International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 07 Issue: 05 | May 2020 www.irjet.net p-ISSN: 2395-0072

TERRORISM DETECTION THROUGH SOCIAL MEDIA ANALYSIS

Prateek Gangopadhyay[1], Rohan Aggarwal[2], Pooja Saharan[3]

[1] *B.Tech.(CSE) IVth Year, ABES Engineering College, Ghaziabad*
[2] *B.Tech.(CSE) IVth Year, ABES Engineering College, Ghaziabad*
[3] *Assistant Prof. (CSE), ABES Engineering College, Ghaziabad*

Abstract - *The infamous terrorist organizations of the world have received a lot media limelight due to their illegitimate use of the social media to spread their propaganda and recruit foot soldiers. In this project, we utilize to the full, the Twitter Firehose to observe and research on one year of ISIS propagandist and sympathizing activity. We calculate the scope of ISIS presence on Twitter, the possible amount of funding it received, and its shared impact over time. We find that ISIS was able to gather a relatively partial share from the total impact mass on Twitter and that this influence lessened over time. Moreover, ISIS showed a leaning towards enticing communications from other similar pro-ISIS accounts, while appealing to only a partial anti-ISIS sentiment. We found that majority of the interactions ISIS received on Twitter in the year 2015 eventually came from accounts that were consequently suspended by Twitter and that only about 8% of the communications they received were anti-ISIS. In addition, we have created a unique dataset of 37,000 ISIS-related tweets posted in 2015 which we make available for research purposes upon request.*

Key Words*: *Terrorism, Sentiment Analysis, Social Media Analytics, Multinomial Learning Algorithm, Text classification problem.

1.INTRODUCTION

A Prediction Model uses data mining, statistics and probability to forecast an outcome. Every model has some variables known as predictors that are likely to influence future results. The data that was collected from various resources then a statistical model is made. The problem statement of our project required to be handled with an algorithm that can best handle the classification of text into suspicious or non-suspicious text. The proposed algorithms that we had chosen for this were a. Random Forest b. K-nearest-neighbour classification and c. Multinomial Naïve Bayes classification algorithm.

The past few years have proven that social media can serve as a breeding ground to all sorts of hate, and terror too. Notorious terror groups like ISIS have illegitimately though efficiently used social media to propagate their plans and recruit new members, also command their sleeper cells. As these terror groups (ISIS and other groups) have embraced the social media platforms like Twitter, it has become a battleground for these groups with the existing governments. Social media has enabled terrorists the ability to directly interact with their target audience and either spread terror or recruit new foot soldiers.

The main of the project is to create a server that can take text-based input and categorize them (on the basis of probabilistic analysis) to be dangerous text or not. This can help in creating a filter for social media sites where textual posts are permissible to sieve out dangerous text and review and remove them before posting. Therefore, containing the propagation of hate, propaganda and terror preaching.

In the present scenario, if a post is reported on social media, it is reviewed and removed manually which is time consuming and also prone to manual error due to misjudgement.

And also, most messages and posts these organizations use to spread propaganda is encrypted or code-based which can only be understood by someone with a background of intelligence departments. Our model can be trained with a similar experience; therefore, these texts can be flagged as soon as they are posted and hence no delay in surveillance or monitoring the owner accounts.

We will use two server communication for our project. The first server will accumulate and create a dataset and send it. The second server will train the ML model with the algorithm giving the most optimized results and host the final endpoint for detection. A frontend will be integrated for easy GUI experience.

2. Literature Survey

In the process of developing this product we had to go through two very in depth works. First is a thesis submitted to Graduate Faculty of Auburn University by Amanda Guthrie. Second is a paper written by statistical experts Majid Afifi and Fred Morstatter by the name of "Measuring the impact of ISIS Social Media Strategy". Both the works describe efficiently (through statistics) the way social media was exploited to sympathize with terrorist organizations, spread propaganda ad conduct recruits. The collection of tweets in comparison to these works is less as that dataset is classified and available only through special requests.

3. Proposed Methodology

The paper will be comparing different prediction models and deduce their limitations as well as advantages. Since all the research papers used different sets of data to infer the accuracy and for cross validation of data, the authors have used the same data for all the models which will give a clearer view on their performance and lead to a better comparison of the same. On the basis of the results, a modified prediction model will be created to ensure maximum accuracy and performance.

4. System Architecture

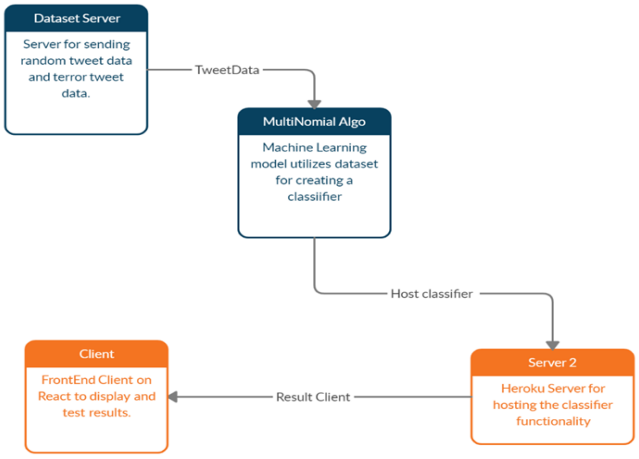

Fig. 1: System Architecture

4.1 Data Collection and Importing

The initial data collection was done keeping in mind the three vectors required to train a model, i.e. TW-RAND, TW-PRO, TW-ANTI. These three collections are of tweets and other related metadata (not relevant mostly). These are collection of random tweets, collection of tweets sympathizing with ISIS and collection of tweets that are anti ISIS and condemns their preaching, provocation and recruitment.
Data was collected from two methods:
i. Kaggle: tweets related to terrorism (37,500) tweets was collected from Kaggle, converted into json and sent via an endpoint call.
ii. Twitter API: random tweets and anti-ISIS tweets were collected from twitter by using trending streams and anti-ISIS hashtags for searching.

Importing the datasets

```
In [5]: data1 = pd.read_excel('/kaggle/input/tweezer/tweetsterror.xlsx')
        data2 = pd.read_excel('/kaggle/input/tweezer/randomtweets.xlsx')

In [6]: data1.head()
```

Fig 2: Data Importing

4.2 Preprocessing Dataset

Since the dataset contains non-pure words that are part of SMS language, we need to pre-process the text to ensure uniformity. To pre-process the text we will use lemmatize, we will also remove common stop words and punctuations.

```
def lemmatizer_preprocessing(mess):
    nopunc = [char for char in mess if char not in string.punctuation]
    nopunc = ''.join(nopunc)
    nopunc = [lemmatizer.lemmatize(word) for word in nopunc.split()]
    nopunc = [word for word in nopunc if word.lower() not in stopwords.words('en
    temp =  ' '.join(nopunc).strip()
    return re.sub(r'[^\w]', ' ', temp)
```

Fig.3 Preprocessing Dataset

4.3 Training the Model

To make hassle free training we will use pipeline technique to streamline the process. Pipeline also has inherent benefit as they can be easily modified, moreover, getting results from them is easier as pipeline provides simple one-use methods.

```
[79]: pipeline1.fit(X_train,y_train)

[79]: Pipeline(memory=None,
              steps=[('bow',
                      CountVectorizer(analyzer=<function lemmatizer_preprocessing at
      0x7fde65a76c80>,
                                      binary=False, decode_error='strict',
                                      dtype=<class 'numpy.int64'>, encoding='utf-8',
                                      input='content', lowercase=True, max_df=1.0,
                                      max_features=None, min_df=1,
                                      ngram_range=(1, 1), preprocessor=None,
                                      stop_words=None, strip_accents=None,
                                      token_pattern='(?u)\\b\\w\\w+\\b',
                                      tokenizer=None, vocabulary=None)),
                     ('classifier',
                      MultinomialNB(alpha=1.0, class_prior=None, fit_prior=True))],
              verbose=False)
```

Fig. 4: Training the model

4.4 Results

After training the model we got following results:

```
In [80]: prediction1 = pipeline1.predict(X_test)

In [82]: print(confusion_matrix(prediction1,y_test))
         print(classification_report(prediction1,y_test))

         [[42838  2750]
          [  254  1562]]
                       precision    recall  f1-score   support

                    0       0.99      0.94      0.97     45588
                    1       0.36      0.86      0.51      1816

             accuracy                           0.94     47404
            macro avg       0.68      0.90      0.74     47404
         weighted avg       0.97      0.94      0.95     47404
```

Fig. 5: Results

3. CONCLUSIONS

The result of the final testing of our project shows that the text entered via the form will be analyzed by a server hosted on Heroku.
The project efficiently shows the communication between two servers for separate functionality i.e. for collection of data and to analyze the text and classify it dangerous or non-dangerous texts.
The website is hosted as a walkthrough of the project from the initial stages i.e. research, study coding and hosting.

ACKNOWLEDGEMENT

It gives us a great sense of pleasure to present the report of the B. Tech Project undertaken during B. Tech. Final Year. We owe special debt of gratitude to Professor Pooja Saharan, Department of Computer Science & Engineering, ABESEC Ghaziabad for his constant support and guidance throughout the course of our work. Her sincerity, thoroughness and perseverance have been a constant source of inspiration for us. It is only his cognizant efforts that our endeavors have seen light of the day.
We also take the opportunity to acknowledge the contribution of Professor (Dr.) Pankaj Sharma, Head, Department of Computer Science & Engineering, ABESEC Ghaziabad for his full support and assistance during the development of the project.
We also do not like to miss the opportunity to acknowledge the contribution of all faculty members of the department for their kind assistance and cooperation during the development of our project. Last but not the least, we acknowledge our friends for their contribution in the completion of the project.

REFERENCES

[1] Interpol **Project First** : Link to Site - https://www.interpol.int/en/Crimes/Terrorism/Identifying-terrorist-suspects
[2] Majid Afifi and Fred Morstatter, "A large-scale study of ISIS social media strategy: community size, collective influence, and behavioural impact , Auburn University