 10.1016/j.dss.2018.06.009.

[10] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta, and W. Z. Khan, "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021, doi: 10.1016/j.future.2020.11.022.

[11] S. J. Malla and P. J. A. Alphonse, "Fake or real news about COVID-19? Pretrained transformer model to detect potential misleading news," *European Physical Journal: Special Topics*, vol. 231, no. 18–20, pp. 3347–3356, 2022, doi: 10.1140/epjs/s11734-022-00436-6.

[12] I. Ahmad, M. Yousaf, S. Yousaf, and M. O. Ahmad, "Fake news detection using machine learning ensemble methods," *Complexity*, vol. 2020, 2020, doi: 10.1155/2020/8885861.

[13] J. E. Sembodo, E. B. Setiawan, and M. A. Bijaksana, "Automatic tweet classification based on news category in Indonesian language," *2018 6th International Conference on Information and Communication Technology, ICoICT 2018*, pp. 389–393, 2018, doi: 10.1109/ICoICT.2018.8528788.

[14] D. A. N. Taradhita and I. K. G. D. Putra, "Hate speech classification in Indonesian language tweets by using convolutional neural network," *Journal of ICT Research and Applications*, vol. 14, no. 3, pp. 225–239, 2021, doi: 10.5614/itbj.ict.res.appl.2021.14.3.2.

[15] S. D. Swamy, A. Jamatia, B. Gambäck, and A. Das, "NIT_Agartala_NLP_Team at SemEval-2019 task 6: An ensemble approach to identifying and categorizing offensive language in twitter social media corpora," *NAACL HLT 2019 - International Workshop on Semantic Evaluation, SemEval 2019, Proceedings of the 13th Workshop*, pp. 696–703, 2019, doi: 10.18653/v1/s19-2124.

[16] M. A. Fauzi and A. Yuniarti, "Ensemble method for Indonesian Twitter hate speech detection," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 11, no. 1, pp. 294–299, 2018, doi: 10.11591/ijeecs.v11.i1.pp294-299.

[17] A. Febriany and D. N. Utama, "Analysis model for identifying negative posts based on social media," *International Journal of Emerging Technology and Advanced Engineering*, vol. 11, no. 10, pp. 96–103, 2021, doi: 10.46338/IJETAE1021_12.

[18] A. Ahraminezhad, M. Mojarad, and H. Arfaeinia, "An intelligent ensemble classification method for spam diagnosis in social networks," *International Journal of Intelligent Systems and Applications*, vol. 14, no. 1, pp. 24–31, 2022, doi: 10.5815/ijisa.2022.01.02.

[19] R. M. K. Saeed, S. Rady, and T. F. Gharib, "An ensemble approach for spam detection in Arabic opinion texts," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 1, pp. 1407–1416, 2022, doi: 10.1016/j.jksuci.2019.10.002.

[20] M. Z. Ansari, M. B. Aziz, M. O. Siddiqui, H. Mehra, and K. P. Singh, "Analysis of political sentiment orientations on Twitter," *Procedia Computer Science*, vol. 167, pp. 1821–1828, 2020, doi: 10.1016/j.procs.2020.03.201.

[21] M. M. R. Mamun, O. Sharif, and M. M. Hoque, "Classification of textual sentiment using ensemble technique," *SN Computer Science*, vol. 3, no. 1, 2022, doi: 10.1007/s42979-021-00922-z.

[22] Ankit and N. Saleena, "An ensemble classification system for Twitter sentiment analysis," *Procedia Computer Science*, vol. 132, pp. 937–946, 2018, doi: 10.1016/j.procs.2018.05.109.

[23] Z. H. Kilimci and S. Akyokus, "Deep learning- and word embedding-based heterogeneous classifier ensembles for text classification," *Complexity*, vol. 2018, 2018, doi: 10.1155/2018/7130146.

[24] A. Khakharia, V. Shah, and P. Gupta, "Sentiment analysis of COVID-19 vaccine tweets using machine learning," *SSRN Electronic Journal*, 2021, doi: 10.2139/ssrn.3869531.

[25] A. Sarker, M. S. U. Zaman, and M. A. Y. Srizon, "Twitter data classification by applying and comparing multiple machine learning techniques," *SSRN Electronic Journal*, 2020, doi: 10.2139/ssrn.3509207.

[26] Y. Chen, W. Hou, X. Cheng, and S. Li, "Joint learning for emotion classification and emotion cause detection," *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*, pp. 646–651, 2018, doi: 10.18653/v1/d18-1066.

[27] N. F. F. Da Silva, E. R. Hruschka, and E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles," *Decision Support Systems*, vol. 66, pp. 170–179, 2014, doi: 10.1016/j.dss.2014.07.003.

[28] A. Sarker, M. R. Islam, and A. Y. Srizon, "A comprehensive pre-processing approach for high-performance classification of Twitter data with several machine learning algorithms," *2020 IEEE Region 10 Symposium, TENSYMP 2020*, pp. 630–633, 2020, doi: 10.1109/TENSYMP50017.2020.9230590.

[29] J. Samuel, G. G. M. N. Ali, M. M. Rahman, E. Esawi, and Y. Samuel, "COVID-19 public sentiment insights and machine learning for tweets classification," *Information (Switzerland)*, vol. 11, no. 6, 2020, doi: 10.3390/info11060314.

[30] K. A. Abdullah, S. O. Folorunso, O. O. Solanke, and S. M. Sodimu, "A predictive model for tweet sentiment analysis and classification," *Anale. Seria Informatica*, vol. XVI, 2018.

[31] O. AlZoubi, S. K. Tawalbeh, and M. AL-Smadi, "Affect detection from arabic tweets using ensemble and deep learning techniques," *Journal of King Saud University - Computer and Information Sciences*, vol. 34, no. 6, pp. 2529–2539, 2022, doi: 10.1016/j.jksuci.2020.09.013.

[32] A. Alessa and M. Faezipour, "Tweet classification using sentiment analysis features and TF-IDF weighting for improved flu trend detection," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10934 LNAI, pp. 174–186, 2018, doi: 10.1007/978-3-319-96136-1_15.
[33] S. Setty, R. Jadi, S. Shaikh, C. Mattikalli, and U. Mudenagudi, "Classification of facebook news feeds and sentiment analysis," *Proceedings of the 2014 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2014*, pp. 18–23, 2014, doi: 10.1109/ICACCI.2014.6968447.
[34] S. J. Malla and A. P.J.A., "COVID-19 outbreak: An ensemble pre-trained deep learning model for detecting informative tweets," *Applied Soft Computing*, vol. 107, 2021, doi: 10.1016/j.asoc.2021.107495.

BIOGRAPHIES OF AUTHORS

Tooba Siddiqui is a master's holder in Computer Science & Information Technology from NED University of Engineering & Technology, and completed her bachelor's in engineering in the field of Computer & Information System, also faculty of National Center of Big Data & Cloud Computing (NED University) where she works as Research Associate and handles NED OBE MIS System which facilitates teaching faculty of NED University. She has also worked as Analyst Software Engineer at IBEX Global (TRG). Her research interest covers the topic of Natural Language Processing, Computer Vision, Machine Learning, Deep Learning and Web Development. She can be contacted at email: toobas.siddiqui@yahoo.com.

Saman Hina is an Associate Professor in the Department of Computer Science IT at NED University of Engineering and Technology, Pakistan where she is serving since 2006. She received her Master's from the same University and completed her Ph.D. from the University of Leeds, the UK in 2013. Her Ph.D. research was in AI in general and Natural Language Processing in particular. Other research interests include areas of Artificial Intelligence and its applications in interdisciplinary domains. She can be contacted at email: samhaq@neduet.edu.pk.

Raheela Asif received her PhD degree in Computer Science & I.T. in 2017 from NED University of Engineering & Technology, Karachi, Pakistan. She is currently serving as Associate Professor in the Department of Software Engineering at NED University. Her research interests include Data Mining and Learning Analytics, Data Bases and Artificial Intelligence. She can be contacted at email:

Saad Ahmed received MS. degree in Computer Science from Hamdard University, Karachi Pakistan, in 2012 and a Ph.D. degree in Computer Science from the NED University of Engineering and Technology Karachi Pakistan 2019. He currently works as an assistant professor at the Department of Computer Science, IQRA University Karachi Pakistan. His current research interests include Natural Language Processing, Data mining, and Big Data Analysis. He can be contacted at email: saadahmed@iqra.edu.pk (Corresponding author).

Munad Ahmed received Master's degree in Management Science from Federal University, Karachi Pakistan and completed Bachelor's degree in Computer Science from SAL University Pakistan. He currently works as a Research Associate at the Department of Research & Development, MSM360.pk Karachi Pakistan. His current research interests include Data mining, ERP and Data Analysis. He can be contacted at email: munadahmed@hotmail.com.