Helping Cyber Forensics for Writer Identification by Identifying Important Features

Akmal Ziyad Gyasoddin[1], Vishal Pushkarnath Marhatta[2], Manoj Vasudev Baviskar[3]

[1,2,3] *Marathwada Mitra Mandal's COE, Pune*

Abstract- **The cover of mystery gave by mobile phones paid early SIM cards, open Wi-Fi hotspots, and flowed frameworks like Tor has profoundly tangled the task of recognizing customers of online life in the midst of logical examinations. From time to time, the substance of a singular posted message will be the principle snippet of data to a maker's character. By what strategy would we have the option to exactly envision who that maker might be the time when the message may never outperform 140 characters on an organization like Twitter? For up to 50 years, language masters, PC specialists, and analysts of the humanities have been as one making automated methodologies to perceive essayists in light of the style of their composed work. All journalists have characteristics of inclination that effect the casing and substance of their created works. These traits can routinely be assessed and evaluated by using AI computations.We will try to analyze the result after classification and find out which word-n-gram and char-n-gram plays most important role.For this evaluation we will use an application developed by us named TweetiFi.**

I. INTRODUCTION

As we see the history of the twitter there are many incidences in which some tweets have caused harm to individual, community and organization directly or indirectly similar cases comes in the department of Cyber Investigation and becomes though to analyze so many tweets of the suspected authors and based on that see which suspected author is having matching pattern with the tweets that have caused harm.

In this scenario we need a system that will help cyber investigator to find out true author behind that tweet.

To build a system that will understand the pattern of the authors and than gives us the result saying that tweet matches with xyz author we are going to use machine learning.

There are many ways by which we can classify text but in this we are planning to use SVM (Support Vector Machine) and Random Forest Classification to train and test to get the true author behind the tweet.

To train a model on some authors we need to Preprocess those tweets.

There are some issues that comes when we talk about tweet data

Tweets with different languages

Tweets with few words and retweets

Tweets with irrelevant data

Tweets with different languages

Twitter supports 34 languages for a user to tweet as of now[8], so it becomes difficult to train on different languages as the percentage of tweets in English language is very high as compared to others[8] and English also being a Global language we are choosing English as our primary language on which we are going to train and test models

Tweets with few words and retweets

As we are supposed to make system know the pattern of the user in that case tweets with very few words are not worth of including in our training dataset as it may mislead accuracy of our model.

So they must be removed and Retweets should also be worthless as they exhibit someone else's pattern that's why they must be removed from the dataset of the author on which we are training.

Tweets with irrelevant data

Tweet include irrelevant data as time, dates, urls etc. that will not help us in predicting the correct author behind a tweet because we are supposed to find out the pattern of the user and these type of irrelevant data will not constitute in the accuracy of our model.

Features Generation

In this step we will generate features that will be used in while training the model and testing the model.
We have used word-n gram generations for producing n-grams for our preprocessed dataset.

II. LITERATURE SURVEY

SVM

In machine learning, support vector machines (SVMs, also support vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis. Given a set of training examples, each marked as belonging to one or the other of two categories, an SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier (although methods such as Platt scaling exist to use SVM in a probabilistic classification setting).

SVM model speaks to the vectors or focuses in space, mapped with the goal that the vectors are assembled as a bunch and the division hole in the middle of two bunches is most stretched out conceivable. At that point, classification is finished by mapping the test vector; the vector is anticipated to have a place with that specific class depending on which side of the hole it is. SVMs are likewise equipped for performing non-direct classifications however for our model it gets the job done by utilizing the direct classification.[2]

Support Vector Machines is another prevalent classification system that examines information. SVM unravels classification issue in less time. It utilizes an adaptable portrayal of the class limits. It can likewise illuminate an assortment of issues with exceptionally less or numerous parameters. It can separate a lot of tests having various classes. For advancement, the best portion is picked to group tweets. Portion, a likeness work which acknowledges two data sources furthermore, consequently gives similitude between the examples. Part capacity speaks to a speck result of info information focuses that are illustrated into the higher dimensional highlight space by change φ. We utilized SVM with straight, outspread premise work and sigmoid bits. Here, parameters gamma ( ) and C are customizable as indicated by our dataset where gamma estimates how far the impact of a solitary preparing test comes to. Gamma esteems mean closeness and farness particular to their high and low values. Parameter C is the expense of arrangement, exchanges off misclassification of preparing tests.[1].

Random Forest

Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees' habit of overfitting to their training set.

A Random Forest is a meta estimator that fits various Decision Tree classifiers on different subsamples of the dataset and utilizations averaging to improve the prescient exactness and authority over-fitting. The Random The backwoods calculation is a group learning calculation, which means the calculation is really comprised of numerous other essential AI calculations. To foresee a class, every essential machine calculation votes in favor of a class and after the majority of the fundamental calculations have cast a ballot, the class with the most votes is the class the troupe calculation predicts. With Random Forests, the basic calculation utilized is the Decision Tree classifier, subsequently the "Backwoods" in Random Forest. The Random Forest calculation carries irregularity into the model when it is developing the Decision Trees. Rather than hunting down the best characteristic while part a hub, or choice in the tree, it scans for the best quality among an arbitrary subset of traits. This procedure makes a decent variety among trees and considers each tree to be based upon various properties, which for the most part results in a superior model. We pick the Random Forest calculation on account of its general viability when arranging numerical sources of info.[9]

III. REVIEW OF METHODS

Filtering Language

Twitter bolsters 34 dialects for a client to tweet as of now[8], so it ends up hard to prepare on various

dialects as the level of tweets in English language is high when contrasted with others[8] and English likewise being a Global language we are picking English as our essential language on which we are going to prepare and test models.
So the question arises that how to identify the language?

For identifying language of the text we are using language library 0.2 python allows us to import its functions and use them to detect language for the specified text and if we get output as english then only we will proceed it further.

Tweets with few words and retweets
As we should make the framework know the example of the client all things considered tweets with not many words are not deserving of incorporating into our preparation dataset as it might delude precision of our model.

So they should be evacuated and Retweets ought to likewise be useless as they display another person's example that is the reason they should be expelled from the dataset of the creator on which we are preparing.

The answer to tweet proportion is processed a similar way utilizing the COUNT() total, and it lets us know the inexact level of tweets presented that answers on different records. This rate is just founded on the example of tweets we have accessible yet gives important understanding into the record's general propensities. The determined component subset can be seen in.[9]

Retweet tag has been checked and after that it is removed if the tweet is a retweet. Similarly for the tweet for few words, here we are removing tweets with 3 or less than 3 tweets which will allow us to reduce the false positive rate.

Tweets with irrelevant data
Tweet incorporate insignificant information as time, dates, URLs, and so on that won't help us in foreseeing the right creator behind a tweet since we should discover the example of the client and these sort of superfluous information won't comprise in the exactness of our model.
This kind of irrelevant data has to be removed to make our prediction better because also that data may belong to the tweets but at the end that data will not help us in knowing the pattern of the user and it will reduce the final accuracy of our model.

Word-N-Grams Generation
If we took a shot at email accumulations furthermore, made a few analyses with word n-grams and character n-grams for origin recognizable proof. Their tests demonstrated that utilizing word bigrams is more fruitful than utilizing word trigrams. In their outcomes, we moreover see that utilizing character trigrams is more effective than utilizing character bigrams. Another examination on composing a print recognizable proof by utilizing n-grams was finished by Sun et. al.[10]. They played out a few investigations with the variable length of character n-grams and saw that 3-grams produce more precise outcomes than 4-grams and 5-grams. They contemplated on 20 client's online audits of Amazon and the normal number of messages per creator were 30. They acquired about 90%-95% exactness on investigations with character 3-grams.[10]

IV. APPLICATIONS & LIMITATIONS

Applications
It will be used in department of cybercrime in identifying the person behind a tweet.
It will be easy to find out that the tweet belongs to that person or not.
It will give the match percentage and will demonstrate who is responsible for a particular tweet.
Limitation
100 % accuracy cannot be achieved till today’s scenario.

Forensic department will not be able to conclude their case with the use of this system only they must have to validate other parameters.
If someone copy pastes someone else’s tweet that also becomes limitation for TweetiFi.

V. DESIGN

Use Case Diagram

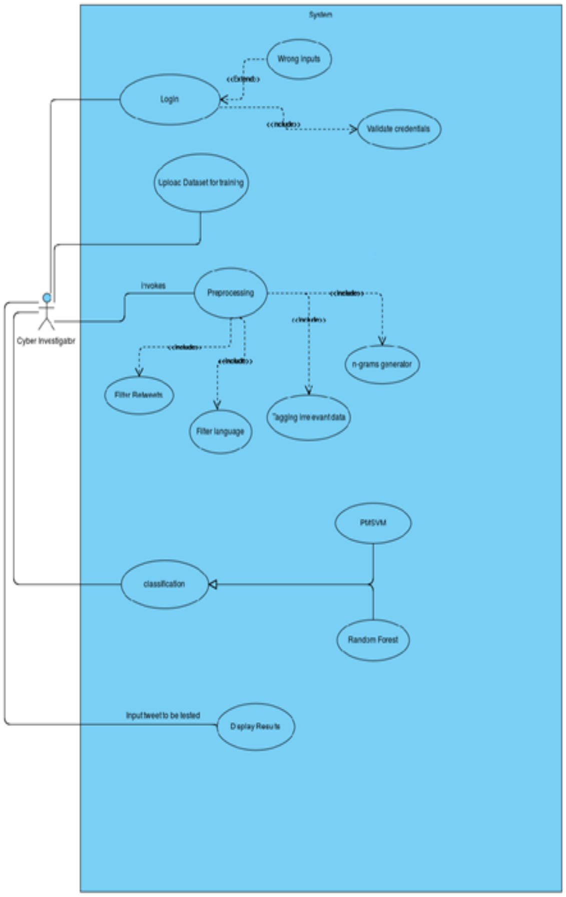


Fig 5.1 Use Case DIagram of Model

User(Cyber Investigator) logins to the Application, his credentials are evaluated based on that he gets message as Wrong or Right.

User will upload the dataset of suspects tweets and then he will click on Preprocess.

Tweets will be filtered based on language.

Retweets and tweets with few words will be filtered.

Irrelevant data is also tagged so that it won't be a part of classification.

N-Grams are generated and stored.

User will click on Classification and then he will upload the tweet on which we have to test those suspects.

After that user will hit on classify which will generate n-grams of test data and check to which that tweet belongs to.

Both SVM and Random Forest will predict the Results.

Architecture Diagram

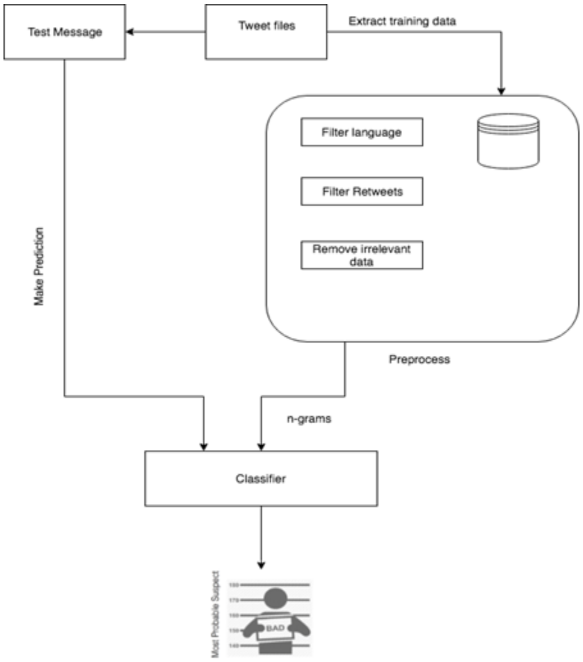


Fig 5.2 Architecture diagram of model

VI. RESULTS

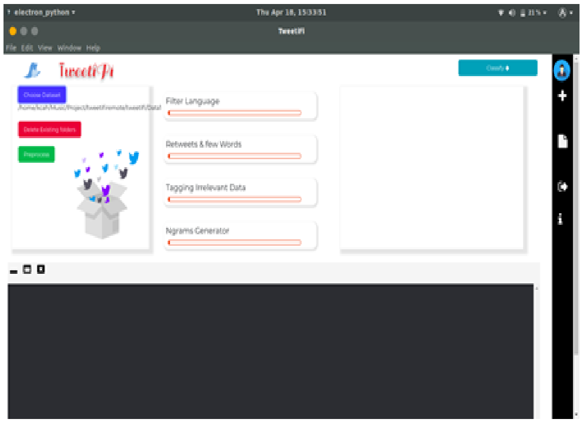
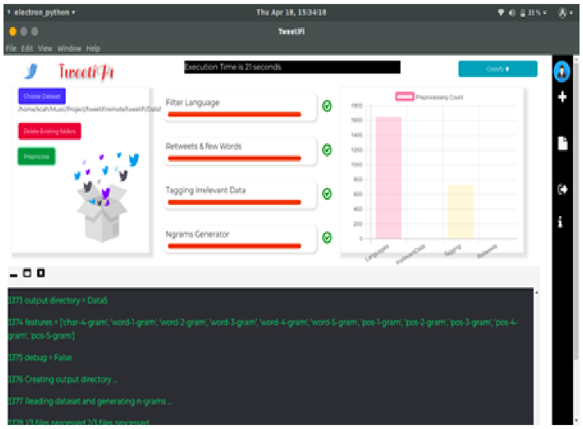


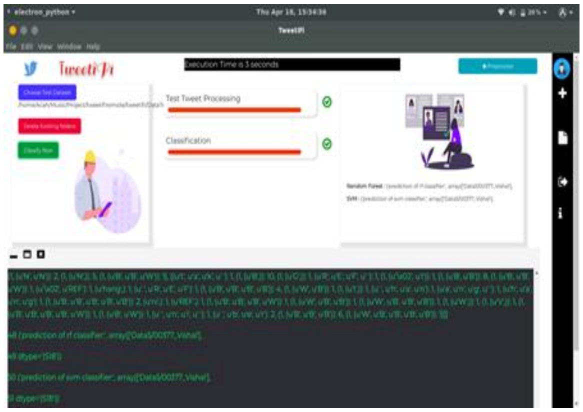
Fig: Upload Suspected Authors Data



Execution Time is 21 seconds

```
3373 output directory = Data5
3374 features = ['char-4-gram', 'word-1-gram', 'word-2-gram', 'word-3-gram', 'word-4-gram', 'word-5-gram', 'pos-1-gram', 'pos-2-gram', 'pos-3-gram', 'pos-4-gram', 'pos-5-gram']
3375 debug = False
3376 Creating output directory ...
3377 Reading dataset and generating n-grams ...
3378 1/3 files processed 2/3 files processed
```

Fig: Getting Analysis Graph & Preprocessing Results



Execution Time is 3 seconds

Random Forest : (prediction of rf classifier', array(['Data5/00377_Vishal'],

SVM : (prediction of svm classifier', array(['Data5/00377_Vishal'],

```
[illegible]

(prediction of rf classifier', array(['Data5/00377_Vishal'],

dtype='[S8]')

(prediction of svm classifier', array(['Data5/00377_Vishal'],

dtype='[S8]')
```

Fig: Recognition Complete of the True Author With Displaying Important feature in recognition

VII. CONCLUSION

After preprocessing 200 author's data having an average of more than 900 tweets we got to know that char-4-gram, word-1-gram, word-2-gram, pos-1-gram, pos-2-gram, pos-3-gram plays an important role among other features. After using random forest the conclusion is that it performs better and predicts the author to which the tweet belongs.

REFERENCES

[1] Authorship Attribution for textual data on Online Social Networks Ritu Banga Pulkit Mehndiratta Student, Department of CSE Jaypee Institute of Information Technology (JIIT) Noida, India Assistant Professor, Department of CSE Jaypee Institute of Information Technology (JIIT) Noida,

[2] Twitter based model for emotional state classification Ravinder Ahuja Dept. of Computer Science Jaypee Institute of Information Technology Noida, India

[3] SOURCE CODE AUTHORSHIP ANALYSIS FOR SUPPORTING THE CYBERCRIME INVESTIGATION PROCESS Georgia Frantzeskou Laboratory of Information and Communication Systems Security, Aegean University Department of Information and Communication Systems Engineering, Karlovasi, Samos, 83200, Greece

[4] Online Social Network Information Forensics A survey on use of various tools and determining how cautious facebook users are? Amber Umair Priyadarsi Nanda Xiangjian He School of Computing and Communications Faculty of Engineering and IT University of Technology, Sydney Australia

[5] Authorship Attribution for Twitter in 140 characters or less Robert Layton Internet Commerce Security Laboratory University of Ballarat

[6] Identifying Actionable Information from Social Media for Better Government-Public Relationship Himani Garg, Charu Bansal, Rishabh Kaushal and Indra Thanaya Indira Gandhi Delhi Technical University for Women, Kashmere Gate, Delhi, India

[7] Text Mining of Tweet for Sentiment Classification and Association with Stock Prices Siddhaling Urolagin Department of Computer Science, Birla Institute of Technology and Science – Pilani ,Dubai, UAE

[8] https://developer.twitter.com

[9] Random Forest Twitter Bot Classifier James Schnebly Computer Science and Engineering University of Nevada, Reno Reno, USA

[10] N-gram Based Approach to Recognize the Twitter Accounts of Turkish Daily Newspapers øslam Mayda Computer Engineering Department Yildiz Technical University østanbul, Turkey