Web Mining To Detect Online Spread of Terrorism

J. RACHANA
Assistant Professor
Department of CSE
Rachanajakkula@tkrec.ac.in
Teegala Krishna Reddy Engineering College, Hyderabad

BEKKAM SOUSIL KUMAR
Department of CSE
sousilkumarbekkam@gmail.com
Teegala Krishna Reddy Engineering College, Hyderabad

K.GIPSON
Department of CSE
Gipsongideon1483@gmail.com
Teegala Krishna Reddy Engineering College, Hyderabad

N.JAHNAVI
Department of CSE
Jahnavi2000n@gmail.com
Teegala Krishna Reddy Engineering College,Hyderabad

G. PALLAVI
Department of CSE
Pallavir390@gmail.com
Teegala Krishna Reddy Engineering College,Hyderabad

Abstract - In the recent times, terrorism has grown in an exponential manner in certain parts of the world. This enormous growth in terrorist activities has made it important to stop terrorism and prevent its spread before it causes damage to human life or property. With development in technology, internet has become a medium of spreading terrorism through speeches and videos. Terrorist organizations use the medium of the internet to harm and defame individuals and also promote terrorist activities through web pages that force people to join terrorist organizations and commit crimes on the behalf of those organizations. Web mining and data mining are used simultaneously for the purpose of efficient system development. Web mining even consists of many different text mining methods that can be helpful to scan and extract relevant data from unstructured data. Text mining is very helpful in detecting various patterns, keywords, and significant information in unstructured texts. Data mining and web mining systems are used for mining from text widely. Data mining algorithms are used to manage organized data sets and web mining algorithms can be helpful in mining and extracting from unstructured web pages and text data that is available across the web. Websites built in different platforms have varying data structures and that makes it quite difficult to read for a single algorithm.

I INTRODUCTION

Terrorist organizations are using the internet to spread their propaganda and radicalize youth online and encourage them to commit terrorist activities.In order to minimise the online presence of such harmful websites we need to devise a system which detects specific keywords in a particular website. The website should be flagged inappropriate if the keywords are found for efficient system development. Data mining consists of text mining methods that help us to scan and extract useful content from unstructured data. Text mining helps us to detect keywords, patterns and important information from unstructured texts. Hence, here we plan to implement an efficient web data mining system to detect such web properties and flag them for further human review. Data mining is a technique used to extract patterns of relevant data from large data sets and gain maximum insights to the obtained results. Web mining as well as data mining are used simultaneously for efficient system development. The literature survey shows the previous work that has been carried out on this subject. The existing systems have been explained in detail in the paper.The system that we propose to implement significantly improves the current system and eliminates the flaws that exist in the existing system.The methodology and results that we achieved after the implementation of the proposed system have also been explained in brief further. This system should be helpful in anti-terrorism and cyber security response departments. The system should help the cops to track communication held between terrorists and should detect web pages developed in different platforms.

II LITERATURE SURVEY

With the tremendous increase in the use of social network sites like Twitter and Facebook, online community is exchanging information in the form of opinions, sentiments, emotions, and intentions, which reflect their affiliations and aptitude towards an entity, event and policy [1,2,3]. The propagation of extremist content has also been increasing and being considered as a serious issue in the recent era due to the rise of militant groups such as Irish Republican Army, Revolutionary Armed Forces of Colombia (FARC), Al Quaeda, ISIS (Daesh), Al Shabaab, Taliban, Hezbollah and others [4]. These groups have spread their roots not only at the community levels but also their networks are gaining control of social networking sites [5]. These networking sites are vulnerable and approachable platforms for the group strengthening, propaganda, brainwashing, and fundraising due to its massive impact on public sentiments and opinions.

Opinions expressed on such sites give an important clue about the activities and behavior of online users. Detection of such extremist content is important to analyze user sentiment towards some extremist group and to discourage such associated unlawful acts. It is also beneficial in terms of classifying user's extremist affiliation by filtering tweets prior to their onward transmission, recommendation or training AI Chatbot from tweets [6].

The traditional techniques of filtering extremist tweets arenot scalable, inspiring researchers to develop automated techniques. In this study, we focus on the problem of classifying a tweet as extremist or non-extremist. The task faces different challenges, such as different kinds of extremism, various targets and multiple ways of representing the same semantics. The existing studies of extremism informatics are based on classical machine learning techniques [7, 8] or use classical feature representation schemes followed by a classifier.

To overcome the aforementioned limitations of state of the art study [7], we investigate deep learning-based sentiment analysis techniques, which have already shown promising performance across a large number of complicated problems in different domains like vision, speech and text analytics [9, 10]. We propose to apply LSTM-CNN model, which works as follows: (i) CNN model is applied for feature extraction, and (ii) LSTM model receives input from the output of the CNN model and retains the sequential correlation by taking into account the previous data for capturing the global dependencies of a sentence in the document with respect to tweet classification into extremist and non- extremist.

We take the task of extremist affiliation detection as a binary classification task. We take the training set Tr = {t1, t2, t3,…..tn} and class tags (labels) has Extrimist_affliation = {yes, no}. Each tweet is assigned a tag. The aim is to design a model which can learn from the training data set and can classify a new tweet as either extremist or non-extremist. The Twitter-based messaging is a major element of communication among individuals and groups, including extremists and extremist's groups. Using this sort of communication, future terrorist activities can potentially be traced. We propose a technique to identify tweets containing such content. Additionally, we classify sentiments of users in terms of emotional affiliations expressed towards individuals and groups having extremist thoughts. For this purpose, we apply IBM Watson API for tone analysis [11].

In this work, we experiment with multiple Machine Learning (ML) classifiers such as Random Forest, Support Vector Machine, KN-Neighbors, Naïve Bayes Classifiers, and deep learning (DL) classifiers. The feature set for such classifiers is encoded by task-driven embedding trained over different classifiers: CNN, LSTM, and CNN + LSTM. As baselines, we compare with feature set which consists of n-grams [12], TF–DF, and bag of words (BoW) [13].

With the development of machine learning, it has gradually been applied to the analysis of extremist content and sentiments. Ferrara et al. [5] applied machine learning techniques on social media text to detect the interaction of extremist users. The proposed system has experimented on a set of more than 20,000 tweets generated from extremist accounts, which were later suspended by Twitter. The main emphasis was on three tasks, namely: (i) detection of extremist users, (ii) identifying users having with extremist content, and (iii) predicting users' response to extremists' postings. The experiments are conducted in two dimensions, i.e. time-independent and real-time prediction tasks. An accuracy of about 93% is achieved with respect to extremist detection. With the same purpose, a machine learning-based technique is proposed by [7] for classifying of extremist affiliations. The Naïve Bayes algorithm is applied

with the classical feature set. The system is based on the classification of user reviews into positive and negative classes with less focus on identifying, which sentiment class (positive or negative) is associated with extremist communication. In contrast to Ferrara et al. [5] extremist affiliations in the context of social media content are also noticeable in illegal drug usage. For example, in their work on marijuana-related microblogs, Nguyen et al. [14] collected more than thirty thousand tweets pertaining to marijuana during 2016. The text mining technique provides some useful insights to the acquired data such as (i) user attitude can be categorized as positive or negative, (ii) more than 65% tweets are originated from mobile phones, and (iii) frequency of tweets on weekend is higher than other days.

III EXISTING SYSTEM

Terrorism detection has been evolving with time. In the beginning we used to find these more in direct speech. But today due to the impact of social media terrorism and it's supporters existence is everywhere. So the tracking systems has grown weaker to work with such a huge data..

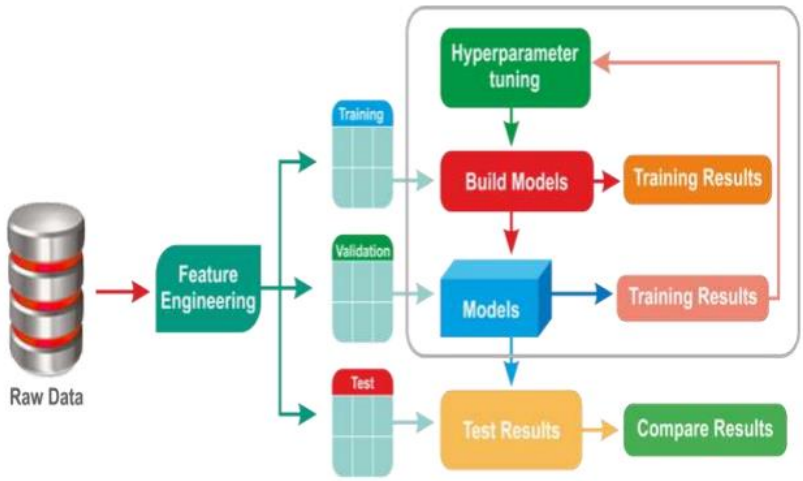
IV PROBLEM STATEMENT

With data revolution in boom, many techniques were introduced to detect the terror speech but not all them made to top of the list. There are several techniques like SVM, NB models etc. which does a decent job but not a great job

V PROPOSED SYSTEM

Overcoming the research from existing system we have noted certain points that could clearly impact the model's accuracy. One major of such is model itself. So we propose to implement the detection of terrorist activity using other models such as logistic regression, Random forest, gradient boosting and decision tree classifier. In the end we compare and contrast the outcomes of every model. As add on we do a comparitive analysis.

VI SYSTEM ACHITECTURE



System Architecture

II. CONCLUSIONS

III. Here is our project that address problems with existing system and solves them effectively. In the end, we have achieved a fully functional model that efficiently extracts fake reviews from given website. Below is the result analysis of the model we developed.

Decision Tree Classification

```
[55]: 1 from sklearn.tree import DecisionTreeClassifier
      2
      3 DT = DecisionTreeClassifier()
      4 DT.fit(xv_train, y_train)
```

```
[55]: DecisionTreeClassifier()
```

```
[56]: 1 pred_dt = DT.predict(xv_test)
```

```
[57]: 1 DT.score(xv_test, y_test)
```

```
[57]: 0.632
```

```
[58]: 1 print(classification_report(y_test, pred_dt))
```

```
              precision    recall  f1-score   support

           0       0.61      0.59      0.60        58
           1       0.65      0.67      0.66        67

    accuracy                           0.63       125
   macro avg       0.63      0.63      0.63       125
weighted avg       0.63      0.63      0.63       125
```

Fig 7.1 Result analysis of Decision Tree

Gradient Boosting Classifier

```
]: 1 from sklearn.ensemble import GradientBoostingClassifier
   2
   3 GBC = GradientBoostingClassifier(random_state=1, n_estimators=2)
   4 GBC.fit(xv_train, y_train)
```

```
]: GradientBoostingClassifier(n_estimators=2, random_state=1)
```

```
]: 1 pred_gbc = GBC.predict(xv_test)
```

```
]: 1 GBC.score(xv_test, y_test)
```

```
]: 0.664
```

```
]: 1 print(classification_report(y_test, pred_gbc))
```

```
              precision    recall  f1-score   support

           0       0.72      0.45      0.55        58
           1       0.64      0.85      0.73        67

    accuracy                           0.66       125
   macro avg       0.68      0.65      0.64       125
weighted avg       0.68      0.66      0.65       125
```

Fig 7.2 Result analysis of Gradient Boosting

Random Forest Classifier

```
]:  1 from sklearn.ensemble import RandomForestClassifier
    2
    3 RFC = RandomForestClassifier(random_state=0)
    4 RFC.fit(xv_train, y_train)
```

```
]: RandomForestClassifier(random_state=0)
```

```
]:  1 pred_rfc = RFC.predict(xv_test)
```

```
]:  1 RFC.score(xv_test, y_test)
```

```
]: 0.648
```

```
]:  1 print(classification_report(y_test, pred_rfc))
```

	precision	recall	f1-score	support
0	0.61	0.66	0.63	58
1	0.68	0.64	0.66	67
accuracy			0.65	125
macro avg	0.65	0.65	0.65	125
weighted avg	0.65	0.65	0.65	125

Result analysis of Random forest

Logistic Regression

```
;1]:  1 from sklearn.linear_model import LogisticRegression
      2
      3 LR = LogisticRegression(solver='newton-cg',class_weight='balanced')
      4 LR.fit(xv_train,y_train)
```

```
;1]: LogisticRegression(class_weight='balanced', solver='newton-cg')
```

```
;2]:  1 pred_lr=LR.predict(xv_test)
```

```
;3]:  1 LR.score(xv_test, y_test)
```

```
;3]: 0.704
```

```
;4]:  1 print(classification_report(y_test, pred_lr))
```

	precision	recall	f1-score	support
0	0.69	0.66	0.67	58
1	0.71	0.75	0.73	67
accuracy			0.70	125
macro avg	0.70	0.70	0.70	125
weighted avg	0.70	0.70	0.70	125

Fig 7.4 Result analysis of Logistic Regression

VII REFERENCES

[1] Aakash Negandhi, Soham Gawas, Prem Bhatt , Priya Porwal Detect Online Spread of Terrorism Using Data Mining.IOSR Journal of Engineering Volume 13,17 April 2019. So here they propose an efficient web data mining system to detect such web properties and flag them automatically for human review. Keywords: Anti-Terrorism, Data Mining, Online, Terrorism,World

[2] Avishag Gordon The spread of terrorism publications: A database analysis,Terrorism and Political Violence journal publishedin Dec 2007.This research note focuses on the spread of terrorism publications from 1988 to 1995 compared to their frequency of appearance from 1996 to 1998. It also identifies the core journals of this research field.

[3] A.Sai Hanuman, G.Charles Babu , P.Vara Prasad Rao, P.S.V.Srinivasa Rao ,B.Sankara Babu A Schematic Approach on Web Data Mining In Online Spread Detection of Terrorism,International Journal of Recent Technology and Engineering Volume-8, Issue-1, May 2019.So here they have propose a compelling web data mining structure to recognize such web properties and standard them thusly for human review. Index Terms: web data mining, terrorism, web structure mining, dread monger affiliations.

[4] Counter Terrorism on Online Social Networks Using Web Mining Techniques Fawad Ali, Farhan Hassan Khan, Saba Bashir, and Uzair Ahmad, Department of Computer Science, Federal Urdu University of Arts, Science and Technology (FUUAST), Islamabad, Pakistan.In this paper some major web mining techniques have been discussed which can be helpful to identify such people and terrorism may be countered from OSN. Each technique is discussed thoroughly, and effectiveness along with its pros and cons are also presented.

[5] Chen, H.. "entiment Analysis in Multiple Languages: Feature Selection for Opinion Classification in Web Forums." ACM Transactions on Information Systems, forthcoming,June 2008.In this study the use of sentiment analysis methodologies is proposed for classification of Web forum opinions in multiple languages. The utility of stylistic and syntactic features is evaluated for sentiment classification of English and Arabic content.

[6] J. Kiruba, P. Sumitha, K. Monisha, S. Vaishnavi Enhanced Content Detection Method to Detect Online Spread of Terrorism,International Journal of Engineering and Advanced Technology Volume-8, Issue-6S3, September 2019.They proposed a system delivery event notification which is used to monitor the activities and delivers notification according to the investigation knowledge. Alert reporting system is developed that takes earthquakes from websites and a message is sent the registered user.

[7] Michael Grenieri, Anthony Estrada Down Converter Characterization in a Synthetic Instrument Context 2006 IEEE Autotestcon.This paper provides an overview of the need for a common set of specification parameters to characterize a down converter in a synthetic instrument (SI).The paper then provides an in-depth technical discussion of two of the less understood down converter related intermediate frequency (IF) output parameters: group delay and phase linearity.

[8] Naseema Begum Detection of online spread of terrorism using web data mining A. Institute of Engineering and Technology, Coimbatore, Tamil Nadu, International Journal of Advance Research, Ideas and Innovations in Technology- Volume 5, Issue

[9] The basic idea of this project is to reduce or stop spreading of terrorism and to remove all these accounts

[10] T.Anand,S. Padmapriya,E. Kirubakaran Terror Tracking Using Advanced Web Mining 2009 International Conference on Intelligent Agent & Multi-Agent Systems.Web mining techniques can be used for detecting and avoiding terror threats caused by terrorists all over the world.