Authorship Detection in Cyber World

Vijeet Rastogi*, Shivam, Prajakta M

Dept. of Information Science & Engineering, The National Institute of Engineering

* Corresponding author email: vijeet272@gmail.com

Abstract

The main goal of this paper is to summarize and propose the idea of detecting fraud in cyber world. Here specifically we are dealing with fraud in text messages. The concept of data mining has been implemented where we have created dataset of user's texting pattern and that data has been trained. The concept of supervise learning also comes into act, if the texting pattern of the sender doesn't match its training data pattern then the receiver will be indicated regarding the doubtfulness of the user. This paper comes under the societal category. The application which we will be creating is going to be a cutting edge in the field of cyber world. Our application can be an answer for new authorship attribution algorithm which can exploit context, can process multi-modal data. For decades computer scientists, scholars have been jointly developing automated method to identify author based on the style of the writing. All authors possess characteristics of habit that influence the form and content of their written work. These characteristics can often be measured and quantified using machine learning methods. A comprehensive review of the method of authorship detection can be applied to the problem of social media forensics. We have provided step by step explanation for several scalable approaches. In some cases, the text of a single posted message will be only clue to an author's identity.

Index Terms- Authorship, Cyber World, Authorship Detection, Supervised Learning, Machine learning

1 INTRODUCTION

The paper lies under the domain of data mining. In this modern era there has been an exponential growth in the amount of data being used and processed. The main reason behind explosion was ignited by the success of relational model for storing data and enhancing the technology of data retrieval and manipulation. Less emphasis has always been laid on analysis of data until recently when companies realized that within these masses of data was a resource that has always been ignored.

The database technologist has always had a concern to find efficient way of storing, retrieving and manipulating data, but the major concern of the machine learning community was to develop methodologies for learning and extracting knowledge from data. Data mining involves concept from ML, DT statistics, clustering, mathematics.

Cyber authorship detection is the process of inferring something about the characteristics of an author from the form and content of their writing present in collection of evidence. Social media is one of the root of primary mode of communication and has challenged the traditional assumption that a forensic investigation will have access to long form of writing.

In this paper we have framed the problem as a computational pipeline in which characteristics are extracted from very small sample of text and supervised learning is deployed to make prediction about unknown samples. At the most basic level the words of a text are useful feature for authorship detection, however all words cannot be considered as feature thus it is common to discard the function words, those words that occurs most frequently but carry little if any semantic meaning to isolate a more stable signal however function words can be coupled with most frequent punctuation or other stable feature becoming more flexible and discriminative, but preserves the underlying statistic of language.

We carry the classification process to detect and clarify whether the message is sent by real author or an imposter then the verification process is carried out comparing the messages with the authors pattern. Finally, if we found any rumour messages we have to identify the author of the messages and therefore indicating the receiver regarding the non-authenticity of the sender. Functional requirement document defines the functionality of a system. The platform used is JVM and the functional user requirement should also describe clearly about the system services in detail. The major components of JVM include constant pool which contains pool of constants, class loader which is responsible for loading the class file, heap area where objects have been stored.

2   LITERATURE SURVEY

Many works have been done in the field of authorship attribution in recent times. Harald Baayen et al. [1] proposed that authorship attribution can be done using the textual pattern and concluded that every author has an authorial structure in written texts. Anderson Rocha et al. [2] showed that authorship attribution can also be applied in social media environment and after processing the messages/comments the author of the message/comment can be determined. Roy Schwartz et al. [3] used character n-grams, word n-grams and flexible patterns for authorship detection in micro-messages. Summing up all the author's references we have used, for a social media application we compare the textual patterns to determine the author of the sent message and also keep updating the textual patterns taking reference from the current conversations. We try to implement a model which compares the keywords of user's text to user's textual pattern to identify the correctness of the authorship of the message sender. We will use Netbeans to create the servlet files and MySQL for the database and we will deploy the database on the local server (or localhost). Eclipse ADT has been used to create the android application. The application will have to be connected to the same network as of the servers for the proper functioning.

Proceedings of the 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA 2018)

319

3   SYSTEM DESIGN

This paper presents the idea to deal with authorship detection. Based on the idea of authorship detection, we are proposing an android application which will deal with non-authenticity of the text messages. In this application, in order to start messaging, a user will have to register himself/herself. After registering, the user can send request to other registered users before starting any conversation. Once the request has been accepted, the two users become friends and now they can start conversing. Whenever a user sends a message, at first the message is preprocessed. The preprocessing of data fetches the keyword and clears the data. After preprocessing, the data is classified using the bag of words model. Then the keywords are compared with the user's text pattern. If the pattern matches, then it is the legitimate user sending the message. If not, then it is someone else. The message will be sent to the receiver but with some extra information which will let the receiver know about the doubtfulness of the sender.

The messages sent by a user are single units and self-contained. Many users tend to adopt a unique style when sending short messages. This style can be detected by any strong classification algorithm such as SVM, and will be sufficient in identifying the author even from a single message. A use case diagram is a type of behavioral diagram created from a Use-case analysis. Its purpose is to present a graphical overview of the functionality provided by a system in terms of actors, their goals (represented as use cases), and any dependencies between those use cases. The use case diagram of our system is represented in Figure 1.
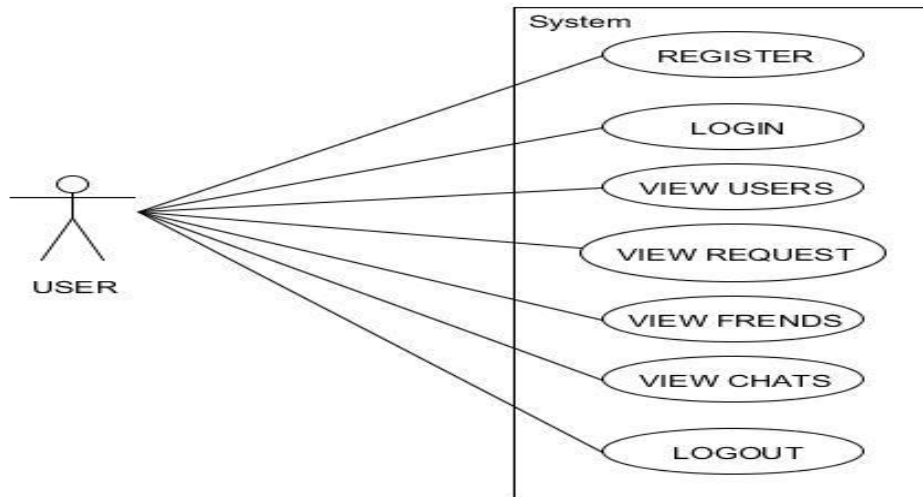

Figure 1: Use Case Diagram of Proposed System

A sequence diagram in Unified Modelling Language (UML) is a kind of interaction diagram which depicts how processes operate with each other and in what order. It is a construct of Message Sequence Chart. The sequence diagram of proposed system is shown in Figure 2.
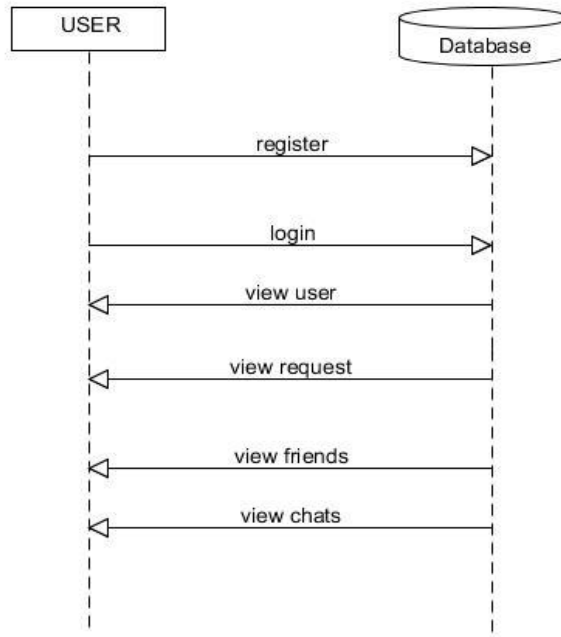
Figure 2: Sequence Diagram of Proposed System

A data-flow diagram (DFD) is a graphical representation of the "flow" of data through an information system. DFDs can also be used for the visualization of data processing (structured design). On a DFD, data items flow from an external data source or an internal data store to an internal data store or an external data sink, via an internal process. The data-flow diagram of the proposed system is shown in Figure 3.
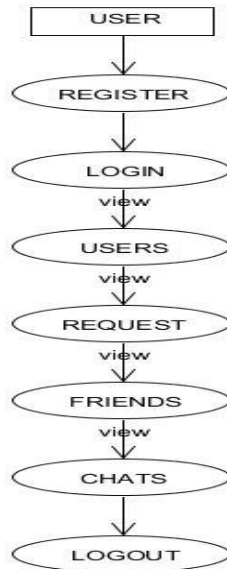

Figure 3: Data-Flow Diagram of the Proposed System

Proceedings of the 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA 2018)

321

Activity diagrams are graphical representations of workflows of stepwise activities and actions with support for choice, iteration and concurrency. In the Unified Modeling Language, activity diagrams are intended to model both computational and organizational processes (i.e. workflows).Activity diagrams show the overall flow of control. The activity diagram of the proposed system is shown in Figure 4.
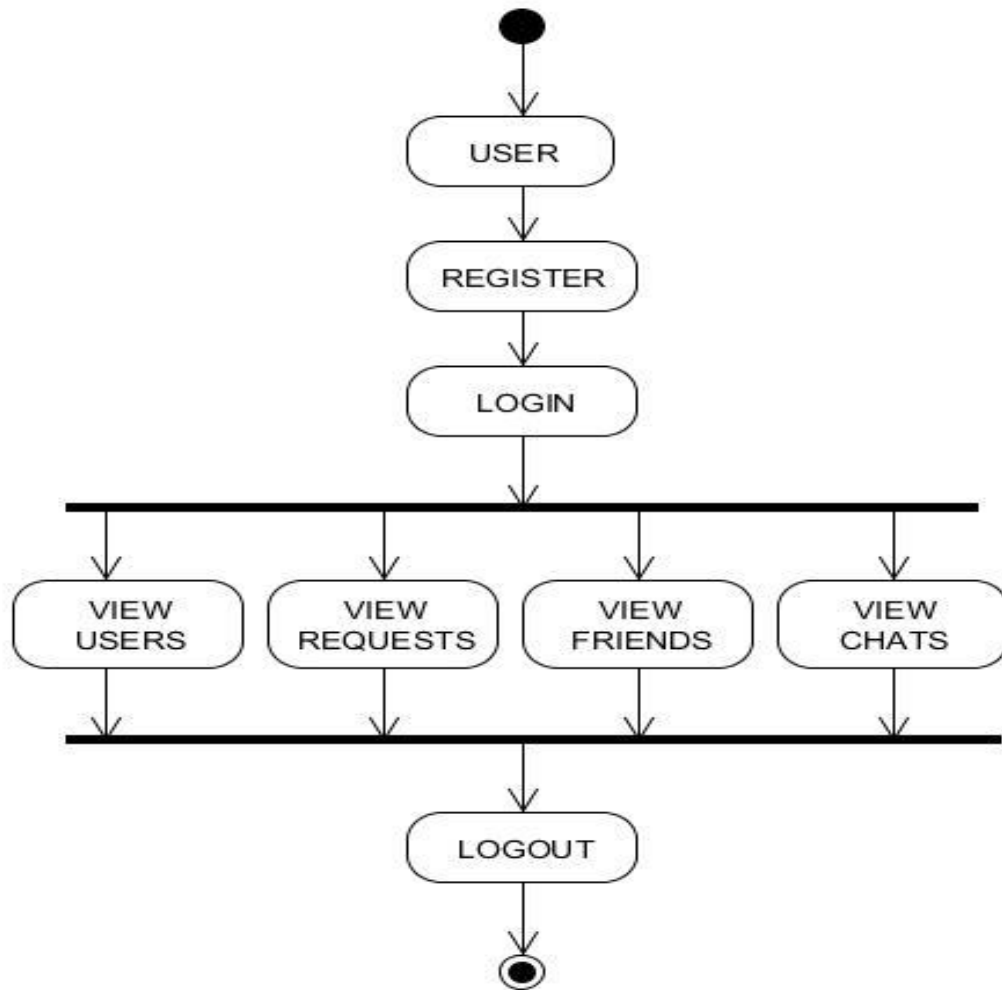

Figure 4: Activity Diagram of Proposed System

A flow chart is a type of diagram that represents an algorithm, workflow or process, showing the steps as boxes of various kinds, and their order by connecting them with arrows. This diagrammatic representation illustrates a solution model to a given problem. Flowcharts are used in analyzing, designing, documenting or managing a process or program in various fields. The flowchart diagram of the proposed system is shown in Figure 5.
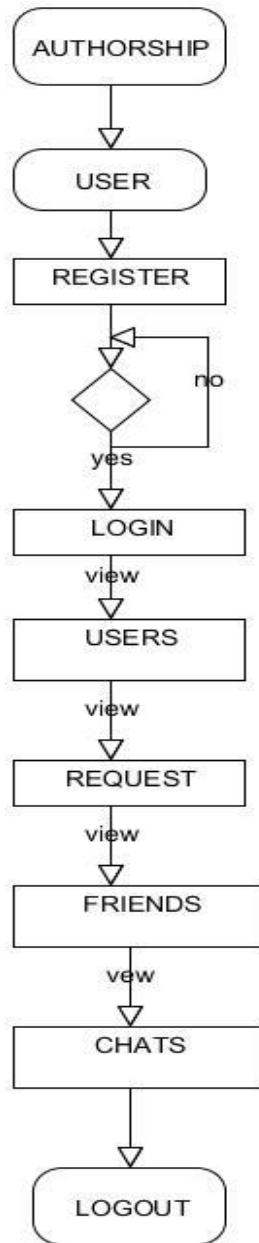
Figure 5: Flowchart Diagram of Proposed System

4   CONCLUSIONS

The proposed system first makes it mandatory for the user to register himself/herself before sending any text message. The pattern of the user is monitored accordingly whenever he/she attempts send any message. The messages are extracted by using lexical analysis, broken down

Proceedings of the 3rd National Conference on Image Processing, Computing, Communication, Networking and Data Analytics (NCICCNDA 2018)

323

into keywords and matched with the user's texting pattern. If the pattern matches, the user is the author else not.

References

[1]   Harald Baayen; Hans van Halteren; Anneke Neijt; Fiona Tweedie "An experiment in authorship attribution".2002 In Journal of Textual Data Statistical Analysis.

[2]   Anderson Rocha; Walter J. Scheirer; Christopher W. Forstall; Thiago Cavalcante; Antonio Theophilo; Bingyu Shen; Ariadne R. B. Carvalho; Efstathios Stamatatos "Authorship Attribution for Social Media Forensics". 2016 IEEE Transactions On Information Forensics and Security.

[3]   Roy Schwartz; Oren Tsur; Ari Rappoport; Moshe Koppel "Authorship Attribution of Micro-Messages" 201 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).