ISSN XXXX XXXX © 2019 IJESC

Research Article Volume 9 Issue No.4

Detection of Social Network Based Cyber Crime Forensics Using Apache Spark

Deepika Chaturvedi[1], Ankita Jangade[2], Harshada Aietawr[3], Aishwarya Pharate[4], Farhana Shaikh[5]
Department of School of Computer Engineering &Technology
MITAOE, Pune, India

Abstract:
Social networking has provided platform to cyber criminals to mask their criminal activities online which poses different challenges on law in tracking and uncovering the fake accounts as most hints are hidden within the posts. Simultaneously, this makes difficult for the forensic experts to classify between real and fake accounts. This is also because of the continuous growth in volume of data that is being generated on regular basis. Handling this huge data with a single node is difficult due to shortage of memory, processors and storage resources. The collection, storage and analysis of this huge data is not possible with the traditional forensic methods and tools. Apache Spark is one of the open source platforms to handle such huge data. Use of Apache Spark and analysis of Big Data in forensic analysis helped in detection of different crimes on social networking platform. This helped in overcoming the challenges faced by the traditional forensic tools used in data analytics. The issues lie with identification of accounts since it is difficult task to classify them. Twitter generates large amount of data through posts, tweets, re-tweets, sharing of the posts etc. There are very large number of Twitter accounts which include both real and fake ones. Our aim is to develop a tool using Apache spark for analyzing the Big Data generated by Twitter to identify the real authorized account by considering different constraints such as number of followers and the number of users the account follows, number of re-tweets and tweets, number of times the tweet is shared and we compared the performance of a wide range of conventional machine learning algorithms.

Keywords: Machine Learning, Parallel Computing, Spam Detection, Scalability, Twitter, Apache Spark, Big Data, Cybercrime, Data Analytics, Distributed Computing.

I. INTRODUCTION(HEADING 1)

In recent years, tremendous amount of data has been generated by information systems. This huge data is mined further by the discovery of smart devices, social networking sites and internet of things with many devices inter connected with internet. This has also seen an increase in cyber threats caused by either individuals or Organized Criminal Groups (OCG) with the intent to break security of information systems. Frequency of cybercrime is increased and their activities has also advanced with advancement in technology. With this tremendous data, forensic analysts face challenges to deal with such huge data with available storage, memory and processing power in terms of processors. Twitter sentiment analysis is useful to detect hate speech in the tweets posted [1]. Main focus of traditional forensic tools is on structured data for analysis which is usually in relational or hierarchical database. Traditional forensic tools lack in some of the features to handle huge data as it does not undergo through any architectural change. However, the huge stream of data generated online from social networking calls for research and design of new generation of forensic analytic methods and tools that can effectively process and co-relate digital evidence found in big data more often in real-time[2].Apache Spark is a distributed computing framework which provides in-memory large data processing ,while at the same time enabling real time analytics and development of programs that can run in parallel within cluster of nodes. To achieve this, the framework abstracts the tasks of system resource scheduling, job submission, job execution, tracking, and communication between cluster nodes. Again, Apache spark offers excellent large data streaming of live data which makes it suitable for streaming social network data and support big data analysis that can be distributed across nodes within a cluster [3]. The study will contribute the following knowledge: a) Demonstrate how distributed computing frameworks can be used in collecting, storing and analyzing big data from social networking sites that has become difficult for forensics to store, collect and analyze on a standalone computer node. b) This research will increase awareness of applications of big data analytics solutions and data science techniques within the digital forensic investigators and to show how they can be utilized in solving large data set challenges and supplement to traditional forensic tools in investigations involving big data. c) The research will also help investigators during Big data forensics to find links between evidences that are hidden within big data sets and which can be easily be overlooked by a forensic investigator especially because of the huge data involved. d) The forensic tool will help law enforcers in investigation involving social media to uncover and correlate evidence found on suspected cyber criminals like cyberbullies and hate speech. Assumptions and Limitations of the Study a) The language used in Twitter sometimes consists of words and phrases that are not formal language (Sheng slang) which burdens to classify sentiments, the study will be limited to phrases made in English. b) There exists a lot of cyber crime related to social network sites including spreading hate speech, cyberbullying, identity theft, harassment, terrorist recruitment and organized criminal groups among others. The study will be limited to forensics involving identification of hate speech and cyberbullying crimes in Twitter social network.

II. PROJECT IDEA

Keeping in mind the forensic challenges observed in Big Data Forensics, the big data generated by social networking sites and limitations on traditional forensic tools. Our research proposed forensic tool based on Apache Spark to support forensic investigation and big data. Using traditional tools and techniques, it is not suitable to collect, store and analyze huge data on single node. Apache spark has in built tools which can handle huge data. It can also easily handle live streaming of data which is suitable for social network data.

The forensic tool utilizes the public available RESTAPI to stream data from the Twitter social media site. Apache Spark Streaming Module is connected to the REST API and the extracted streaming data which is saved to MongoDB database for preservation and later use for training Twitter forensic classification model. Two MongoDB documents was maintained within a single collection. The forensic data which was extracted went through tokenization, lamentation and stop word removal using Apache Spark feature transformers. The resulting data set was subjected through Spark MLLib library for feature extraction. The feature vector was classified using Naive Bayes algorithm with Spark MLLib module to identify and classify tweet/post as either hate speech or cyberbullying. The evidence was then stored within MongoDB document as evidence which was going to be used for evidence visualization on web-based interface. The concept undertook utilized Apache Spark for Big Data processing and Spark Streaming module to capture on live Twitter updates so as to collect/capture the digital evidence near live.

Mining social network data for forensic investigations can be difficult and provides online evidence Which is far different from the conventionally accepted evidence. Digital forensic investigators and law enforcers had begun to consider a new approach for effective ways to bring data from SNSs into investigations and develop standards to enhance its authenticity before a court of law. Even though social network investigators can apply and learn many things from digital forensics disciplines, investigation requires different tools and techniques for social network forensics to ensure the evidence (artifacts) are authentic and the process is done forensically. In order to achieve the objectives as well as storage, collection and analysis of Twitter data stream, our study will consist of a quantitative and explanatory research design. The study consists data mining techniques to get an insight regarding cybercrime from big data collected from Twitter.

III. RELATED WORK

Social networking sites has increased the number of malicious activities performed by hackers or fake users to conduct their wrong deeds for their benefit or to cause harm to one's reputation [6]. Different researches have been made to detect fake users on the social networking platform. Social networking has provided platform to cyber criminals to mask their criminal activities online which poses different challenges on law in tracking and uncovering the fake accounts as most hints are hidden within the posts.

It was witnessed that there is tremendous amount of data which is generated by information systems that has being mined further by the discovery of smart devices, social networking sites and internet of things with many devices inter connected with internet [7]. This has also seen an increase in cyber threats caused by either individuals or Organized Criminal Groups (OCG) with the intent to break security of information systems [6]. These activities must be stopped in order to protect one's personal information. This can be done by detecting the fake users who try to perform these activities. Different social networking platforms like Facebook, Twitter, Snapchat, etc. have millions of users among which there are many fake accounts [6]. Previously there has been work done on detecting the fake accounts where they have explained the fake account detection using Naive Bayes algorithm where each attribute is checked independently using Bayesian theorem, they bring attention to the fact that fake users behave different from authorized users [7]. Similarly in [8] the importance of study of the Naive Bayes algorithm is given for data mining. [9] has explained the importance of data reduction for our work so that important data can be made available for our detection work removing the unwanted data.

This can help in processing our work fast with minimum delay for fetching the data. Study of [11] gives us the explanation of two classification algorithms of K-nearest neighbor and C.45 classification techniques used for classifying our data. Studying these algorithms and comparing them gives us the knowledge of the most efficient algorithm among the three. The term Big Data and its analysis which is often used along with social networking sites uses Hadoop and Spark Shelly Garion, Apache spark for managing big data. Apache spark is the recent most studied in managing big data using various machine learning tools [7],[12]. Sentiment analysis in [2], gives the knowledge for analyzing sentiments which forms the basis on which we will analyze our data set to detect the fake accounts on twitter social networking data. Twitter data contains a large number of unwanted data which gives us an option of data mining for different events and their classification [13]. The study of different forensic tools which can be used to mine data can be useful to understand different approaches to analyze data [21]. Study of [13], [14] gives idea for big data analysis which is necessary for data classification and data mining.

IV. DISCUSSION

In this project, Apache Spark is used because it is 100 times faster as compared to HDFS. It performs in-memory computation due to which it is very fast [15]. There are different types of algorithms in this project. Naive Bayes is used because this is a supervised algorithm and it is based on conditional probability [16]. This algorithm calculates probability by counting the frequency of values and combinations of values in the historical data [17].

Twitter sentiment analysis is a process in which collected data is classified into different groups like positive, negative or neutral tweets. In proposed system, the process of Twitter spam detection by using machine learning algorithms. Before classification, a classifier that contains the knowledge structure should be trained with the prelabeled tweets. After the classification model gains the knowledge structure of the training data, it can be used to predict a new incoming tweet.

The whole process consists of two steps: learning and classifying. Features of tweets will be extracted and formatted as a vector. The class labels i.e. spam and non-spam could be obtained via some other approaches. Features and class label will be combined as one instance for training. One training tweet can then be represented by a pair containing one feature vector, which represents a tweet, and the expected result, and the training set is the vector. The training set is the input of machine learning algorithm, the classification model will be built after training process. In the classifying process, timely captured tweets will be labeled by the trained classification model.

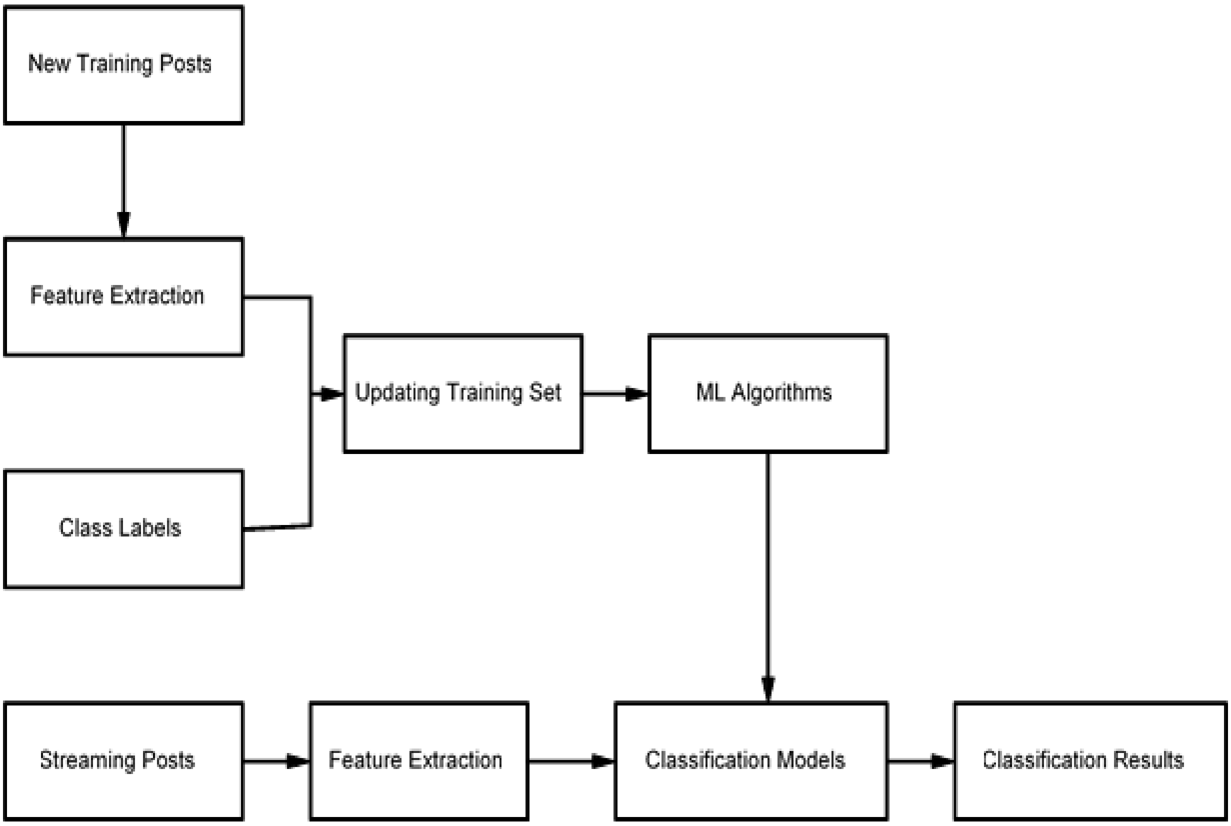

Figure.1.System Architecture

V. MODELLING TOOLS AND TECHNIQUE

Tool	Description
Ubuntu LTS	This formed the operating system upon which Apache spark cluster was setup and web server for front-end reporting interface.
MongoDB	To store tweet data and classified tweets, the project utilized open source NOSQL database MongoDB for easy storage and retrieval
ApacheSpark Framework	This was used to setup big datadist ributed computing cluster which was used to stream Twitter data and Big data analysis.
Apache Server	This was used to setup web server to serve front end reporting module.
Flask Frame- work	Flask framework was used to design the front-end reporting web interface.
PyCharmIDE	This formed the IDE for developing the front-end reporting website in python.
IntelliJIDEAIDE	To develop the back end big data program, the project adopted IntelliJ IDEA 20 as Scala IDE.
SCALA	Scala formed the core language for de- veloping distributed computing program which were executed on the apache cluster.
PYTHON	To develop the front-end reporting module, the project utilized python language.
Bootstrap	To apply styling on the front-end reporting module, the project made use of Bootstrap
PyMongo	This was used to connect python scripts to Mongodb and retrieve data. Chart.js Chart.js was used for visualization of project reports and analysis graphs.

VI. METHODOLOGY

As in software development, there are various data mining techniques applied in projects involving data mining. Most widely used methodology in data mining is Cross Industry Standard Process[4]. CRISP-DM methodology consists of step by step approach that can be used in tackling projects involving data mining. In this, process are broken into six major phases where by the phases do not strictly follow a sequence but allow back and forth movement between the project phases [4].Objective ensures big data is of quality and dependable so that the results obtained from Data Mining can be relied upon in solving problems. It also results in reduced skills required, capturing experience for reuse and general purpose. For projects involving data mining, this particular methodology was an excellent fit because it is robust and well-proven in which data mining tasks can be carried out in many different ways[5].This methodology also proved helpful in allowing one to return to previous phase and return certain tasks. The problem which was being handled in this study involved understanding the problem space and through this, building a forensic tool which required many iterations to understand big data forensic issues. It also aims to find social network crimes and cast a forensic tool which will extract, preserve evidences and easily co-relate crime evidence from big data from Twitter social network.

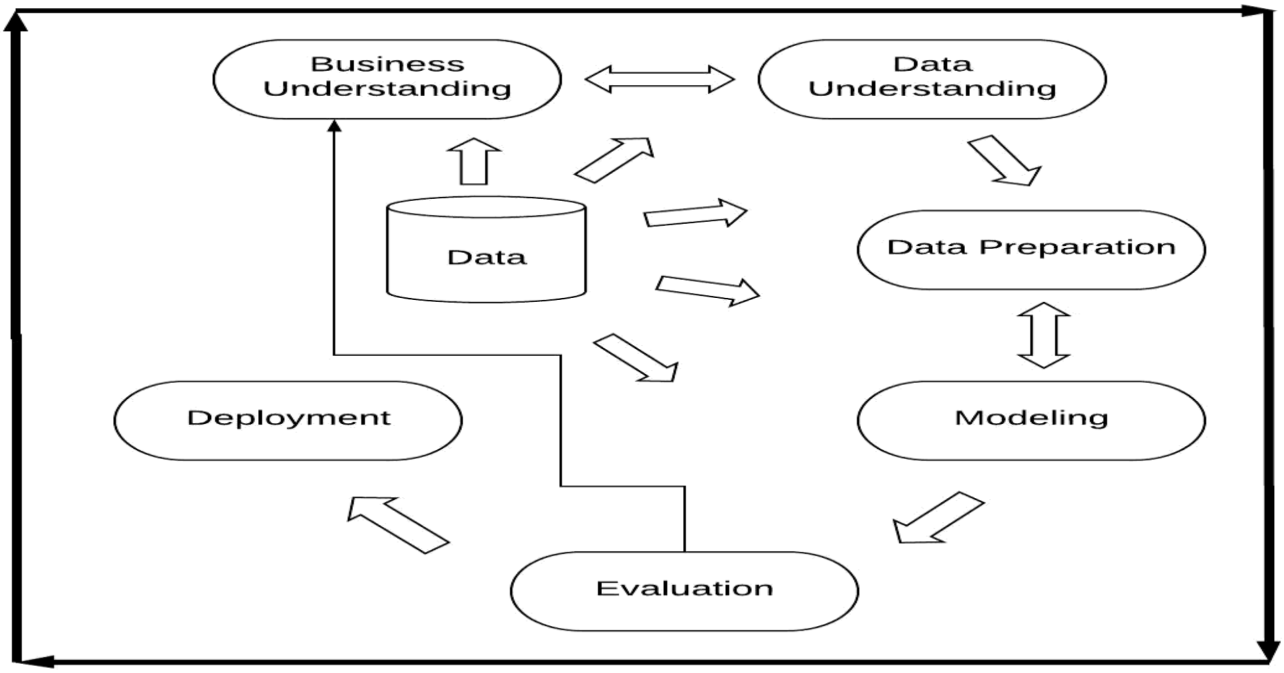

Figure.2.Crisp-Dm Methodology

A. Data Sources

Because of the limited resources and time, the study targeted one of popular social network site, Twitter. Primary data included data collected from actual Twitter pages including tweets/retweets and metadata using Twitter API which enabled us to pull data in real time using Spark Stream module and saving the data into MongoDB for evidence preservation and later text mining/classification using Naive Bayes classifier algorithm and sentimental analysis. In order to have full representation of the entire population, data streaming was carried out using spark stream module to collect real-time tweets and was repeated several times to ensure relatively large volume data set of tweets and posts are fetched on various cybercrime topics. These data were used to extract features for training and modeling the forensic tool which then was used to carry out real time sentiment analysis on Twitter social network site.

B. Data Collection and Data Collection tools

An Apache spark tool for data collection, mining, and cleaning was implemented using Scala programming languages in Apache Spark. The study focused on social media data and metadata from social network site Twitter. Harvesting of social networking forensic data was facilitated through the use of Twitter API which are specific and available to individual social network sites. The study utilized Streaming API by Twitter API to collect datum from all the data sources to ensure complete data was collected for the research. Integration of the forensic tool with Twitter API ensured that key metadata unique to individual account and which is only available through the publisher's API were captured. Scala programming languages together with Spark Stream were used to perform web crawling and scraping. The harnessed data collected was stored in Mongodb before text preprocessing was applied and hashed to ensure its authenticity. The use of Python, Scala in spark and Spark stream module was chosen due to its versatility, agility and previous studies have shown it to be a viable solution for web crawling, indexing and scraping with wider support from data science and development communities. For streaming of data from Twitter, keywords were used particularly the ones oriented to hate speech crimes, bullying like “gun”, “kill”, “murder”, “rape”, “assault”, “kidnap”, “shoot”, “gun”,” crime”,” sinister”, ”bitch” among others. The web crawled data and metadata were stored in the MongoDB database before data preprocessing and transformation was applied using Spark Mlib library. MongoDB was ideal for storing social media API responses since they are designed to efficiently store JSON data while providing powerful query operators and indexing capabilities (Russell 2013). Digital forensics standards dictate that forensic data to be collected in forensically sound manner and to enforce these standards, key information such as SHA-256 hash keys of individual items and logs were maintained to ensure integrity of the evidence collected is verifiable before a jury. Individual item SHA-256 hash values were calculated upon capture and before storage to database and maintained through to analysis. Social media Account metadata unique to individual account and tweets were harvested through integration with REST API's provided by Twitter. The social media account metadata in forensic analysis plays very important role in proofing the authenticity of the evidence collected and help in establishing chain of custody.

C. Data Preparation

The collected data from social media is usually unstructured and contains unwanted characters such as html tags, xml markups, links, exclamation marks, question marks and other irrelevant characters thus required to be prepared, processed and transformed for data evaluation and validation before it can be ingested into spark classification module. Thus, for this study, data collected from Twitter social network site underwent preprocessing with the intent to reduce some noises, incomplete and inconsistent data. The preprocessing included the following tasks:

1) ***Text Preprocessing:*** The collected social media data is usually not only unstructured but also contains other irrelevant and non-textual characters. Text preparation involved cleaning before analysis is performed. The preprocessing is broken down into the following steps:

I. Tokenization

Tokenization involves the process of splitting a text into its desired smaller parts (tokens) seeking to isolate as much sentiment information as possible. Tokenization helps in keeping the vocabulary small as possible. Common with social media, emoticons and abbreviations are identified through the process and treated as individual tokens. This was carried out using Apache Spark machine learning Tokenizer and regex pattern matching. Apache Spark ML come packaged with word tokenization feature (spark. ml. fature. Tokenizer) that was used in tokenizing.

II. Text Normalization

One challenge involving social media data is the abbreviation (e.g. think's, r, u) of texts hence requires text normalization which will involve replacing abbreviated word by the meaning they represent (e.g. thnks=thanks, u=you). This involved text case conversion form caps to lower case and character repetitions are reduced using the Apache spark Normalizer feature transformer to normalize each vector.

III. Stop Word Removal

The data harnessed from Twitter contains words which occurs very frequently but are notuseful as they are used to structure words together in a sentence (e.g. the, at, and, etc.) and they don't contribute to the context or content of textual documents. In text analytics, their high frequency occurrence presents an obstacle in understanding the content of the documents. For this research Stop Words Remover transformers built-in module in Spark ML feature package was used for stop word removal.

Part-of-speech (POS) tagging

Part-of-Speech tagging which is also referred to as grammatical tagging involves assigning words within a sentence their respective part of speech to (such as a verb, noun, conjunctions or adjective) understand its role within the sentence. POS help codetermine what are important keywords within a document or to assist in searching for specific usages of a word in a text document. This involves marking and classification of words in a text sentence based on definition, context, its adjacent relationship with related words in a sentence, or paragraph. Apache Spark includes most popular libraries for NLP in Python among them NLTK, OpenNLP, CoreNLP, WordNet which can be used for text Stemming, lemmatization and POS-tagging.

1) *Bag of Words (Space Vector) Model:*

Bag of words approach is the process of classifying documents where by each word occurrence within the document is used as feature for model training and developing text classifier. A text is represented as a bag of words without paying attention to grammar and even words order but keeping its multiplicity. For document classification, space vector remains the commonly used in method where the frequency or occurrence of each word is used as a feature for training a classifier. This forms part of the text classifiers by taking individual words into account and giving them a specific subjectivity score. Keywords such as gun, kill, murder, rape, assault, fuck, shot, nigga, crime, sinister, bitch among others were used to identify crime-related on Twitter posts by matching the word in the tweets/ posts in the Bag of word dictionary.

D. Data Mining Algorithm and Sentiment Classification

The study employed the use of supervised machine learning approach and specifically Naive Bayes classifier algorithm. Naive Bayes classifier algorithm is simple yet very efficient a kind of classifier which from previous studies has shown to perform well particularly for text classification and sentiment analysis. After the text in Tweet dataset has been segmented into words, the words have been tokenized and normalized, a bag-of-words was modeled by taking individual words and assigning each word a specific subjectivity score whereby if the total score is negative the text was classified as negative and if its positive the text was classified as positive. This bag- of-words formed a dictionary of words and the training dataset for sentiment analysis/text classification. Sentiment analysis was used to determine an author's attitude with respect to either a particular topic or a document's overall contextual polarity.

E. Data Analysis

Twitter streamed data was stored in MongoDB database which formed raw data. The data underwent filtering by using the keywords as search tools which were ran as queries on the database. The filtered data collected was mapped against different crime categories as either cyber bullying, violence based, ethnic based and Sexual based language together with the total counts of each occurrence and exact phrase of crime. Data was analyzed using both Scala and Python libraries and was presented using of graphs and in tabular format. Python was chosen for data analysis because is one of the most popular languages for data analysis and data mining which includes a broad range of libraries suitable for data analysis problems and visualization. It is also an open source software and enjoys a wide support from the data science community. Scala forms one of the mostly used language for programming distributed computing systems because of its scalability and was one of the languages supported by Spark framework. For this study quantitative methods were used for analysis ofthe data. Quantitative analysis was used to graph social media crime incidences categories (e.g. sexual, ethnic, religious, and violence, bullying hate speech) against total tweets.

1) *Feature Selection:* This involved selecting a subset of relevant features that would help in identifying inflammatory or offensive tweets and can be used in the modeling of the classification problems using Naive Bayes model. We did stream the whole Twitter profile account and retrieved all the properties or features making a Twitter Account. This was presented in JSON. The study focused more on Twitter status update which is represented as text. The text field formed the main feature of interest for the study as its Twitters status update for users. For Twitter, forensic analysis we grouped the feature set into two categories i.e. comment based features and metadata-based features. Comment based features involved Twitter comments and replies to the comments and metadata-based features involve Account features such as created at, tweet id, account id, name, screen name, coordinates among others. The Account meta data can be used for account authentication of forensic data and was also focus of the study for forensic evidence preservation and authentication. The figure 36 below shows part of Twitter Account Structure and data types

2) *Training Tweet Labelling:* Naive Bayes machine learning classifier requires that training data set to be labeled as either positive, negative (hate) or neutral. This labeling can be done manually by going through each tweet and labeling it as positive, negative or neutral. Although this will create a highly reliable corpus lexicon, it can be very tedious exercise for large volume softweets. For this study, we adopted an automatic labeling approach by collecting 3,138,367 of tweets and running a Scala function through those tweets which compares each individual tweet word with a list of positive words and a dictionary of hate speech words. The labeled tweets were later fade into Spark ML Pipeline to generate Naive Bayes Classifier model to automatically classify the tweet as either hate speech, bullying in nature are positive/ neutral.

VII. PROPOSED APPROACH

In proposed system, the process of Twitter spam detection by using machine learning algorithms. Before classification, a classifier that contains the knowledge structure should be trained with the prelabeled tweets. After the classification model gains the knowledge structure of the training data, it can be used to predict a new incoming tweet. The whole process consists of two steps: learning and classifying. Features of tweets will be extracted and formatted as a vector. The class labels i.e. spam and non-spam could be get via some other approaches. Features and class label will be combined as one instance for training. One training tweet can then be represented by a pair containing one feature vector, which represents a tweet, and the expected result, and the training set is the vector. The training set is the input of machine learning algorithm, the classification model will be built after training process. In the classifying process, timely captured tweets will be labelled by the trained classification model.

A. Mathematical Model

• **Input:** The large amount of data which is generated from different twitter accounts is then given as input. The data which is collected is very huge in volume and size.

• **Output:** The data which is collected from different sources is then processed in order to find any suspicious activity. Indigital forensics, data mining can be used in identifying correlations or association in big forensic discovering and sorting data into groups based on similar features, discovering insightful patterns in big data that may lead to useful predictions.

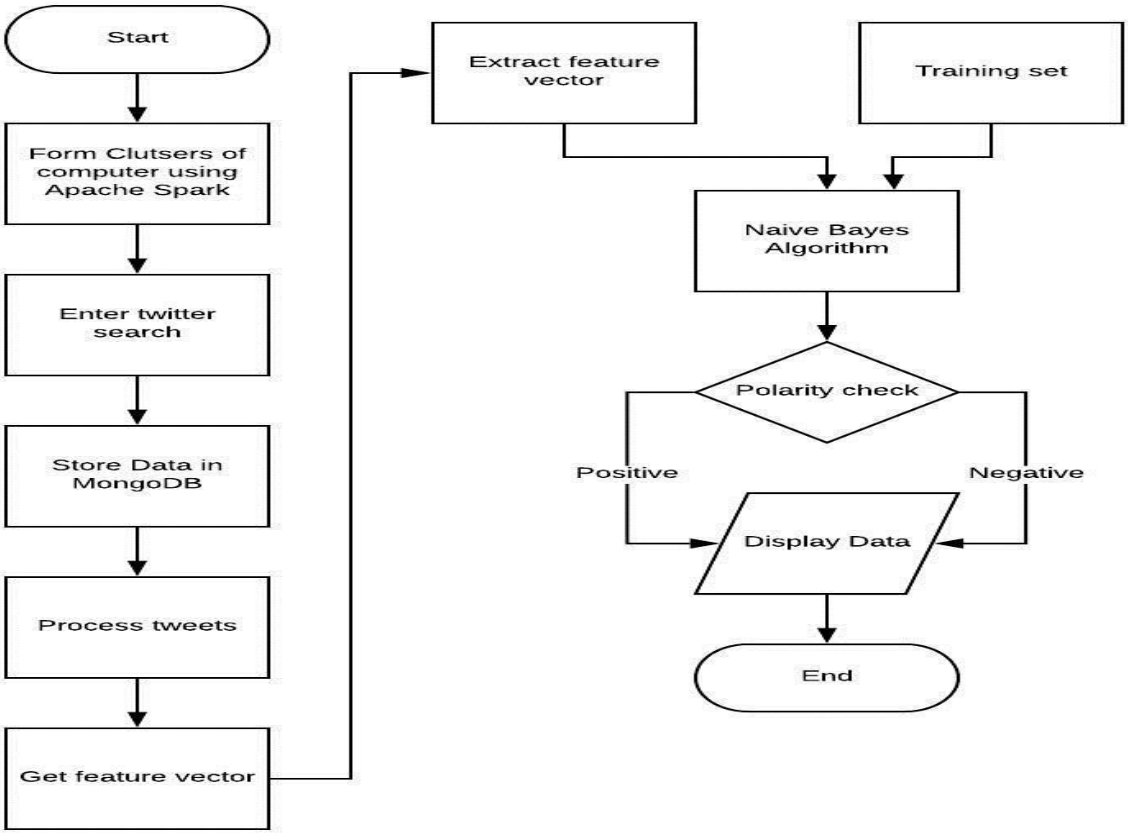

Figure.3.Flow Chart

B. Algorithm

- INPUT: Training data set.
- Given training data set D which consists of documents belonging to different class say Class A and Class B.
- Calculate the prior probability of class A=number of objects of class A/total number of objects
- Calculate the prior probability of class B=number of objects of class B/total number of objects.
- Find NI, the total no of frequency of each class a. Na=the total no of frequency of class A b. Nb=the total no of frequency of class B
- Find conditional probability of keyword occurrence given a class:
  a. P (value 1/Class A) = count/ni (A)
  b. P (value 1/Class B) = count/ni (B)
  c. P (value 2/Class A) = count/ni (A)
  d. P (value 2/Class B) = count/ni (B)
  e. ..
  f. ..
  g. ..
  h. P (value n/Class B) = count/ni (B)
- Avoid zero frequency problems by applying uniform distribution
- Classify Document C based on the probability p(C/W)
- Find P (A/W) =P (A)*P (value 1/Class A)* P (value 2/Class A). P(value n /Class A)
- Find P (B/W) =P (B)*P (value 1/Class B)* P(value 2/Class B). P(value n /Class B)
- Assign document to class that has higher probability.

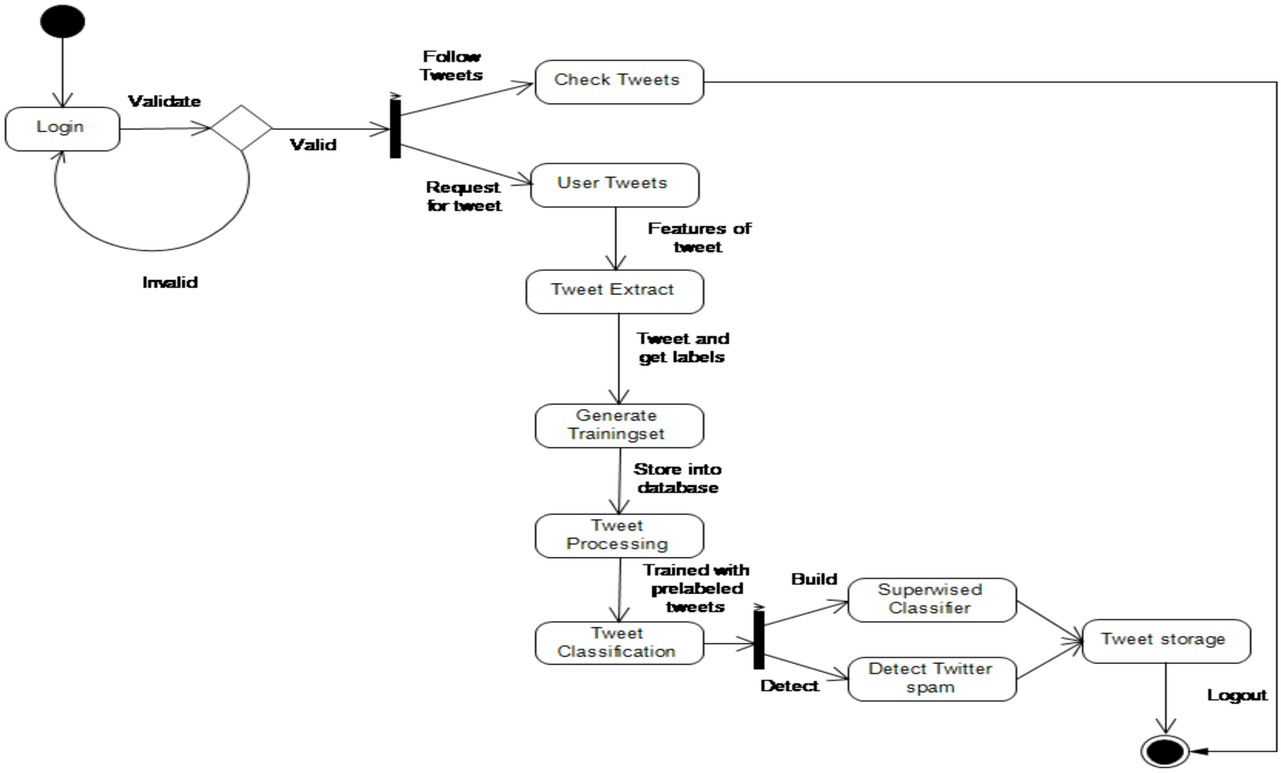

Figure.4.State Daigram

C. Steps for project execution

1) Account Authentication: Through this feature, a user is authenticated into the system. The login credentials of the user must match the information stored in the application. On successful provisioning of the credentials, a user is logged in the system.
2) Cluster formation.
3) Check Tweets: User can check or view tweets by other users.
4) User Tweets: The user of the application has the right to send or post as or to text lines in the form of tweet in the application and stored into the training tweets database.
5) Tweet Extract: At the backend side, extract tweets from training tweets database.
6) Generate Training Set: After extract tweets and get input labels includes no-retweets, no-hashtags, no-user mentions, no-urls, no-chars, and no-digits. While no-chars and no-digits needs a little computing, i.e., counting them from the tweet text, others can also be straightforwardly extracted. And stored in training data set.
7) Tweet Processing: System evaluate the spam detection performance on data set by using machine learning algorithms.
8) Tweet Classification: In the classifying process filtering the spam tweets and Timely captured tweets also label in this classifier.
9) And finally display tweets on the application.

VIII. CONCLUSION

We have implemented and designed a forensic tool which helps in analyzing cyber bullying and hate speech using Apache Spark. Spark ML API, Apache Sparks Machine Learning library can be used to implement the classification algorithm in Apache Spark cluster. Spark streaming API can be used for streaming twitter data and Spark ML API for tweet analysis and classification of this study relies on distributed framework Apache Spark. Data mining techniques namely sentimental analysis helped in finding out cyberbullying and hate speech in the Twitter Social network. The model successfully classified hate speech which was ethnically based. The study showed how MongoDB can be used to preserve the social networking data for further forensic analysis. A SHA256 hash key can be generated for each twitter item within Dstreams and giving a hash key to each individual tweet in the database helps in analyzing changes in the data stream. This can be used for the preservation of forensic evidence. The acquisition of the evidence can be documented preserving tweet stream date and time, improving the chain of custody. Printouts and screenshots of twitter page cannot be useful as an authenticated and allowed evidence in a court as they do not prove or indicate their source, creator. A relevant twitter post metadata must be considered for forensic analysis of twitter post. A lot of hate speech and cyberbullying is done on twitter platform which can be then preserved and used for forensic investigation. The dynamic nature of updates can be a challenge in storage which can be overcome by live streaming of data which in turn may demand large storage space. Apache Spark Streaming API can be used to improve the drawbacks of traditional forensic tools in handling big data problems. the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

ACKNOWLEDGMENT

We are profoundly grateful to Prof. Fahrana Shaikh for her expert guidance and continuous encouragement throughout to see that this project rights its target since its commencement to its completion. We would like to express deepest appreciation towards Dr.YogeshBhalerao, Principal, MIT Academy Of Engineering, Prof. Ranjana Badre, Head of School of Computer Technology and Prof. Sumit Khandelwal, Project Coordinator whose invaluable guidance supported us in completing this project. At last we must express our sincere heartfelt gratitude to all the staff members of School of Computer Technology who helped me directly or indirectly during this course of work.

IX. REFERENCES

[1]. Sentiment analysis in twitter using machine learning techniques-M S Neethu ; R Rajasree

[2]. A study of forensic analysis tools-Sriram Raghavan ; S V Raghavan Department of Computer Science Engg., IIT Madras, Chennai, INDIA

[3]. Ecient kNN Classication Dierent Numbers of Nearest NeighborsShichao Zhang ; Xuelong Li ; Ming Zong Guangxi Key Laboratory of MIMS, College of Computer Science and Information Technology, Guangxi Normal University, Guilin, China ; Xiaofeng Zhu ; Ruili Wang

[4]. High Performance Data Mining Algorithms and Similarity Models Research.- Shengjun Xue ; Hongtao Wang ; Tan Ran

[5]. Data Mining with Big Data-Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding

[6]. Fake news detection using Naive Bayes Classier-Mykhailo Granik

[7]. Social media threats; https://www.calyptix.com/top-threats/socialmediathreats-facebook-malware-twitter-phishing/

[8]. The study of Naive Bayes algorithm online in data mining.-Song Chunyue,Song Zhihuan,Li Ping [9] Data Reduction by removal of lurkers in OSN.-Sumith Nireshwalya, Annappa Basava

[10]. Towards Detecting Compromised Accounts on Social

[11]. K-nearest neighbor and C4.5 algorithms as data mining methods: advantages and diculties.-M.E. Yahia ; B.A. Ibrahim

[12]. Big data machine learning using apache spark MLlib-Mehdi Asse ; Ehsun Behravesh ; Guangchi Liu ; Ahmad P. Tafti Biomedical Informatics Research Center, Marsheld Clinic Research Institute, WI 54449, USA

[13]. Detection of Fake Twitter Followers using Graph Centrality Measures;Ashish Mehrotra, Mallidi Sarreddy and Sanjay Singh

[14]. Big Data Analytics -Hadoop and Spark-Shelly Garion

[15]. Twitter Data Mining for Events Classication and Analysis-Nausheen Azam ; Jahiruddin ; Muhammad Abulaish ; Nur Al-Hasan Haldar

[16]. Nave Bayes Text Classier;Haiyi Zhang, Di Li

[17]. Big Data analytics-Sachchidanand Singh ; Nirmala Singh

[18]. The Hadoop Distributed File System-Konstantin Shvachko, Hairong Kuang, Sanjay Radia, Robert Chansler

[19]. Twitter Fake Account Detection;Buket Erahin, zlem Akta, Deniz Kln, Ceyhun Akyol

[20]. Support vector machines-M.A. Hearst ; S.T. Dumais ; E. Osuna ; J. Platt ; B. Scholkopf

[21]. K-Means Clustering Algorithms: Implementation and ComparisonGregory A. Wilkin ; Xiuzhen Huang

[22]. Random Walk based Fake Account Detection in Online Social Networks; Jinyuan Jia, Binghui Wang, Neil Zhenqiang Gong Networks; Manuel Egele, Gianluca Stringhini, Christopher Kruegel, and Giovanni Vigna.

[23]. Improving Spam Detection in Online Social Networks; Arushi Gupta,Rishabh Kaushal

[24]. FakeBook: Detecting Fake Proles in On-line Social Networks; Mauro Conti,Radha Poovendran,Marco Secchiero

[25]. Recognizing Fake Identities In Online Social Networks Basedona Finite Automaton Approach; Mohamed Torky, Ali Meligy, Hani Ibrahim