Apache Spark based Big Data Analytics for Social Network Cybercrime Forensics

Simon Mulwa Kiio
School of Computing and Informatics,
University of Nairobi, Kenya

Elisha O. Abade
School of Computing and Informatics,
University of Nairobi,
Kenya

ABSTRACT

The anonymity of social networks makes its attractive for cyber criminals to mask their criminal activities online posing a challenge to law enforcers in tracking and uncovering the perpetrators as most evidence is hidden within big data. With this ever-increasing volume of data, forensic analyst faces challenges in investigations involving huge data volumes while at the same time limited by computer processor, memory and storage resources of a single computer node. With increased social media data and the high rate of production, it has become difficult to collect, store and analyze such big data using traditional forensic tools. This study involved the application of apache spark and big data analytic in forensic analysis of social network cybercrimes such as hate speech, cyberbullying and demonstrated the application of data analytics in supplementing the challenges of traditional forensic tools in investigations involving Big Data. The study developed an apache spark based forensic tool to stream and analysis social media data for hate speech and cyberbully cybercrimes while diving to investigate relevant artifacts found on Twitter social network and ways to collect, preserve and ensure authenticity of the evidence. The study employed Naïve Bayes algorithm within Spark ML API to automatically classify and categorize hate speech and cyberbullying found within Twitter social media. The study showed that by generating SHA-256 Hash key for each tweet item within DStreams and storing tweet data together with corresponding Hash key in MongoDB can be used in tweet evidence preservation and authentication. Again, by streaming full tweet Account metadata, the study revealed that such metadata can be used in authenticating the creator, source, date and time for a given hate speech tweet.

General Terms

Social media forensics, Big data analytics, Apache Spark, Cybercrime forensics

Keywords

Big data forensics, Social network forensics, Hate speech, Big data analytics, Mongodb, Apache Spark, Social network cybercrimes, Spark Streaming

1. INTRODUCTION

In the past few years, the world has witnessed an exponential growth in volume of data generated by information systems that has being fueled further by the discovery of smart devices, social networks sites, and internet of things with many devices connected to the 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. Cybercrimes have also increased in frequency and their degree of sophistication has advanced with advancement in technology. With this increasing volume of data, forensic analyst faces challenges when dealing with investigation involving large volumes of forensic data while at the same time constrained by computer processing power in terms of processor, storage and available memory space. Traditional digital forensic tools focus on transactional data commonly known as structured data for forensic analysis that is normally in relational or hierarchical database. Again most widely used traditional forensic tools have not undergone any major architectural change [1] hence they lack suitable features to handle big data forensic investigation. The use of traditional tool to analysis Big Data is time consuming, resources intensive and correlation of evidence from multiple source is not feasible. The ability to derive insights and correlate artifacts found in such big data become difficult using the traditional forensic tools. The range of data from social network sites for forensics increases considerably and increases further with numerous participants involved in social media resulting into challenges in carrying forensics investigation involving these large volumes of data. With this increased social network data, it has become difficult to collect, store and analyze such big data on a single computer node. Traditionally, digital forensic methodologies (identification, preservation, extraction, interpretation, and documentation) would include taking the suspect system offline and removing hard drives from suspected computer system containing source evidence[2], making bit copy of the original hard disk, calculating MD5/SHA-1 checksums, and performing physical collections that capture all metadata. The forensic analyst would then work from this copy, leaving the original hard disk unchanged. However, big data system limitations prevent investigators from applying these forensic methodologies and as such alternative methods for identifying, collecting, storing, analyzing and documenting such data are required. With people, businesses and organization revealing more personal information and business activities online, social networking sites have often been targeted as a platform for committing crimes, including gang recruitment, identity theft, or online harassment and cyberbullying.

The anonymity of social network sites makes its attractive for cyber criminals to mask their criminal activities online posing a challenge to law enforcers in tracking and uncovering the perpetrators. Cyber criminals leave electronic traces as part of their social networks activities and interactions which are contained within an enormous big datasets that are difficult to filter, analyze and correlate evidence using traditional forensic tools [3]. These evidences are not often visible but hidden within large dataset in the form of patterns and correlations. Forensic analysis of social networks and the associated metadata can help forensic investigators understand and solve various cybersecurity problems, including uncovering the online networks of extremists, organized criminal groups, hate speech and cyberbullying [4]. However, the huge stream of data generated from online social networks calls for research and design of new generation of forensics analytics methods and tools that can effectively process and correlate digital evidence found in big data more often in real-time or near real-time and within digital forensic standards. Big Data is defined by the three attributes commonly known as 3V's i.e. Volume, Velocity and Variety in which for data to be categorized as Big data, the data must poses the three V's [5]. With cybercrime increasingly expanding from structured to unstructured data, forensic analysts need new tools and methods to get insights of large volume of data and correlate artifacts from multiples data sources. In order to collect, store and analyze such data fast and effectively, Apache Spark a leading distributed computing framework come in handy with features that can process voluminous amount of data that can range from terabytes to petabytes of data.

2. RELATED WORK

2.1 Social Network Sites Forensics

The identification and collection of digital evidence from big data systems has become challenging for forensic investigators and especially investigation involving cloud based systems and social networks sites. This have been made difficult by the fact that most forensic artifacts are not stored on hard disk rather the data shared on social media sites is largely volatile with no guarantee of later retrieval as it can be deleted or updated. Social network analysis and data visualization techniques can significantly help in the discovery of social media evidence and collection by identifying and understanding relationships and data flow between individuals and events within social networks like Facebook, Twitter. SNA is defined as "a multidisciplinary area involving social, graph theory, statistical and computer science". It uses analytical techniques to discover social relationships that are formed from individuals and groups, the structure of those relationships, and how relationship and their structure influence (or are influenced by) social behavior, attitudes, beliefs and knowledge.

SNA have been used in a wide range of interdisciplinary studies. For example, this approach has been used to discover and analysis individuals in organized criminal groups [3]. In his study, graph based algorithms and methods were used to analyze network structures in identifying interesting and central individuals within a criminal network. An automatic analysis tool for [6] social media posts was proposed to understand the functions, structure, operations of gangs within streets of Chicago, IL region. It involved using Twitter as a source of data to captures tweets posted by gangs and used an automated analysis to discover gang structures, functions, and operations. Intelligent social media analysis and other types of media data can help in understanding and solving various cybersecurity problems including uncovering online terrorist networks and radicalization. [7] in his study, applied social media analysis and machine learning in discovering and predicting civil unrest and online radicalization detection. Structural analysis of social networks sites like Facebook can provide forensic insight about how people relate to one another and where they fit within the larger social network. The social network sites can be exploited by criminals to commit several cybercrimes among them identity theft, cyberbullying, sexual harassment to children and spreading hate speech. These cybercrimes require forensics analysis tools that can effectively be used in identifying the perpetrators and collect the evidence needed for prosecution. SNA has previously been used to uncover such cybercrimes for example a study by [8] who applied text analysis methods in detecting offensive social media contents in protecting the safety of adolescent. By using Lexical and Syntactical Feature he was able to identify content which is offensive in social network sites, and also predict user's potential to send out contents that are offensive. Social networks analysis can also been used for analysis of fraud as more often fraud is committed through illegal set-ups with many accomplices hence social network analysis might give new insights by investigating how people influence each other in what is called guilt-by-associations, where it is assumed that fraudulent influences run through the network [9].

2.2 Legal Challenges to Social Media Evidence Authentication

With the increased use of social networking sites and its target by cybercriminals, social media evidence is becoming highly relevant in cybercrimes investigations, legal disputes and broadly discoverable, but challenges lies in evidence authentication as there is lack of best practices, technology and processes. Social media status updates, posts and photographs on Social networking sites are increasingly denied admission as evidence in criminal litigation with courts citing issues with the evidence authentication. An article by [10] states that "Given the transient and cloud-based nature of social media data, it generally cannot be collected and preserved by traditional computer forensics tools and processes. Full disk images of computers in the cloud is effectively impossible and the industry has lacked tools designed to collect social media items in a scalable manner while supporting litigation requirements such as the capture and preservation of all key metadata, read only access, and the generation of hash values and chain of custody." With these challenges, social media evidential data must be properly identified, collected and preserved in a manner that is consistent with digital forensics best practices so at to ensure all available circumstantial evidences are collected, including account metadata and a proper chain of custody established through the evidence collection. With this in place and associated account metadata preserved, it become easier to establish or reveal authenticity of the evidence. For example, metadata fields for individual Facebook account posts such as status updates, photographs among others can provide important information to reveal the authenticity of the Facebook posts when collected and preserved using best digital forensic standards. In the evidence authentication process, actor accounts metadata can be examined to establish authenticity of the content whereby hidden metadata fields that are not visible on the face of a social media site (including dates, URLs, IDs, usernames among others) can be used to reveal authenticity and hence crucial for proper preservation and production of social media evidence.

3. METHODOLOGY

The study used both quantitative and Exploratory research design to collect, store and analyze Twitter stream data. These research design was useful in exploring the application of big data solutions and distributed computing frameworks in the field of digital forensics to collect, store, process and analyze big data. The study employed data mining methodologies to get insightful information regarding cybercrime from big data collected from Twitter social network site. The most popular methodology used in data mining is Cross Industry Standard Process for Data Mining[11]. CRISP-DM methodology describes step by step approaches that can be used in tackling projects involving data mining. In this methodology, the data mining process are broken into six major phases where by the phases do not strictly follow the sequence but allows for back and forth movement between the project phases [11].

3.1 Data Sources

Primary data was used to get forensic data for evidence retrieval and Twitter social network site was used as source of data for the study. Primary data included data collected from actual Twitter pages including tweets/retweets and account 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 classification using Naive Bayes classifier algorithm and sentimental analysis. 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 dataset of tweets were fetched on various cybercrime topics. These data were used to extract features for training and modeling the forensic tool which was then used to carry out real time sentiment analysis on Twitter social network site.

3.2 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. 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 Streaming API were used to perform web crawling and scraping. For streaming of data from Twitter, keywords were used particularly the ones oriented to hate speech crimes and cyberbullying like “gun”, “kill”, “murder”, “rape”, “assault”, “kidnap”, “shot”, "gun," "crime," "sinister", “bitch among others. Individual item SHA-256 hash key was 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.

3.3 System Design

For this study, an Apache Spark Standalone cluster was setup and a web based forensic tool was implemented using Apache Spark which offered a high scalable data intensive processing which is suitable for big data processing like Twitter data. In addition, Spark offers scalable real live streaming module (Spark Streaming) for data, making it suitable for use with Twitter API. Scala and Python programming language was used for both development of logic applications and interfacing. MongoDB was used as the back end for storing stream data and the data fetched. MongoDB is ideal for storing social media API responses since they are designed to efficiently store JSON data while providing powerful query operators and indexing capabilities. The implementation comprised of a forensics data streaming module, MongoDb database, apache spark classifier and front end web interface. It also provided additional information of Twitter account metadata and location where the crime was committed or uttered. The dictionary of words was mainly a dataset of potential crime feature words such as “kill”, “murder”, “rape”, “kidnap”, “shot”, "gun," "crime," "sinister, “bitch” and others which were used as the baseline for assessing the crime forensic data collected.

3.4 Architectural Design

The apache spark forensic system was implemented as a three-tier application using apache spark, MongoDB, Scala, Python programming languages. The Figure 1 below shows an overview of the system components and the interconnections between them which enabled it to extract, store, transform and carry forensic analysis of twitter posts.

a) Spark Streaming Module

Implemented using Spark Streaming module of Apache Spark, this module captures live streams from Twitter API, parses the data in JSON format to the desired format and stores the data in MongoDB database. The data contains all the information about the original tweets, time it was streamed, Analyzer collecting the data and metadata required for authenticity of evidence.

b) Apache Classifier Module

This module was implemented using Scala programming languages in apache spark. The module incorporated classification algorithms specifically Naive Bayes Classifier that is available in Spark ML. Spark ML pipeline was used in providing a set of tokenization, stemming, tagging and stop words removal. Spark ML is used for sentiment analysis and classification of data streams before they are stored on MongoDb and based on the analysis, each tweet is classified as positive (crime) or negative (not crime) sentiment and the result is then persisted into MongoDb as crime evidence which is used later by Report Module.

c) MongoDb Backend Module

This was implemented using MongoDb and it was responsible for storing data streamed from Twitter social network site. It stores raw tweets from Twitter in the form of JSON document format. MongoDb also stores twitter data which has been classified as of criminal in nature which was later used by reporting module. Bag of words which contains hate/bully words which are used for feature extraction and tweet classification was also stored in MongoDb database.

d) Report Module

This is reporting module which provides visualizations of data classification results as forensic report showing Tweets which were identified as of criminal nature in form of percentages and charts.

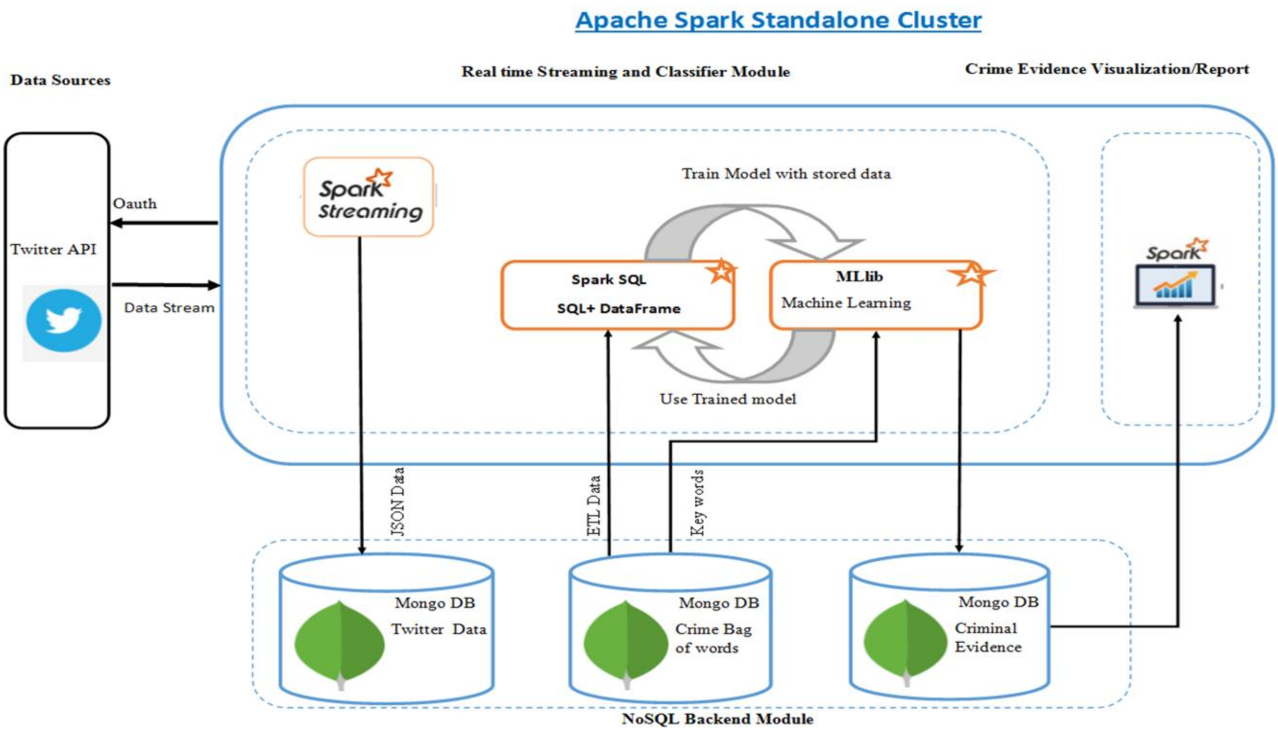


Fig 1: Spark Forensic tool architectural design

4. SYSTEM IMPLEMENTATION

The study setup an apache standalone cluster and designed an apache spark twitter streaming tool to collect forensic data which was subjected to sentimental analysis to identify hate speech and cyberbullying using Spark ML API. The model was trained using 3,138,367 million tweets and used to correctly classify twitter data according to the three categories namely positive, neutral, negative (hate speech/cyberbullying).

4.1 Forensic Tool Module Analysis

The forensic tool was designed and programmed using both Scala and Python programming languages and is composed of mainly nine main components:

a) Training Twitter Streaming

This was Scala based module implemented Spark Streaming API and was responsible for streaming live tweets and storing them on local hard disk under TwitterJson files. These tweets were used for training Naïve Bayes model.

b) Spark Naïve Bayes Model Creator

This was Scala based module which implemented Spark ML API and Naïve Bayes classifier pipelines. The module worked by loading locally stored JSON tweets and training Naïve Bayes classifier model which was saved on local disk as “NaiveBayes Classifier Model”.

c) Tweet streaming model

This was live tweet streaming module which was implemented using Spark ML API and utilized earlier saved trained model to stream live tweets and classifying them as hate speech or bullying. It was also responsible for categorizing the tweets in different categories as either bully, ethnicity, sexual, religious and others for tweets which didn’t fall under the defined categories.

d) Flask Report Viewer

Implemented using flask, HTML and JavaScript, this module was responsible for presenting forensic hate speech reports and analysis graphs.

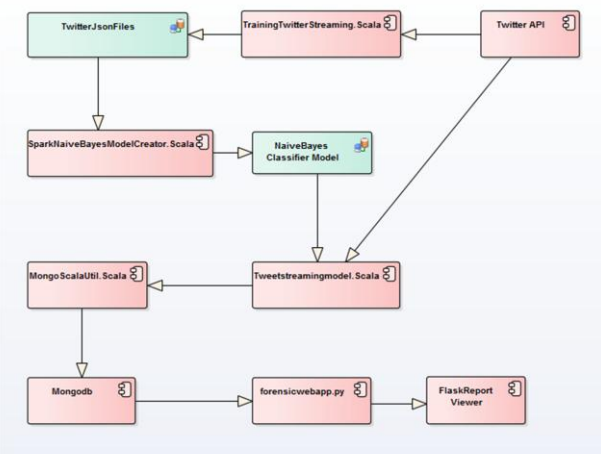

TwitterJsonFiles
TrainingTwitterStreaming.Scala
Twitter API
SparkNaiveBayesModelCreator.Scala
NaiveBayes Classifier Model
MongoScalaUtil.Scala
Tweetstreamingmodel.Scala
Mongodb
forensicwebapp.py
FlaskReport Viewer

Fig 2: Forensic Tool Module Analysis

e) Mongo Scala Util

This module implemented Mongodb connector for spark and utilized MongoDB Scala Driver for storing of live classified tweets to forensicdb with Mongodb. It was also responsible to persisting raw tweets to Mongodb database for evidence preservation and authenticity.

f) Mongodb

This module formed the backend storage for preservation of evidence and classified tweets. The study utilized Mongodb 3.4 version for tweet storage and preservation.

g) Forensic web app

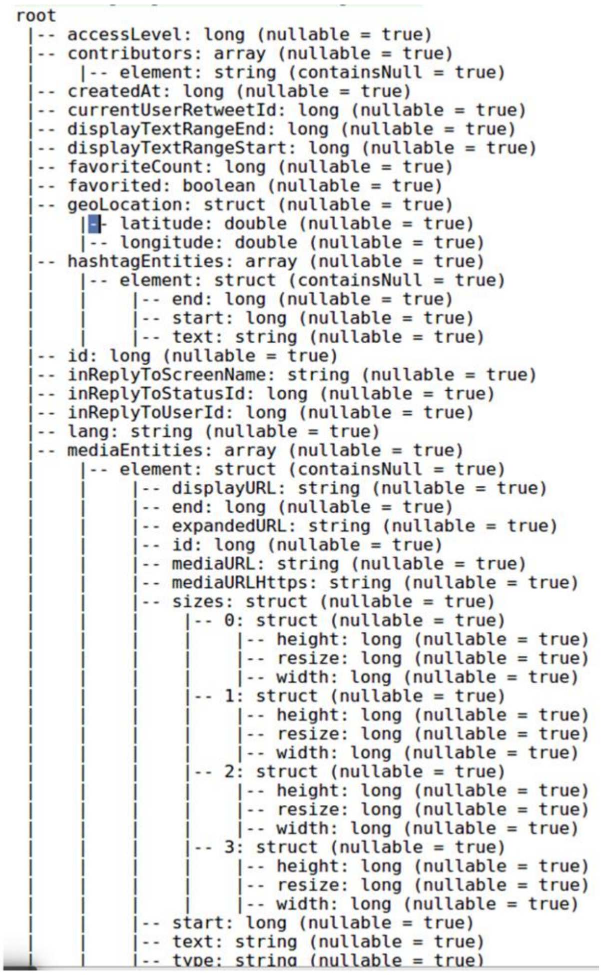
This is a Flask based web application which gets data which connects to MongoDB using PyMongo module to retrieve classified tweets for presentation. The module was also used for tweet analysis using graphs and charts.

4.2 Data Collections

Data for the project was streamed from Twitter social network site and the study utilized publicly accessible Twitter API with Spark streaming API. The study streamed 3,138,367 tweets which were stored locally on local hard disk. The tweets were filtered using a set of keywords that are viewed as offending, insulting, and intimidating to people or of inflammatory in nature or bullying. For this study, several keywords were used among them- hate you, nigga, stupid, idiot, fuck you, faggot, kill, bitch, dyke, gay, black nigger, white people, black people, ugly, terrorists. These keywords were also categorized into different groups based on their biasness i.e. ethnicity, religious, sexual, violence, bully. The collected Tweets formed our tweets dataset which were later used to train Naïve Bayes Model to analyze and classify tweets as either positive sentiments, hate (negative) sentiments, or neutral sentiments.

4.3 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 Naïve Bayes model. The study 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 Twitter's status update for users. For Twitter, forensic analysis the study 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, screenname, and coordinates among others. The Account metadata can be used for account authentication of forensic data and was also focus of the study for forensic evidence preservation and authentication. Figure 3 below shows part of Twitter Account Structure and data types.



```
root
 |-- accessLevel: long (nullable = true)
 |-- contributors: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- createdAt: long (nullable = true)
 |-- currentUserRetweetId: long (nullable = true)
 |-- displayTextRangeEnd: long (nullable = true)
 |-- displayTextRangeStart: long (nullable = true)
 |-- favoriteCount: long (nullable = true)
 |-- favorited: boolean (nullable = true)
 |-- geoLocation: struct (nullable = true)
 |    |-- latitude: double (nullable = true)
 |    |-- longitude: double (nullable = true)
 |-- hashtagEntities: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- end: long (nullable = true)
 |    |    |-- start: long (nullable = true)
 |    |    |-- text: string (nullable = true)
 |-- id: long (nullable = true)
 |-- inReplyToScreenName: string (nullable = true)
 |-- inReplyToStatusId: long (nullable = true)
 |-- inReplyToUserId: long (nullable = true)
 |-- lang: string (nullable = true)
 |-- mediaEntities: array (nullable = true)
 |    |-- element: struct (containsNull = true)
 |    |    |-- displayURL: string (nullable = true)
 |    |    |-- end: long (nullable = true)
 |    |    |-- expandedURL: string (nullable = true)
 |    |    |-- id: long (nullable = true)
 |    |    |-- mediaURL: string (nullable = true)
 |    |    |-- mediaURLHttps: string (nullable = true)
 |    |    |-- sizes: struct (nullable = true)
 |    |    |    |-- 0: struct (nullable = true)
 |    |    |    |    |-- height: long (nullable = true)
 |    |    |    |    |-- resize: long (nullable = true)
 |    |    |    |    |-- width: long (nullable = true)
 |    |    |    |-- 1: struct (nullable = true)
 |    |    |    |    |-- height: long (nullable = true)
 |    |    |    |    |-- resize: long (nullable = true)
 |    |    |    |    |-- width: long (nullable = true)
 |    |    |    |-- 2: struct (nullable = true)
 |    |    |    |    |-- height: long (nullable = true)
 |    |    |    |    |-- resize: long (nullable = true)
 |    |    |    |    |-- width: long (nullable = true)
 |    |    |    |-- 3: struct (nullable = true)
 |    |    |    |    |-- height: long (nullable = true)
 |    |    |    |    |-- resize: long (nullable = true)
 |    |    |    |    |-- width: long (nullable = true)
 |    |    |-- start: long (nullable = true)
 |    |    |-- text: string (nullable = true)
 |    |    |-- type: string (nullable = true)
```

Fig 3: Twitter Account Attributes

4.4 Social Media Evidence Identification

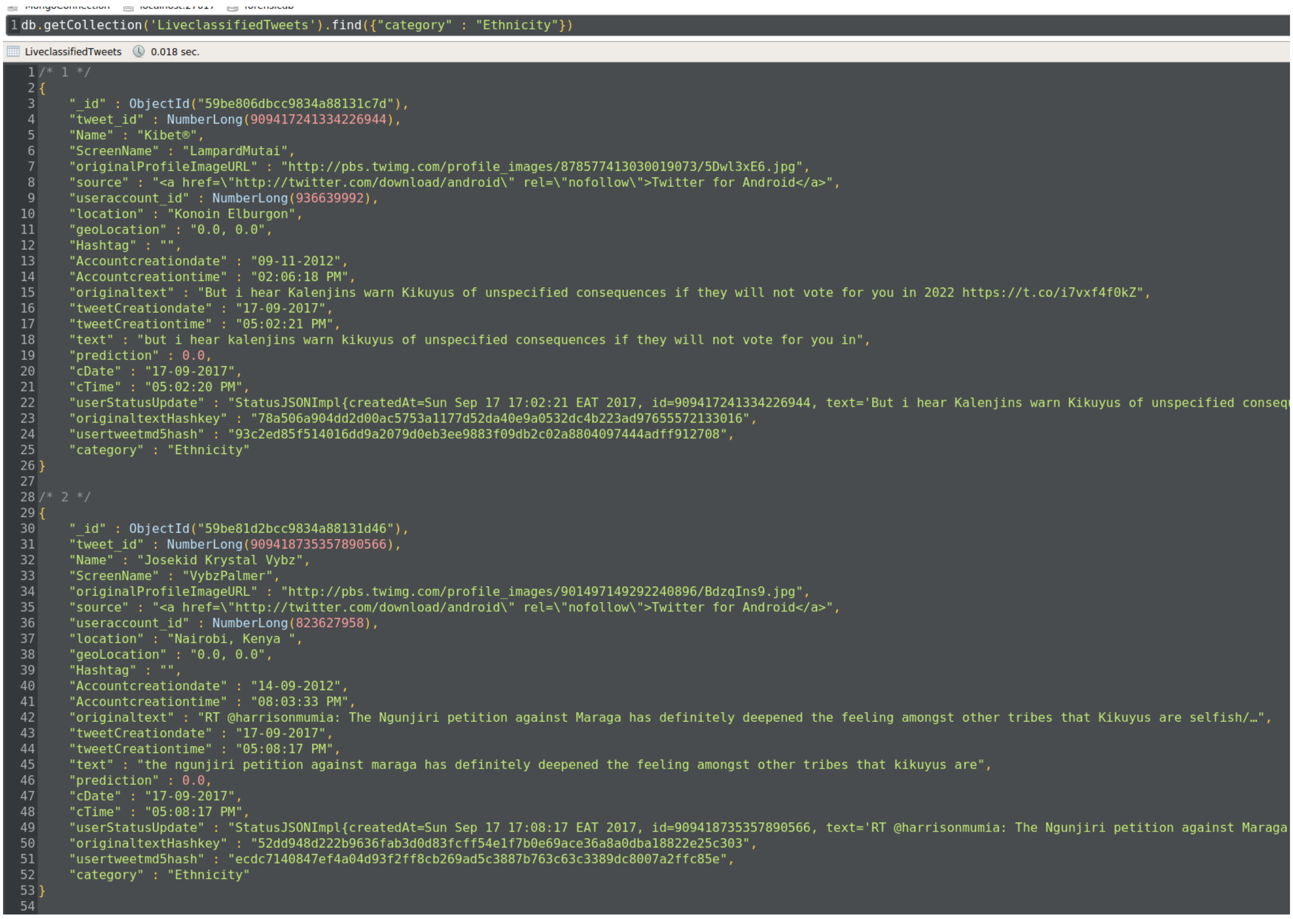
Social media users create massive amounts of data which becomes challenging when trying to extract evidence as it's hidden with enormous big data. In this study, 3,138,367 tweets were streamed and analyzed to identify hate speech and bullying tweets. This involved identifying which attributes might be used as evidence for commitment of cybercrime in Twitter Social Network. The study went ahead to identify attributes which might be used in supporting the evidence and whether such tweet account was used to commit the said cybercrime or hate speech. In this study, it was noted that user status updates what is commonly referred as tweets are used to express hate speech or bully comments. The status updates (tweet) are represented as text field in twitter account structure as shown in table below. Twitter evidence can also include photographs which might carry out inflammatory messages or contents which might be of hate speech or bullying in nature. With the development of GPS smartphones and location-based services, Twitter enables user to tag or provide location information which might be used during search warrant of a culprits in case of cybercrime commitment on Twitter.

Table 1: Twitter Account Forensic Attributes

Twitter Field Name		**Evidence/Comments**
Twitter Status Update (Twitter Text field)		Tweet updates which indicate user's/Account status updates or posts
originalProfileImageURL		This shows users profile picture as uploaded by the user
created_at		This can be used to show when such tweet was created and can be used to authenticate when tweet which has be categorized as hate speech/bullying was created or posted.
id		This is unique identifier for the tweet in question and can be used to uniquely identify each individual tweet.
Users	created_at	The time when the account in question was opened with twitter.
	id	A unique identifier which represents twitter account user.
	location	This defines the user location for this account's profile and might be used to identify the location of the user or where user might have created the account in.
	Name	The name of the user, as they've defined it
	profile_image_url	The account user's profile image which can be used to identify the user physically.
	screen_name	An alias which the user identifies himself with.

4.5 Evidence Retrieval

Social media users create massive amounts of data which becomes challenging when trying to extract evidence as it's hidden with enormous big data. As highlighted in table 1 above, Twitter Account profile encompasses many fields which are not possible to be retrieved by snapshotting and printing of Twitter web pages. This invalidates such evidence collected by screen shots or printing web pages as they lack the supporting metadata. To improve on evidence validity, this study focused on an automatic retrieval of Tweet user status updates together with the Account metadata as supporting evidence which might be relevant in proofing authenticity before a court of law. To achieve this, the study designed spark based forensic tool to retrieve Twitter status updates together with account metadata. The table 1 above shows Twitter account schema attributes which were automatically retrieved and which the study felt can be used to authenticate the evidence received. The Forensic tool streams live tweets using Twitter API together with Spark Stream API. SHA-256 hash key for each individual tweet was generated and stored in MongoDb together with each individual user tweet both accompanied by Twitter Account unique identifier. The following figure 4 shows sample tweets which were classified as hate speech and stored within the MongoDb together with SHA-256 hash key for full tweet and SHA-256 hash key for tweet text status update.



```
db.getCollection('LiveclassifiedTweets').find({"category" : "Ethnicity"})
```

```
/* 1 */
{
    "_id" : ObjectId("59be806dbcc9834a88131c7d"),
    "tweet_id" : NumberLong(909417241334226944),
    "Name" : "Kibet®",
    "ScreenName" : "LampardMutai",
    "originalProfileImageURL" : "http://pbs.twimg.com/profile_images/878577413030019073/5Dwl3xE6.jpg",
    "source" : "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "useraccount_id" : NumberLong(936639992),
    "location" : "Konoin Elburgon",
    "geoLocation" : "0.0, 0.0",
    "Hashtag" : "",
    "Accountcreationdate" : "09-11-2012",
    "Accountcreationtime" : "02:06:18 PM",
    "originaltext" : "But i hear Kalenjins warn Kikuyus of unspecified consequences if they will not vote for you in 2022 https://t.co/i7vxf4f0kZ",
    "tweetCreationdate" : "17-09-2017",
    "tweetCreationtime" : "05:02:21 PM",
    "text" : "but i hear kalenjins warn kikuyus of unspecified consequences if they will not vote for you in",
    "prediction" : 0.0,
    "cDate" : "17-09-2017",
    "cTime" : "05:02:20 PM",
    "userStatusUpdate" : "StatusJSONImpl{createdAt=Sun Sep 17 17:02:21 EAT 2017, id=909417241334226944, text='But i hear Kalenjins warn Kikuyus of unspecified conseq
    "originaltextHashkey" : "78a506a904dd2d00ac5753a1177d52da40e9a0532dc4b223ad97655572133016",
    "usertweetmd5hash" : "93c2ed85f514016dd9a2079d0eb3ee9883f09db2c02a8804097444adff912708",
    "category" : "Ethnicity"
}

/* 2 */
{
    "_id" : ObjectId("59be81d2bcc9834a88131d46"),
    "tweet_id" : NumberLong(909418735357890566),
    "Name" : "Josekid Krystal Vybz",
    "ScreenName" : "VybzPalmer",
    "originalProfileImageURL" : "http://pbs.twimg.com/profile_images/901497149292240896/BdzqIns9.jpg",
    "source" : "<a href=\"http://twitter.com/download/android\" rel=\"nofollow\">Twitter for Android</a>",
    "useraccount_id" : NumberLong(823627958),
    "location" : "Nairobi, Kenya ",
    "geoLocation" : "0.0, 0.0",
    "Hashtag" : "",
    "Accountcreationdate" : "14-09-2012",
    "Accountcreationtime" : "08:03:33 PM",
    "originaltext" : "RT @harrisonmumia: The Ngunjiri petition against Maraga has definitely deepened the feeling amongst other tribes that Kikuyus are selfish/…",
    "tweetCreationdate" : "17-09-2017",
    "tweetCreationtime" : "05:08:17 PM",
    "text" : "the ngunjiri petition against maraga has definitely deepened the feeling amongst other tribes that kikuyus are",
    "prediction" : 0.0,
    "cDate" : "17-09-2017",
    "cTime" : "05:08:17 PM",
    "userStatusUpdate" : "StatusJSONImpl{createdAt=Sun Sep 17 17:08:17 EAT 2017, id=909418735357890566, text='RT @harrisonmumia: The Ngunjiri petition against Maraga
    "originaltextHashkey" : "52dd948d222b9636fab3d0d83fcff54e1f7b0e69ace36a8a0dba18822e25c303",
    "usertweetmd5hash" : "ecdc7140847ef4a04d93f2ff8cb269ad5c3887b763c63c3389dc8007a2ffc85e",
    "category" : "Ethnicity"
}

```

Fig 4: Sample Classified Twitter Tweet

4.6 Evidence Preservation

Spark Framework provides Streaming API which divides data in stream of batches in every predefined time internal normally in seconds called Discretized Stream (DStream). In Apache Spark, these sequence represents RDDs. In this study, the study utilized stream interval of five seconds and in every batch, an SHA-256 hash key for the tweet status update was calculated and each tweet post (text) hash key was also generated. The Spark forensic application processes the received RDDs using Spark APIs, and the processed results of the RDD operations are returned in batches. Figure 5 and figure 6 below shows stream batches arriving in time interval.

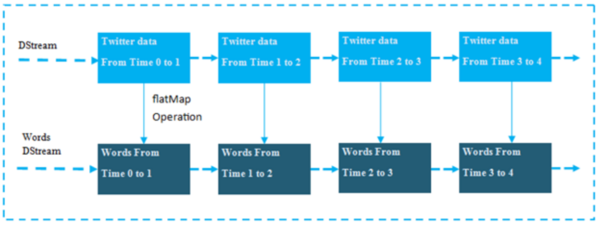


Fig 5: Spark Streaming Dstreams

Discretized Stream (DStream) forms the basic abstraction provided by Spark Streaming and represents a continuous stream of data which is received from a data source or a processed data stream generated by transforming the input stream.

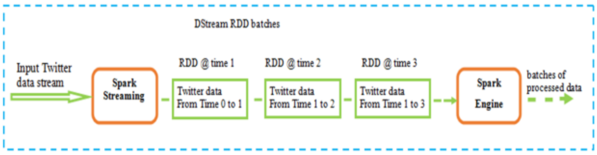

DStream RDD batches
Input Twitter data stream
Spark Streaming
RDD @ time 1
Twitter data From Time 0 to 1
RDD @ time 2
Twitter data From Time 1 to 2
RDD @ time 3
Twitter data From Time 1 to 3
Spark Engine
batches of processed data

Fig 6: Spark Streaming Dstreams RDDs

To enable repeatability and reproducibility of the captured data and evidence preservation, the system utilized spark streaming Dstreams (RDD) which are generated in batches at time interval as indicated in figure 43 and figure 44 above. The system generates SHA-256 hash key for each batch and tweet item before it is stored to Mongodb database. This ensures repeatability and reproducibility whereby data gathered can be used to reproduce the same results when using the same method on identical test algorithms or different algorithm on different labs and by different forensic analyst. This can also ensure evidence authentication before court of law to proof that captured data haven't modified after capture.

4.7 Model Design and Classification

The study collected 3,138,367 tweets (24.2GB) which were used for training the Naïve Bayes classifier. The tweets were used to train Naïve Bayes classifier model which was saved on the Spark cluster and later used for streaming live tweets and classify them for hate speech and cyberbullying. To design the model, Spark ML API (spark.ml) which provides ML pipelines (workflow) for creating, tuning, and evaluating of machine learning model was utilized. In Spark ML, a pipeline is defined as a sequence of stages, and each stage is either a Transformer or an Estimator. These stages are run in order, and the input DataFrame is transformed as it passes through each stage. Figure 8 below shows the Spark ML Pipeline stages adopted for the model design.

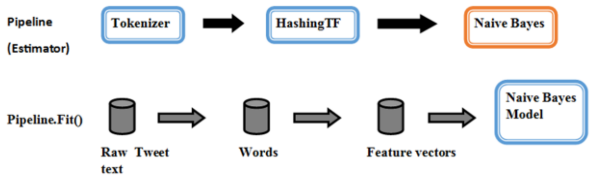


Fig 5: Spark ML Pipeline

For this study, training raw tweets were read from local disk, cleaned by passing through Scala function which removed unnecessary characters. The cleaned tweets were ingested into Spark ML Tokenizer were the tweet text were broken down into their constituent words. The tokenized tweets were again passed through Spark ML Stop Words Remover with a dictionary of stop words. This removed commonly appearing words which does not contribute to the structure of the tweets. The study split the tweet data into two datasets, 70% (2,197,498, tweets) as training dataset and 30% (940,869, tweets) as testing dataset. The training dataset was used to train Naïve Bayes model, and test dataset was used to evaluate the model accuracy. For accuracy of the model, the study used cross validation using Spark evaluation tool namely Multiclass Classification Evaluator within the spark.ml.evaluation.Multiclass Classification Evaluator and apache.spark.ml.tuning. {CrossValidator, ParamGridBuilder} packages. After the model was trained, evaluated and tested with training dataset, the model was saved on local disk within the Apache Spark Cluster. The model was later reloaded for live tweet streaming and tweet hate speech classification and categorization.

4.8 Model Evaluation

To evaluate the performance of the forensic model in terms of quality or predictive effectiveness, different metrics are used. F-measure (F1-score) is a statistical measure of model's test accuracy that is the weighted harmonic mean of precision and recall of the test where recall is the fraction of all samples classified correctly as positive by the model and Precision describes the ratio of all positives samples classified as true positives by the model. For evaluating our model, the study used spark.ml.evaluation packages which provides a suite of metrics that are suitable for evaluating the performance of spark data mining models. For this study, F-measure which is provided in spark.ml.evaluation was used to evaluate the model performance. F-measure was chosen because it includes metrics like precision and recall that are used to take into account of errors that might occur if dataset is highly unbalanced.

$$\text{Accurancy} = \frac{\text{tp} + \text{tn}}{\text{tp} + \text{tn} + \text{fp} + \text{fn}}$$

$$\text{Precision} = \frac{\text{tp}}{\text{tp} + \text{fp}}$$

$$\text{Recall} = \frac{\text{tp}}{\text{tp} + \text{fn}}$$

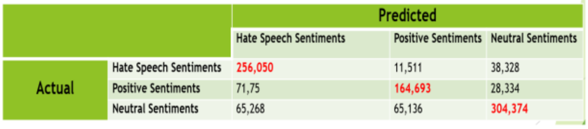
$$\text{F} = 2.\frac{\text{precision. recall}}{\text{precision} + \text{recall}}$$

The accuracy, precision by label, recall by label, and F-measure by label of the model was used to evaluate the performance of the model. Recall metric measures the overall classification correctness, which represents the percent of tweet posts that were correctly identified as hate speech. The false positive (FP) rate represents the percent of tweet posts that are not truly offensive but classified as offensive. The false negative (FN) rate represents the percent of tweets which are offensive but classified as positive tweets. Precision presents the percent of identified tweets that are truly offensive messages, and f-score represents the weighted harmonic mean of precision and recall. For estimating the performance of classification model, the study used Cross-validation (CV) which is a method for evaluating the performance of a model classifier for unseen data. Cross-validation (CV) works by randomly splitting the whole labeled data set K (K-folds) equal partitions. For each data partition, the classifier is trained on the remaining K-1 partitions and is tested on data from that partition and the final accuracy of the model is calculated as the average of all K accuracies. The following table 2 outlines the model performance as evaluated using spark.ml.evaluation library and upon cross validation against 10 folds (K-folds).

Table 2. Model Evaluation

Multiclass Metrics	**Fraction**
Model Accuracy	0.7705571409937039
Weighted precision	0.7776074500404562
Weighted recall	0.7705571409937038
Weighted F1 score	0.770139251942318
Weighted false positive rate	0.11993472033775189

The study also employed use of confusion matrix which is a matrix where rows represent actual classes and columns represent predicted classes to see the classifier effectiveness. The following figure 10 show confusion matrix which was generated using Spark ML API.



		Predicted		
		Hate Speech Sentiments	Positive Sentiments	Neutral Sentiments
Actual	Hate Speech Sentiments	256,050	11,511	38,328
	Positive Sentiments	71,75	164,693	28,334
	Neutral Sentiments	65,268	65,136	304,374

Fig 6: Confusion Matrix

This formed 940,869 (30%) of the testing tweets and as per the above table, 256,050 Tweets were correctly classified as containing words which are offending, insulting, intimidating, inflammatory or bullying in nature while 164,693 Tweets were correctly classified as Positive Sentiments and 304,374 Tweets were classified as of Neutral Sentiments. Out of 940,869 Testing Tweets, 215,752 Tweets were wrongly classified as either hate speech Sentiments, Positive sentiments or Neutral Sentiments which formed 23%.

5. MODEL RESULTS AND ANALYSIS

From the Model reports that was presented in bar charts and pie charts, it was evident that bullying and sexual related offensive language/hate speech was rampant within Twitter social network with 32.8% and 19.2% respectively. They were followed by violence related hate speech that formed 10.7%. Ethnicity and Religious related hate speech formed the lowest with ethnicity representing 7.51% and 0.467% respectively. At the same time, the study had tweets that did not fall with the range of buying, sexual, ethnicity, religious or violence and they were categorized as others and formed 29.4% of the tweets streamed. The following two charts shows sample of the forensic report represented using pie chart and a bar chart.

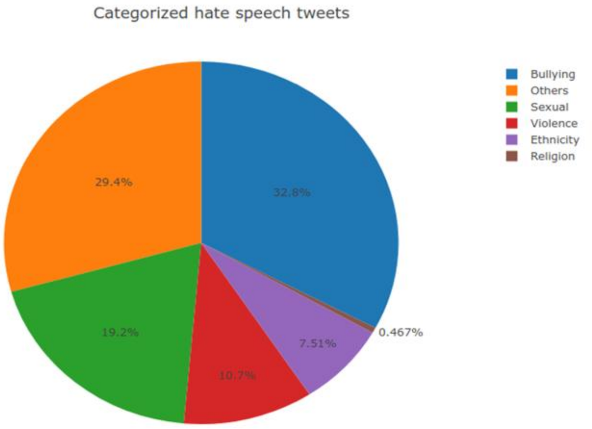


Fig 7: Categorized Hate Speech Tweets Pie Chart

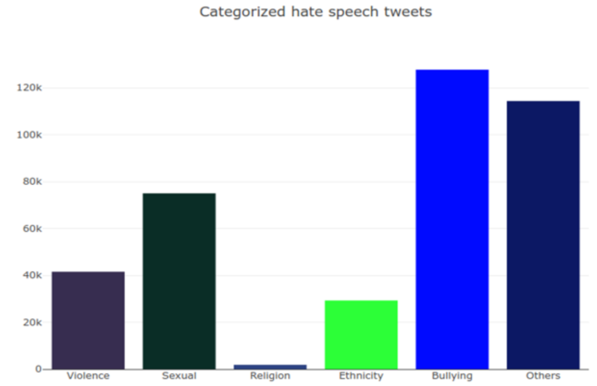


Fig 8: Categorized Hate Speech Tweets Bar Chart

6. CONCLUSION

The study designed a forensic tool that analyzes tweets sentiments for hate speech and cyberbullying using spark machine learning techniques. The classification algorithm was implemented in Apache Spark cluster using the Apache Spark's Machine Learning library, namely Spark ML API. The study relied on distributed contributing framework Apache Spark and made use of Spark streaming API to stream Twitter data and Spark ML API for tweet analysis and classification. the study designed Naïve Bayes model by utilizing a dataset of 2,197,498 tweets to train the model and 940,869 tweets for testing. The model was used to stream and classify hate speech and detect cyberbullying for Kenyan based hate speech during 2017 general election and following the nullification of presidential election. The model was able to successfully detect and classify hate speech which mostly were ethnic based. The study demonstrated how twitter social network data can be collected and preserved within Mongodb database for forensic analysis to ensure its authenticity before court of law and ensure forensic reproducibly. The study shown that by generating SHA-256 hash key for each twitter item within DStreams and saving the hash key with each individual tweet item in database can be used to detect changes to the data stream during analysis or different forensic analyst can verify the twitter data and thus repeatability/reproducibility of the forensic data can be done.

This feature can be used for forensic evidence preservation and ensure changes to the streamed evidence data can be detected by regenerating the SHA-256 Hash key and comparing it with already stored tweet item key in Mongodb database. The study also has shown that by preserving each tweet stream date and time can be used to document the acquisition of the evidence hence improving the chain of custody. The forensic tool was able to ensure chain of custody by maintaining "When" the evidence was captured (Date/Time), when each tweet was created/posted, "Where" the evidence was posted from (source) and tweet ID. Twitter page printouts and screenshot may not be authenticated or allowed as evidence before a court of law because they lack indication or proof of its creator, source, or custodian. The study has demonstrated which twitter account metadata might be relevant in forensic analysis of twitter posts and how it can be captured. It was evident that a lot of cyberbullying and hate speech is rampant in twitter social media and when the data is well retrieved and preserved, it can form basis for forensic investigation. However, the issue of the dynamic nature of the updates makes it a challenge which calls for real live streaming of social media data which might demand large storage space.

As future work, the study plan to extend the forensic tool to include all other social media with capability to provide hot maps which indicates the specific region where such hate speech was posted on social media. It will also involve using other machine learning algorithms to try and increase the effectiveness of the tool in identifying and categorizing hate speech and cyber bullying on social network sites

7. ACKNOWLEDGMENTS

I wish to convey my appreciation and special thanks to my supervisor Elisha O. Abade for his dedication and assistance throughout the research process, the members of the panel whose knowledge and experience in this field has been of great help to my research and the whole School of Computing and Informatics for their support that made me deliver in this work.

8. REFERENCES

[1] Edwards, D., Computer Forensic Timeline Analysis with Tapestry. 2011, SANS Gold Paper accepted November.

[2] Press, E.-C., Computer Forensics: Investigation procedures and response. Course Technology Cengage learning, USA, 2010.

[3] Johnsen, J.W., Algorithms and Methods for Organised Cybercrime Analysis. 2016.

[4] Gupta, R. and H. Brooks, Using Social Media for Global Security. 2013: John Wiley & Sons.

[5] Berman, J.J., Principles of big data: preparing, sharing, and analyzing complex information. 2013: Newnes.

[6] Wijeratne, S., et al. Analyzing the social media footprint of street gangs. in Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on. 2015. IEEE.

[7] Agarwal, S. and A. Sureka, Applying social media intelligence for predicting and identifying on-line radicalization and civil unrest oriented threats. arXiv preprint arXiv:1511.06858, 2015.

[8] Chen, Y., et al. Detecting offensive language in social media to protect adolescent online safety. in Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom). 2012. IEEE.

[9] Baesens, B., V. Van Vlasselaer, and W. Verbeke, Fraud Analytics Using Descriptive, Predictive, and Social Network Techniques: A Guide to Data Science for Fraud Detection. 2015: John Wiley & Sons.

[10] Patzakis, J. Overcoming Potential Legal Challenges to the Authentication of Social Media Evidence. 2012 [cited 2016 17/12/2016]; Available from: https://articles.forensicfocus.com/2012/04/02/overcoming-potential-legal-challenges-to-the-authentication-of-social-media-evidence/.

[11] Ncr, P.C., et al., CRISP-DM 1.0. 1999.