Automated Crime Tweets Classification and Geo-location Prediction using Big Data Framework

Dr.K. Santhiya[1], Dr.V. Bhuvaneswari[2], V.Murugesh[3]
[1] Assistant Professor of Information Technology, Sri Krishna Adithya College of Arts and Science, Coimbatore.
[2] Associate Professor, Department of Computer Applications, Bharathiar University, Coimbatore, Tamil Nadu, India
[3]Assistant Professor of Information Technology, Sri Krishna Adithya College of Arts and Science, Coimbatore.
[1]krsanthiya@gmail.com [2] bhuvanes_v@yahoo.com [3]murugesh74@gmail.com

Abstract: This paper investigates with the automated classification of tweets which turns out to be a very complicated problem because of its nature, heterogeneity and the amount of data. According to internet live stats, nearly 500 million tweets are tweeted per day, where the user's opinion about different topics is shared. An automated decision support system is developed to analyze the tweets related to crime against women and children. The problem is viewed in a big data perspective because of the nature of data. The proposed work focuses on developing two systems: Hadoop MapReduce and Apache Spark framework for programming with Big Data. The algorithm based on hierarchical domain lexicon classifies different types of crime in a parallel and distributed manner. Moreover, the crime classification tool is based on hybridized Machine Learning techniques combined with Natural Language Processing techniques. To predict the location of twitter users, multinomial Naive Bayes classifier trained on Location Indicative terms and other vital parameters (such as city/country names, #hash tags and @mentions) is implemented. Our approach outperforms in terms of classification accuracy, mean and median error distance when compared with other algorithms based on parameters such as Location Indicative terms, #hash tags and city/country names.
Keywords: Crime; Twitter; Naive Bayes; Map Reduce; Spark; Natural language Processing; Geo-location.

I. INTRODUCTION

Women of all ages are highly vulnerable to crimes and have increased across years. Women found her totally suppressed and subjugated heavily under patriarchal-male-dominated and male-identified society. Their lives in these kinds of society are the harshest. These harsh conditions often include crimes perpetrated by men against women including humiliation, harassment, rape, torture, murder and exploitation of women. With the advent of social media services such as Twitter, substantial amount of public information is available,

which reflects the real opinion of people on different aspects of life. It is used by large group of users to express their opinions on various social issues (Pang et al. 2008). Therefore, for classifying crime into different categories, the data generated by social media are considered to be a powerful tool. In this work, tweets related to crime committed against women and children are extracted and analysed in order to provide valuable insights to crime analyst regarding the location where the crime occurred most frequently, the type of crimes committed, the opinion of the people and the time when more crime tweets are tweeted. Our approach differs from the conventional crime classification model by taking into consideration the opinion of social media data. As the data generated by social media are purely unstructured, there lies another big challenge with respect to storage and processing of these unstructured free texts. Since twitter is also a real time unstructured data, it is very much essential to define highly scalable solutions.

Problem Statement

Automated crime report analysis and classification is one of the strongest tools that can be used by both data journalist and crime analyst to process and analyze anonymous crimes efficiently within a short span of time (Ku et al. 2011). However, traditional crime classification models with tweets as primary input possess certain limitations on reflecting real time criminal incidents against women and children. In order to enhance the classification accuracy, maximize the power of categorization and to lessen the time and storage space requirement, we set five objectives.

- Extract the real time tweets using Twitter Streaming API
- Pre-process the tweets in order to remove punctuation, stop words and URLs. Classify the tweets based on emoticons and hash tags as sentiment labels.
- Perform NLP techniques and one of the machine learning approaches called classification to effectively categorize the crime labels.
- Proposed a novel distributed framework implemented in an open source platform Hadoop as well as in Spark.
- Integrate bloom filters to improve the performance of the algorithm.

The rest of the article is organized as follows: Section 2 discusses on related work with respect to crime related applications, Geo-location prediction and text classification. Section 3 explains the text-oriented decision support framework developed and the system components. The experimental evaluation and the results obtained are detailed under Section 4. Eventually, Section 5 explains conclusions and directions for future work.

II. RELATED WORK

Crime Related Applications

Crime Analysis is an efficient combating tool used in law enforcement agencies and government. Lot of research work is done by various researches in

this domain due to tremendous increase in crime rate in the last few years. Accurate real time crime predictions help to decrease the crime rate. Therefore analysis and prediction of crime is a vital activity that can be optimized using various techniques and processes. Now-a-days most of the researchers focused on crime trend analysis, pattern discovery, crime link analysis, geo-spatial visualization and spatio-temporal crime analysis.

HiteshKumarReddy ToppiReddy et al. (2018) proposed a crime prediction and monitoring framework for predicting the crime and to help the law enforcement agencies. The proposed framework used various visualization techniques to map the crime trends and different machine learning algorithms to predict the crimes using Google Maps and various R packages. In this work, the author concentrated only on the location prediction of the crime. Ahishakiye E et al., (2017), Nasridinov, A et al. (2013), Iqbal, R. Murad et al., (2013) focused on the implementation of decision tree classifiers for crime prediction. But the authors focused on the structured data.

Arushi Jain et al. (2016) proposed a big data analytic framework to analyze crime trends . The author captured the data using Flume and Scoop, stored the data in HDFS and evaluated crime data analytics using pig. Here the author predicted only the count of occurrences of crime. M.S. Gerber et al. (2014) proposed a novel approach for crime prediction using spatio-temporally tagged tweets. In this research work, the author has considered only the geo-tagged tweets. Satya Katragadda et al. (2014) proposed and assessed an unsupervised approach to identify the location of a user absolutely dependent on tweet history of that user. But the author acquired only 62% of accuracy in identifying the location.

Agarwal, P et al. (2012) proposed an approach to extract location information from tweets. The author combined the Stanford NER tool and a concept-based vocabulary to identify the location from tweets. Yu, R et al., (2014) provided the static maps to plot the crime hotspots. Bao Wang et al. (2017) adapted the state-of-the-art deep learning spatio-temporal predictor to predict the distribution of crime over the Los Angeles area. Sheila Kinsella et al. (2011) constructed a language model to predict the location of an individual tweet as well as the location of the user using the coordinates extracted from the geo-tagged tweets.

While the rest of papers concentrated on structured data stored in databases, they often ignore information from unstructured sources. The proposed work concentrates on extricating information related to crime from unstructured data and developed a classifier using big data framework to classify the crime data, identify the crime hotspots and also to assist data journalist and crime analyst by providing them with the crime analysis reports.

Text Classification

The purpose of text classification is to automatically label the text documents into one or more pre-defined categories. Since enormous volume of data gets accumulated from various sources such as emails, blogs, social networks, web

pages and even information produced by big enterprises are also digitized, automatic text classification has gained priority in many research domains. Lee et al. (2011) applied text-based and network-based classification models to classify tweets into 18 trending topics such as sports, politics, etc., Mahendran et al. (2013) applied Naive Bayes and MaxEnt classification algorithms to classify microblogs into pre-defined class labels (positive, negative or neutral) using Bag of Words feature set .Current classification methods such as decision trees (Diao et al. 2000, Vens et al. 2008), k-nearest neighbours (Li et al. 2011, Tan 2006, Wan et al. 2012) , neural networks (Ghiass et al. 2012, Rajan et al. 2009) , support vector machines (Li et al. 2011, Rajan et al. 2009, Wan et al. 2012) and Naive Bayes (Bermejo et al. 2011, Isa et al. 2009, Tian et al. 2009) have been successfully used in automated text classification. Duwairi et al. (2014) conducted a comparison among Naive Bayes, SVM and K-nearest Neighbour classification algorithms to classify sentiments of tweets on various topics like education, sports and politics as positive, negative or neutral .

Research Question

Most of the crime-related articles which was published has focused on text mining operations based on structured data (Piskorski et al. 2010) , crime classification (Borg et al. 2014) , crime analysis and visualization (Kovachev et al. 2008) , and distance measure for determining similarity between criminal investigations (Cocx et al. 2006) . However, most of the methods are designed to handle adequate amount of structured data rather than unstructured crime data (Helbich et al. 2013) . It is also notable that due to lack of a domain-specific lexicon, it is difficult to get deep insights out of unstructured data (Pinheiro et al. 2010) . In addition to it, many of these articles implemented the classification algorithms in conventional platform which could not handle enormous volume of text data. The gaps paved way to study and enquire upon the common research question:

Can a proposed work that integrates the domain-specific lexicon, various corpus and algorithms with classification approaches implemented using Big Data framework achieve high performance in classifying big datasets of crime tweets?

This paper concentrates in the development of a big data framework (an automated DSS) for data extraction from unstructured text. Certain information extraction techniques, domain-specific lexicon embedded in the form of corpus and classification algorithms are required for unstructured data extraction and utilization.

III.SYSTEM DEVELOPMENT AND DESIGN

The architecture of the text-oriented decision support system for processing the crime related tweets using big data framework is presented in Figure 1. The four-layer architecture comprised of information extraction and pre-processing layer, similarity computation layer, text classification layer and presentation layer. To

accomplish the tasks mentioned in each of these layers the system components shown in Figure 2 is needed.

System Components

As we are dealing with massive volume of data which gets generated at a faster rate, we could not rely upon traditional databases or tools to perform computation. So we are in need of big data framework called Hadoop which is suitable for large batch processes. But we could not achieve faster performance using Hadoop, as the data are flushed to the disk instead of memory (White 2012). In contrast, Spark performs better than Hadoop as it maintains the data in the cache memory (Karau et al. 2015)

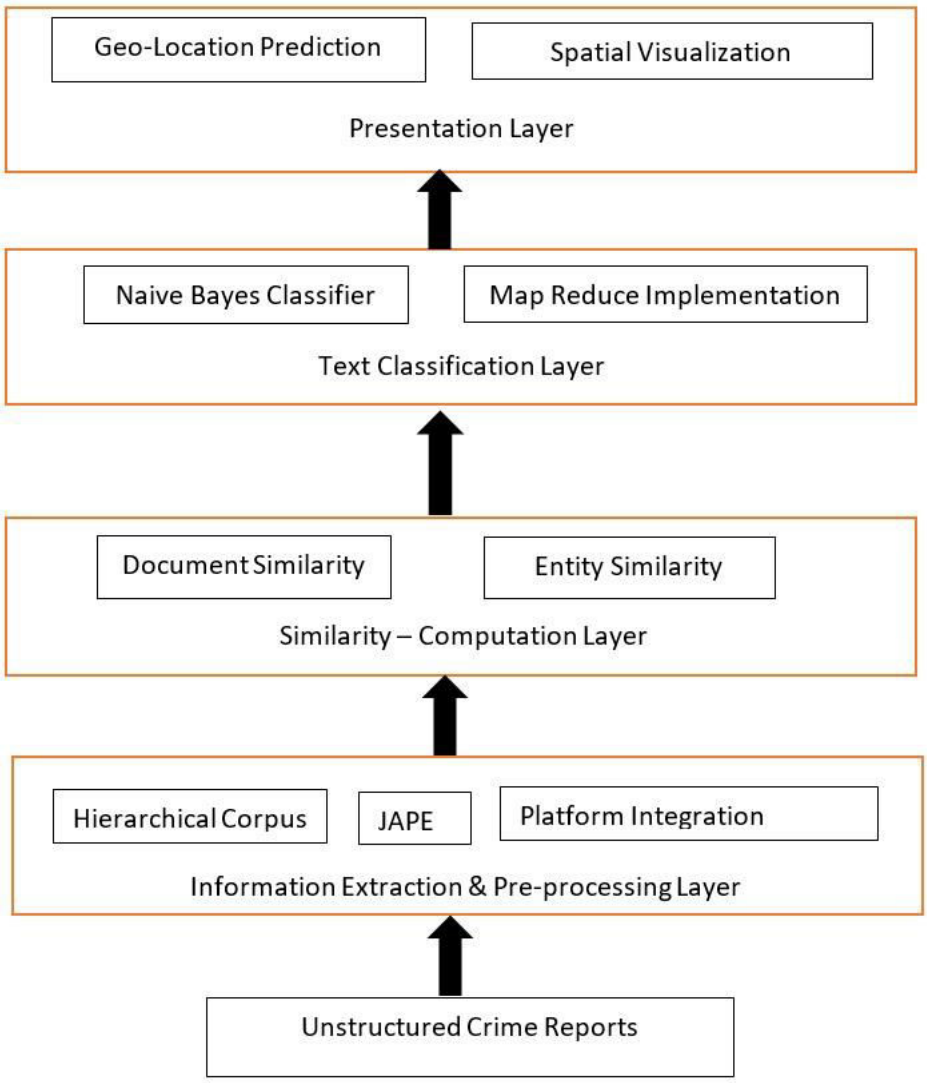

Figure 1. Text-Oriented Decision Support System

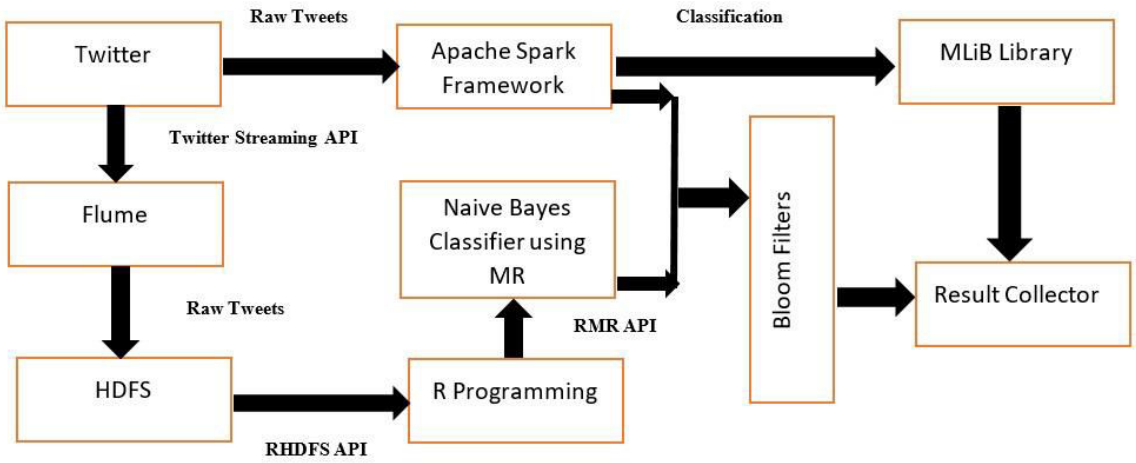

Figure 2. System Architecture

In this proposed work, we make use of these two different big data framework, in order to improve the scalability of the storage space and to reduce the computation time. One of the machine learning algorithms called classification is implemented to classify the stored crime tweets into appropriate crime categories. In order to make the algorithm both time efficient and space efficient, we integrated bloom filters, the data structures proposed by bloom in our big data framework (Nodarakis et al. 2016)

To extract tweets and to store it in Hadoop distributed file system (HDFS) in real time at a faster rate, we make use of twitter streaming API called Flume. We rely upon Apache Flume, a data ingestion mechanism for transporting data from twitter data source into HDFS, as it is highly reliable, distributed and configurable (Katarial et al. 2014). The raw tweets stored in HDFS are pre-processed in R programming environment by integrating the two platforms R and Hadoop by means of an API called RHDFS. After pre-processing, now the tweets are readily available for classification. In this work, we make use of Naive Bayes classifier, as it is robust and from several literature papers we come to know that this algorithm outperforms other classification algorithms like SVM, logistic regression etc. This algorithm is implemented using parallel programming called Map Reduce under two different big data frameworks (Lin et al. 2010)

Text-Oriented Decision Support System

The four-layer architecture mentioned above consists of several internal components to accomplish the task mentioned in each of these layers.

Information Extraction and Pre-processing Layer

With the help of above mentioned system components and APIs we extracted real time tweets which is purely unstructured free flow of text. These tweets consist of both informative and uninformative text (Chowdhury 2003). So, we have adapted and used several components to process the tweets and to extract relevant information out of it. This layer is composed of eight components:

Lowercase Converter, Tokenizer, Sentence splitter, POS tagger, stemmer, gazetteer, Java Annotations Pattern Engine (JAPE) and Information filter (Banerveld et al. 2014). To analyze each and every individual words and sentences in crime tweets, the first four components mentioned in the system are utilized. While the next two components concentrate on extracting relevant noun and verb phrases out of crime tweets. Majority of stop words which is of less important while processing the tweets are removed using Information filter. Table 1 shows the brief overview of these components.

Table 1: Components of Information-Extraction and Pre-processing Layer

Components	**Explanation**
Lowercase Converter	Convert the text into lowercase letters
Tokenizer	Breaks the text into individual words and output a stream of tokens
Sentence Splitter	Identify the boundaries of tweets in text
POS tagger	Each word in the crime tweets are labelled as noun, adverb and adjective using POS tagger
Stemmer	Reduces the inflected words to its core form
Gazetteer	The gazetteer is a geographical dictionary containing group of words or indices to locate entities such as type of crimes and location. The geographical dictionary containing nearly 28,000+ words are organized in the form of hierarchical lexicon. Our new lexicon is represented in the form of 17 semantic trees containing 28,000+ words and phrases. Each tree has one root node and many levels of child nodes. The root node serves as the main class and the child nodes serves as the subclasses of the classification.
JAPE rules	This particular engine is coded to extract entities like locations , age of the victim or perpetrator, names and type of crimes committed.
Information Filter	Removes meaningless words, redundant entries so that relevant information can be maintained.

Similarity Computation Layer

The similarity computation layer consists of two components namely entity similarity and document similarity (Aliguliyev 2009). Our aim is to analyze and categorize the tweets into specific crime types. Beforehand to analyze the tweets completely and to tag them accordingly, we implemented the word matching algorithm, N-gram approach (varying from unigram to trigram) as well as word association.

Entity Similarity

To perform entity similarity (or) word similarity, the word matching algorithm is implemented by making use of the following corpus (Online 2017):
Lexical Corpus: It comprises of most of the English words, through which the tweets can be analyzed and segregated by matching the word in the tweet with the words in the lexical corpus. It also contains common English phrases, headwords, idioms and multiword (Ajinkya Ingle et al. 2015).
Emoticon Corpus: Emoticons essentially portray the tweeter's mood and it gives certain meaning for the tweets (Online 2017). So, with the help of this corpus, the emoticons present in the tweets can be matched and analysed (Yamamoto et al. 2014)
Acronym Corpus: All the acronyms and abbreviations present in the tweets can be elaborated using this corpus and the word can be matched with that of the lexical corpus.
Crime Corpus: It consists of list of all crimes committed against women and the type of crime that occurs in the tweets can be matched and tagged with the help of this corpus.
The raw tweets are pre-processed and then sent through word matching algorithm which performs matching of words in the tweets with that of the words maintained in the corpus list (Chirag Kansara et al. 2016).

Document Similarity
Text-Transformation using N-gram Approach

In this step, tweet contents are represented using vector of features. The frequencies of the single word (unigram) frequency, two-word (bigram) and the three-word (trigram) sequences were determined from the dataset of tweets. The corpus is tokenized into N-grams by setting the unigram as minimum and trigram as maximum gram. We set the sparse value as 0.98 to remove sparse value from these N-grams. So, the terms that occur at most in less than 0.02 corpuses are removed. Then the BOW (Bag of words) vector is created by finding the frequent terms observed in selected terms. Having computed the initial word frequencies, certain other transformations are carried out to encapsulate and aggregate the extracted information.
Log-frequencies

The raw word or term frequencies denote the importance of a word in each document. Prosaically, certain words can also be referred as better descriptors of the contents of that document, if it occurs with greater frequency (Ku et al. 2008). But we cannot declare that word counts themselves are proportional to their importance as descriptors of the documents. Thus, we need to compute the common transformation of the raw word frequency counts (tf) as shown in Eq.1

$$f(tf) = 1 + \log(tf), for tf > 0 \quad \text{---------- Eq. (1)}$$

This transformation will "lessen" the effect of the raw frequency counts on consecutive computations and analyses.
Binary frequencies

The other simpler transformation technique as shown in that can be used to determine whether a term is used in the document.

$f(tf) = 1\ , for\ tf > 0$---------- Eq. (2)

The presence or absence of the respective words is indicated as 1s and 0s accordingly in the resultant term document matrix. Again, this transformation will also reduce the effect of the raw frequency counts on subsequent computations and analyses.

Inverse document frequencies

The inverse document frequency (for the i^{th} word and j^{th} document) contemplates both the terms specificity (document frequencies) as well as the overall frequencies of their occurrences (term frequencies).

$$idf(i,j) = \begin{cases} 0 & if\ tf_{i,j} = 0 \\ (1 + \log(tf_{i,j}))\log\frac{N}{df_i} & if\ tf_{i,j} \geq 1 \end{cases}$$ ---------- Eq. (3)

In this formula, the total number of documents are represented as N, df_i represents the document frequency for the i^{th} word (the number of documents that include this word or term). Hence, it is obvious from the formula that the simple word frequencies can be represented in terms of log functions. The weighting factor is also assigned in such a way that it evaluates to 0 if log (N/N=1)i.e. (the term occurs in all documents), and it is also assigned to a maximum value when a word occurs only in a single document (log(N/1) =log(N)). It is very clear that this transformation will create indices that both reflect the relative frequencies of occurrences of words, as well as their semantic specificities over the documents included in the analysis. This vector is then used to create the N-gram tokenizer of testing dataset.

Text Classification Layer

This section deals with the evaluation of classification process under Hadoop framework. We manipulated the existing map reduce naive bayes classifier to address the essentials of opinion mining problem (Dean et al. 2008). Our algorithm consists of four pipelined map reduce jobs in order to execute four consecutive steps:

Feature Extraction: Extract the features from all tweets in training (T) and test set (TT).

Feature Vector Construction: Construct the feature vectors for training (F_T) and test set (F_{TT}).

Probability Computation: For each vector $v \in F_{TT}$, find the matching vectors in F_T.

Crime Classification: Assign a crime label $\forall tt \in TT$.

The records provided as input to our algorithm have the format <tweet_id, class, text >, where class refers to the crime label for tweets in T. The detail analysis of

Map and Reduce functions that takes place in every MapReduce job are described separately in the following subsections.

Feature Extraction

The features are extracted from the training (T) and the test set (TT) and their weights are calculated in the first MapReduce job (Chen et al. 2009). The inverted index is generated as output of the job in the form of key-value pairs, where the features serves as a key and the value is a list of tweets that contain those features. In the MapReduce Job1, we sum up the Map and Reduce functions of this process.

The records from T and TT are given as input to the Map function and the tweets features are extracted out of it. A key-value record is generated as an output for each feature, where the key represents the feature and the value consists of tweet id, the class to which the tweet belongs to and the number of occurrences of a feature inside the sentence . The above mentioned key-value pairs are given as input to the Reduce function. The Reduce function computes the weight of a feature in each sentence. Then, it forms a list l with the format< t1, w1, c1:...:tx, wx, cx >, where ti is the id of the i-th tweet, wi is the feature weight of i-th tweet and ci is class to which i-th tweet belongs to. For each key-value pair, the Reduce function outputs a record where the feature is the key and the value is list l.

Feature Vector Construction

The feature vector is constructed by combining all the tweets features into one single vector. Moreover, ∀tt ∈TT, we generate a list of training tweets in T that share at least one word or n-gram. Initially ∀f ∈ F, the tweets containing f is separated into two lists—training and test, respectively. And also it put forth a key-value record ∀f ∈ F, where the key is the tweet id that contains f. In addition, the value represents f and weight of f. Next, ∀v∈test generates a record where the key is the id of v and the value is the training list. The key-value pairs with the similar key are gathered and FT as well as FTT are constructed by the reduce function. For each tweet t ∈ T (tt ∈ TT), it outputs a record where key is the id of t (tt) and the value is its feature vector (feature vector together with the training list).

Probability Computation and Crime Classification

We have a document D, and set of classes C. The posterior probability P(C|D) is computed to find the corresponding class to which the document D belongs to. P(C|D) can be computed by Bayes' Theorem as shown in Eq. (4)

$$P(C|D) = \frac{P(D|C)P(C)}{P(D)} \alpha\, P(D|C)P(C)$$ ---------- Eq. (4)

where the prior probability and likelihood can be computed from the labelled dataset. The prior probability $P(C_i)$ can be computed easily, since every class is equally probable. Let X be a corpus, then in the training data $P(w_j \mid C_i)$ represents the probability of j-th word belonging to a class C_i.Let us assume that $TT_i$is the i-th

tweet in my test dataset and $F_j$is the feature vector for this tweet. Then the probability of tweet TT_ibelongs to class C_i can be computed as shown in Eq. (5)

$P(TT_i| C_i) \alpha P(F_j| C_i) = \prod_{j=1}^{|X|}[F_jP (w_j | C_i) + (1- F_j) (1-P (w_j | C_i))]$ ----------Eq. (5)

From the equation it is crystal clear that it represents the multiplication of the probabilities that this tweet is composed by words in the corpus.

Presentation Layer

The presentation layer consists of two sub components namely geo-location prediction and spatial visualization.

Geo-location Prediction

By making use of various feature sets such as location indicative words(LIW), city/country names (CC), mentions and hash tags, a multinomial Naive Bayes classifier is trained (Cheng et al. 2010). Let us assume that C represents set of all cities (i.e., our labels) and T is the set of all tweets in training set. The probability P(c|t) is maximized by geotagging each tweet t ∈ T with a city c ∈ C. Bag of features approach is used and each tweet t is represented as a set of features F_i∈ t (out of N total features), where each feature F_i indicates the number of times (frequency count) that a feature word $F_i$is used in a tweet t. Given that t_c is the set of all tweets that are posted in a specific city c and T is the set of all tweets, the prior probability can be calculated as shown in Eq. (6)

$P(c) = \frac{|tc|}{|T|}$ ---------- Eq. (6)

The geo-location prediction task is performed at two levels (i) at tweet level (ii) at user level. We make use of the following evaluation metrics like accuracy, mean error distance and median error distance.

IV.EXPERIMENTS AND RESULTS

This section highlights our experimental setup, the dataset used in our experiment, the key results of our proposed work and various baselines.

4.1Dataset Description

We have extracted all the tweets between January 2016 to December 2017 by making use of Twitter Search API. The proposed algorithm is evaluated on the twitter datasets. The total dataset comprised 1,48,707 tweets with 79,848 mentions, 26,425 hashtags, 32,798 Quote tweets and 35,974 retweets. Table 2 shows the detailed count of tweets extracted under different crime categories.

Table 2: Tweets collected under different crime categories

Categories	**Tweets**	**Mentions**	**Hashtags**	**Quote Tweets**	**Retweets**
Sexual Harassment	56034	33725	6903	7792	8954
Rape	36542	19032	8463	6932	7521
Dowry Death	20239	12100	6341	4941	5874
Kidnapping &	19362	8469	2809	5501	5603

Abduction					
Stalking	7748	3692	1219	3147	3368
Groping	5857	1983	632	2845	2896
Suicide	2925	847	58	1640	1758
Total	1,48,707	79,848	26,425	32,798	35,974

Certain pre-processing tasks are carried out on the twitter dataset and only the English tweets are retained. We also used an available English dictionary to identify the appropriate English word, and do not include two or more hashtags or emoticons. Moreover, during pre-processing all the URL links, hashtags and references are replaced by URL/REF/TAG meta-words. Then we compute the term frequency as well as inverse document frequency for each feature vector as shown in Figure 3. We have also applied n-gram techniques to find the word matching as shown in Figure 4.

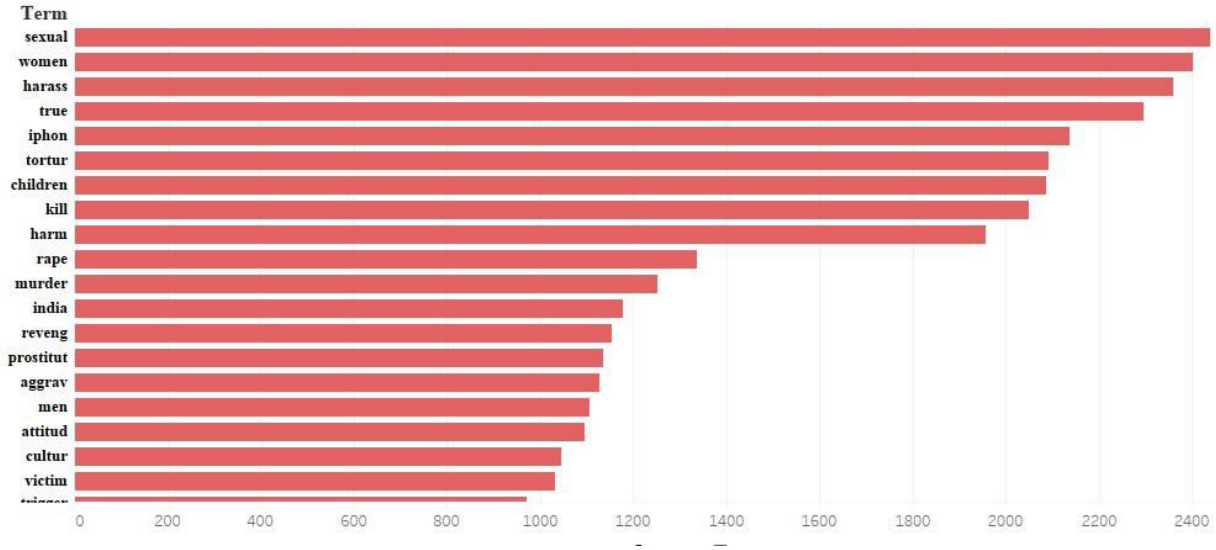

Figure 3: Frequency of terms

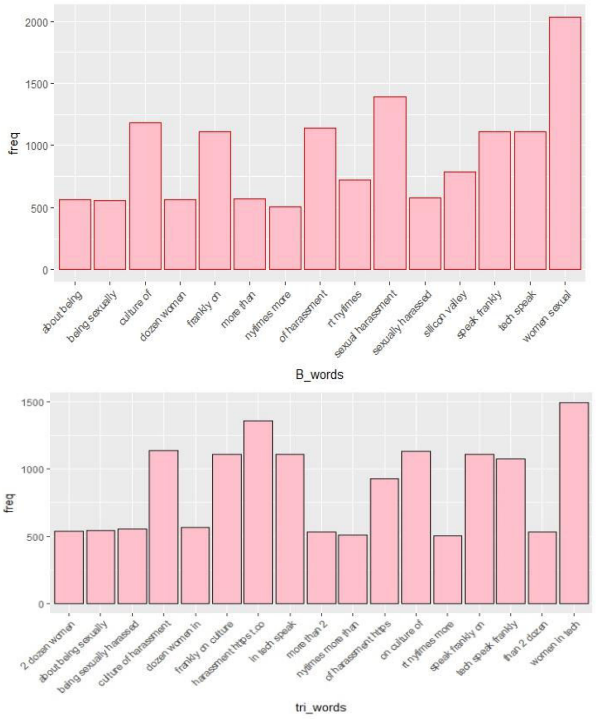

Figure 4: Frequency of Bigram and Trigram words

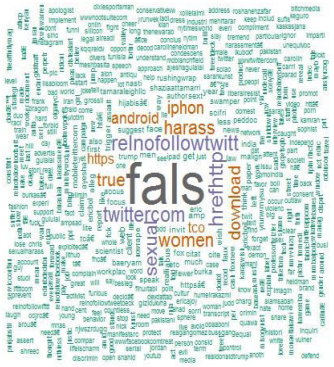

Figure 5: Word Cloud Generation

Figure 5 shows the occurrence of most frequent words in the tweets in slightly larger font and the least occurrence of words in a smaller font. Table 3 shows the probability computation for the crime label "suicide".

Table 3: Probability Matrix for the crime label "suicide"

S.No	Term	Count	Additive	Probability	Inprobabilty
1	Case	22	23	0.001465155	-6.525794379
2	Death	5	6	0.000382214	-7.869529125
3	Commit	4	5	0.000318512	-8.051850682
4	Daughter	4	5	0.000318512	-8.051850682
5	Father	3	4	0.00025481	-8.274994234
6	Police	35	36	0.002293286	-6.077769656
7	Report	10	11	0.000700726	-7.263393322
8	Court	16	17	0.001082941	-6.828075251

9	Sexual	6	7	0.000445917	-7.715378446
10	Violence	15	16	0.001019238	-6.888699872
11	Victim	76	77	0.004905083	-5.317483173
12	Revenge	95	96	0.006115429	-5.096940403
13	Arrest	30	31	0.001974774	-6.22730139
14	Rape	49	50	0.003185119	-5.749265589
15	Abuse	7	8	0.000509619	-7.581847053
16	Threaten	22	23	0.001465155	-6.525794379
17	Dowry	1	2	0.000127405	-8.968141414
18	Harass	1	2	0.000127405	-8.968141414
19	Violent	3	4	0.00025481	-8.274994234
20	Video	3	4	0.00025481	-8.274994234

Figure 6 shows the classification results for the test data. The results mention the classification of crime label "Domestic Violence".

```
> classResults
        vScores     oScores Classification
1   -148.062352 -151.892918       Violence
2   -199.711547 -155.877652          Other
3   -192.457484 -146.807380          Other
4   -215.969949 -175.125334          Other
5   -218.753594 -165.932571          Other
6   -218.363209 -169.947710          Other
7    -77.064869  -79.422597       Violence
8      0.000000    0.000000          Other
9   -124.481206  -70.937751          Other
10   -85.173675  -94.920349       Violence
11  -132.696784  -84.670084          Other
12  -218.553947 -170.820682          Other
13  -104.752276 -114.976223       Violence
14  -113.551581  -58.356590          Other
15  -222.317339 -171.028690          Other
16   -86.767609  -97.916082       Violence
17  -114.893832  -55.998247          Other
18  -229.986164 -179.943611          Other
19  -206.068160 -155.414249          Other
20  -208.643687 -159.425784          Other
21  -200.492939 -178.834235          Other
22  -192.807341 -164.810808          Other
23   -33.104598  -38.126404       Violence
24     0.000000    0.000000          Other
25   -23.071058  -27.890256       Violence
26   -24.992870  -29.499694       Violence
27   -26.938780  -28.113399       Violence
28  -110.239121  -55.060873          Other
29  -233.607015 -187.304891          Other
30  -184.370863 -147.339124          Other
31  -215.190339 -182.695996          Other
32   -17.371103  -19.004311       Violence
33  -202.883334 -150.913087          Other
34   -47.418038  -45.139520          Other
35     0.000000    0.000000          Other
36  -157.970754 -115.227642          Other
```

Figure 6: Tweets classified under crime label "Domestic Violence"

Table 4 shows the geolocation prediction results for the tweet-level, in terms of accuracy, mean and median error distances. The results show that our proposed MNB-ALL algorithm outperforms all baselines for the training and testing datasets, in terms of all three evaluation metrics.

Table 4: Tweet-level Geo-Location Prediction for both training set and test set.

Algorithm	Accuracy	Mean Error	Median Error
MNB-LIW	0.1023	9231.9379	9153.5067
MNB-CC	0.0689	12586.9015	10814.025

MNB-HASH	0.0845	5382.6759	6216.4987
MNB-MENTION	0.0559	11476.6707	9442.6731
MNB-ALL	**0.1153**	**3214.8084**	**4933.7693**

Algorithm	Accuracy	Mean Error	Median Error
MNB-LIW	0.135	7779.5598	7352.0442
MNB-CC	0.0892	11258.3635	9687.0232
MNB-HASH	0.0961	5432.3149	6358.0754
MNB-MENTION	0.0561	9588.3402	9116.1263
MNB-ALL	**0.1372**	**3524.8263**	**5224.7842**

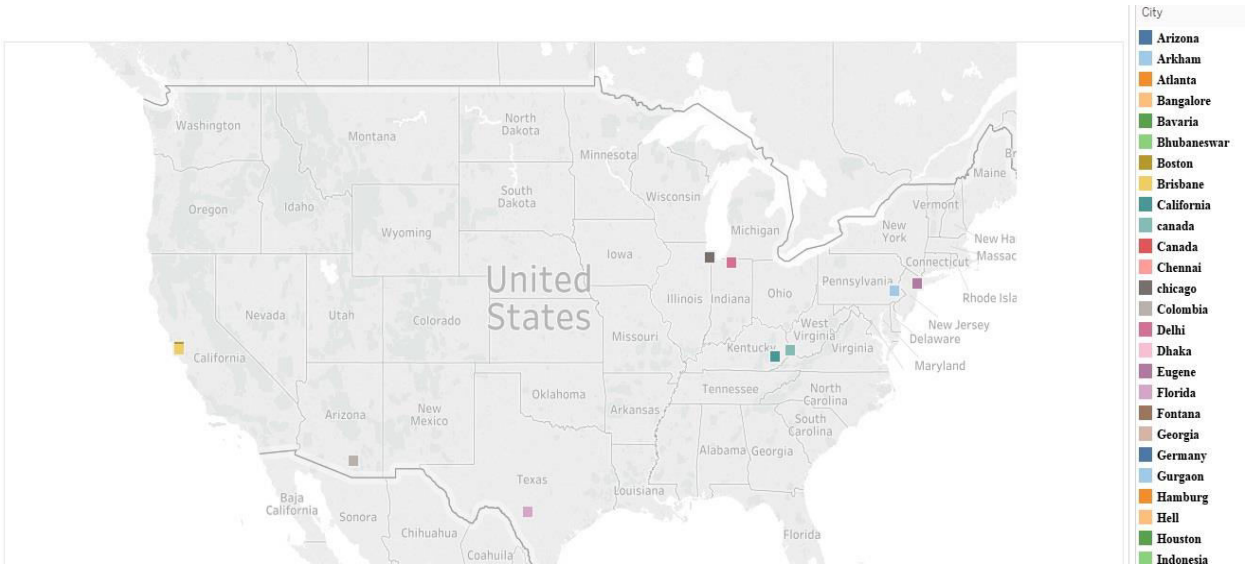

Figure 7. Location Prediction using proposed MNB-ALL algorithm

Figure 7 shows the location where the crime against women occur more frequently compared to any other parts of the world.

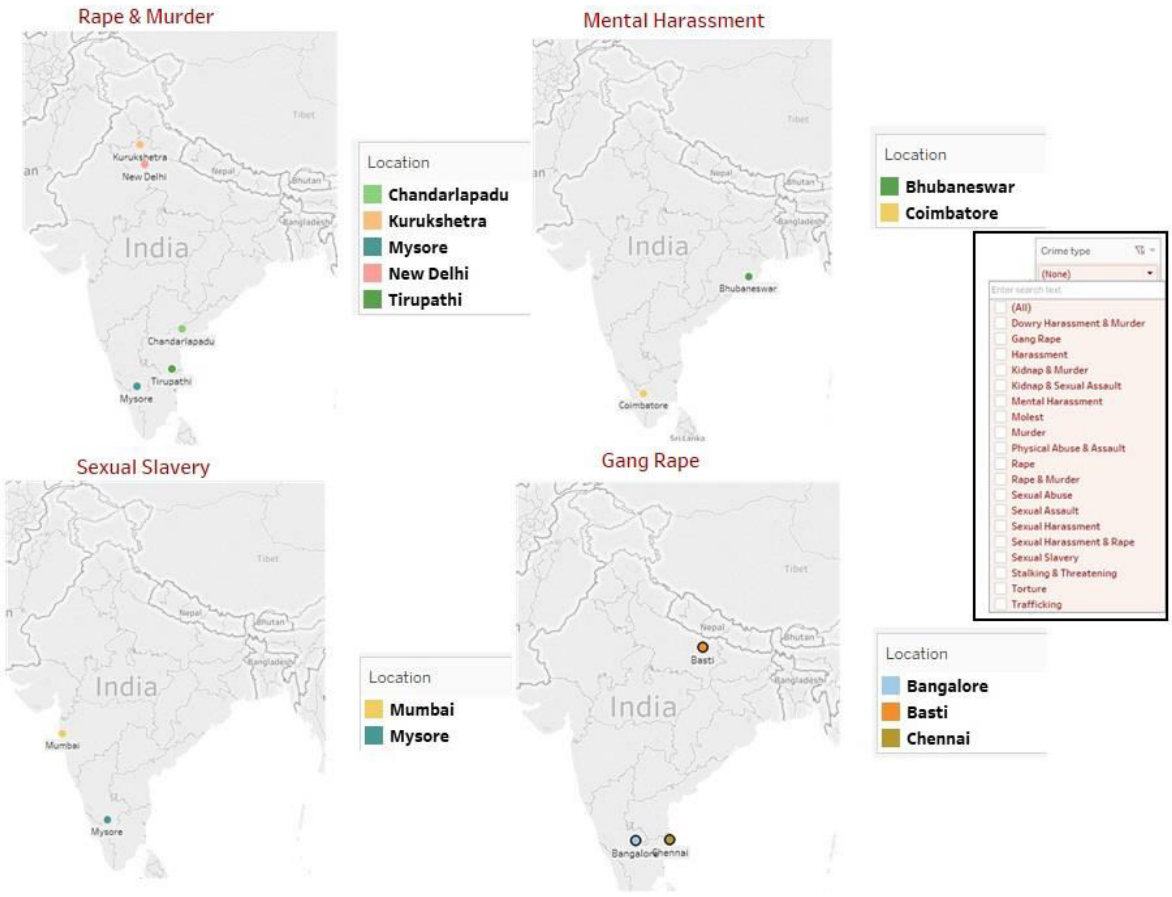

Figure 8. Location Prediction of Crime Hotspots

Figure 8 shows the predicted crime hotspots within India and also the possibility of occurrence of type of crime in that particular area.

4.2 Discussion

The proposed text-oriented decision support system, four-layer architecture effectively classifies the tweets into exact crime labels. From the tweets dataset extracted, it is crystal clear that most of the tweets are related to sexual harassment rather than any other crime labels. When considered the frequency of the term, the word "sexual" occurs more than 5000 times then comes the terms "harass" and "rape". From this, it is found that many of the public opinions are shared mainly on these two categories "sexual harassment" and "rape". From the bigram as well as trigram words, it is found that the terms "women sexual", "sexual harassment" occurs more than 2000 times. In finding the probability matrix of the crime label "suicide", it can be inferred that the occurrence of certain terms like "police", "case", "revenge" and "rape" confirms that the tweets belong to the category "suicide". From the proposed MNB-ALL algorithm for geo-location prediction, it

can be inferred that most of the crimes occur in areas in and around "North America" when compared to other parts of the world.

The results are validated with the help of the report provided by WHO "World Health Organization" under the head "Violence Against Women". Again in order to validate our results, we filtered the tweets and gathered only the tweets that are tweeted within India during the period January 2016 to December 2016. The possibility of occurrence of the type of crime and the location where it occurs are plotted in the map. The results are cross-verified with the report "Crime in India – 2016" published by NCRB (National Crime Records Bureau), India and found that nearly 82% of places were plotted correctly. The type of crime was exactly determined.

V.CONCLUSION AND FUTURE WORK

Now-a-days online crime reporting systems are being used by numerous law enforcement agencies which not only provides vast amount of information but also leads to an accumulation of ever-expanding digital crime reports. So the crime analysts as well as data journalist has to spend more time to analyze crime reports. Many existing research works do exist to analyze and predict the location of crimes, but with certain limitations. For instance, from the literature survey, we can infer that for extraction of tweets, the author relies upon any APIs and retrieved only limited number of tweets. In our work, we make use of Flume with R, to retrieve streaming tweets within few milliseconds.

Many existing research work makes use of machine learning algorithms like Naive Bayes, SVM, Decision trees and Random Forest. Our proposed approach makes use of Multi class, Multi-level NB classifier, which iterate through each and every level to match with the static and dynamic corpus provided, to tag the tweets. Because of tagging, we could enhance the accuracy of classification. Again in order to lessen the time taken by the classifier it is implemented under Map Reduce and Apache Spark's machine learning library entitled MLlib, to identify tweets discussing on different types of crime. The algorithms were fine-tuned with the depth of the hierarchical corpus and expert knowledge coded as rules to scale the probability scores. Through a large-scale implementation, it is found that our system is efficient, robust and scalable.

In the near future, we plan to extend and improve our framework by exploring more features that may be added in the feature vector and will increase the classification performance. Moreover, our work focuses on the use of textual features for geo-location prediction, but there are certain interesting future directions for predicting geo-location using non-textual features such as friendship links, temporal information as well as demographics information.

REFERENCES

1. Ahishakiye, E., Taremwa, D., Omulo, E. O., Nairobi-Kenya, G. P. O., Niyonzima, I. 2017. Crime Prediction Using Decision Tree (J48)

Classification Algorithm. Analysis, 6(03).

2. Ajinkya Ingle., Anjali Kante., ShriyaSamak., Anita Kumari., 2015. Sentiment Analysis of Twitter Data Using Hadoop. International Journal of Engineering Research and General Science , 3(6).
3. Aliguliyev, R. M., 2009. A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Expert Systems with Applications, 36(4), 7764–7772.
4. Arushi Jain., Vishal Bhatnagar., 2016.Crime Data Analysis Using Pig with Hadoop. International Conference on Information Security & Privacy , Procedia Computer Science 78 (2016) 571 – 578.
5. Bermejo, P., Gámez, J. A., Puerta, J.M., 2011. Improving the performance of naive Bayes multinomial in E-mail foldering by introducing distribution-based balance of datasets. Expert Systems with Applications, 38(3), 2072–2080.
6. Borg, A., Boldt, M., Lavesson, N., Melander, U.,Boeva, V., 2014. Detecting serial residential burglaries using clustering. Expert Systems with Applications, 41(11), 5252–5266.
7. Buczak, A. L., Gifford, C. M., 2010. Fuzzy association rule mining for community crime pattern discover. ACM, 1–10.
8. Chen, H., Atabakhsh et al., 2003. Visualization for crime analysis. Annual National Conference on Digital Government Research , 2003 Boston, MA. Digital Government Society of North America, 1-6.
9. Chen, J., Huang, H., Tian, S., Qu, Y., 2009. Feature selection for text classification with naïve Bayes. Expert Systems with Applications, 36(3), 5432–5435.
10. Cheng, Z.,Caverlee, J., Lee, K., 2010. You Are Where You Tweet: A Content-based Approach to Geo-locating Twitter Users. In Proceedings of the ACM International Conference on Information and Knowledge Management (CIKM),Washington, DC, USA, 25–28 July 2010, 759–768.
11. Chirag Kansara., Rakhi Gupta., S.D Joshi., Suhas Patil., Crime Mitigation At Twitter Using Big Data Analytics and Risk Modelling. IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE-2016), 2016, Jaipur, India.
12. Chowdhury, G. G., 2003. Natural language processing. Annual Review of Information Science and Technology, 37(1), 51–89.
13. Cocx, T., Kosters, W., 2006. A distance measure for determining similarity between criminal investigations. Advances in Data Mining , 511–525.
14. Dean, J.; Ghemawat., 2008.S. MapReduce: Simplified Data Processing on Large Clusters. Commun. ACM , 51, 107–113.
15. Diao, Y., Lu, H., Wu, D., 2000. A comparative study of classification based personal email filtering. In T. Terano, H. Liu, A. Chen (Eds.), PADKK '00 Proceedings of the 4th Pacific-Asia Conference on Knowledge

Discovery and Data Mining, Current Issues and New Applications, Kyoto, Japan : Springer, 408-419.

16. Duwairi, R., Qarqaz, I. 2014. Arabic Sentiment Analysis Using Supervised Classification. The 1st International Workshop on Social Network Analysis, Management and Security, Barcelona, Spain.
17. Emoticon. Available from: http://dictionary.reference.com/browse/emoticon [Accessed 2 March 2017].
18. Ghiassi, M., Olschimke, M., Moon, B., Arnaudo, P., 2012. Automated text classification using a dynamic artificial neural network model. Expert Systems with Applications, 39(12), 10967–10976.
19. Helbich, M., Hagenauer, J., Leitner, M., Edwards, R., 2013. Exploration of unstructured narrative crime reports: An unsupervised neural network and point pattern analysis approach. Cartography and Geographic Information Science, 40(4), 326–336.
20. Hitesh Kumar Reddy ToppiReddy., Bhavna Saini , Ginika Mahajan., 2018.Crime Prediction & Monitoring Framework Based on Spatial Analysis. International Conference on Computational Intelligence and Data Science, Procedia Computer Science 132 (2018) 696–705.
21. Iqbal, R., Murad, M. A. A., Mustapha, A., Panahy, P. H. S., Khanahmadliravi, N. 2013. An experimental study of classification algorithms for crime prediction. Indian Journal of Science and Technology, 6(3), 4219-4225.
22. Isa, D., Kallimani, V. P., Lee, L. H., 2009. Using the self organizing map for clustering of text documents. Expert Systems with Applications, 36(5), 9584–9591.
23. Karau, H., Konwinski, A., Wendell, P., Zaharia, M., 2015. Learning Spark: Lightning-Fast Big Data Analysis, O'Reilly Media, Sebastopol, CA, USA.
24. Kovachev, S., Reichert, P., Speck, H., 2008. Crime blips: Web based framework for crime incident analysis and visualization. 10th International Conference on Information Integration and Web-based Applications and Services . Linz,Austria: ACM,2008. 694-697.
25. Ku, C. H., Leroy, G., 2011. A crime reports analysis system to identify related crimes. Journal of the American Society for Information Science and Technology, 62(8), 1533–1547.
26. Ku, C.-H., Leroy, G., 2014. A decision support system: Automated crime report analysis and classification for e-government, Government Information Quarterly .
27. Ku, C. H., Iriberri, A., Leroy, G., 2008. Natural language processing and e-government: Crime information extraction from heterogeneous data sources. International Conference on Digital Government Research. Montreal, Canada: Digital Government Research Center.162-170.
28. Lee, K., Palsetia, D., Narayanan, R., Patwary, M., Agarwal, A.,

Choudhary, A. 2011 . Twitter Trending Topic Classification. 11th IEEE International Conference on Data Mining Workshops.

29. Li, W., Miao, D., Wang, W., 2011. Two-level hierarchical combination method for text classification. Expert Systems with Applications, 38(3), 2030–2039.
30. Li, L., Wang, J., Leung, H., 2009. A knowledge-based similarity classifier to stratify sample units to improve the estimation precision. International Journal of Remote Sensing, 30(5), 1207–1234.
31. Lin, J., Dyer, C., 2010. Data-Intensive Text Processing with MapReduce Morgan and Claypool Publishers: San Rafael, CA, USA.
32. Mahendran A., Duraiswamy A., Reddy A., Gonsalves C., 2013. International Journal of Scientific Engineering and Technology, 2(6),589-594.
33. Munesh Kataria1., Ms. Pooja Mittal., 2014. Big Data and Hadoop with Components like Flume, Pig, Hive and Jaql. International Journal of Computer Science and Mobile Computing, 3(7), 759 –765.
34. Nasridinov, A., Ihm, S. Y., Park, Y. H., 2013. A decision tree-based classification model for crime prediction. In Information Technology Convergence. Springer, Dordrecht, 531-538.
35. Nodarakis, N., Pitoura, E., Sioutas, S., Tsakalidis, A., Tsoumakos, D., Tzimas, G. kdANA., 2016. Rapid AkNN Classifier for Big Data. Trans. Large Scale Data Knowl. Cent. Syst. 23, 139–168.
36. Nodarakis, N., Sioutas, S., Tsakalidis, A., Tzimas, G., 2016. Large Scale Sentiment Analysis on Twitter with Spark. In Proceedings of the EDBT/ICDT Workshops, Bordeaux, France, 15–18 March 2016.
37. Pang, B., Lee, L., 2008. Opinion Mining and Sentiment Analysis. Found. Trends Inf. Retr. **2008**, 2, 1–135.
38. Pinheiro, V., Furtado, V., Pequeno, T., Nogueira, D., 2010. Natural language processing based on semantic inferentialism for extracting crime information from text. IEEE International Conference on Intelligence and Security Informatics (ISI) ,Vancouver, BC.19-24.
39. Piskorski, J., Atkinson, M., Belyaeva, J., Zavarella, V., Huttunen, S., Yangarber, R., 2010. Real-time text mining in multilingual news for the creation of a pre-frontier intelligence picture. ACM SIGKDD Workshop on Intelligence and Security Informatics, Washington, D.C,ACM, 1-9.
40. Rajan, K., Ramalingam, V., Ganesan, M., Palanivel, S., Palaniappan, B., 2009. Automatic classification of Tamil documents using vector space model and artificial neural network. Expert Systems with Applications, 36(8), 10914–10918.
41. Roth, R. E., Ross, K. S., Finch, B. G., Luo, W., MacEachren, A.M., 2013. Spatiotemporal crime analysis in U.S. law enforcement agencies: Current practices and unmet needs. Government Information Quarterly, 30(3),

226–240.

42. Schroeder, J., Xu, J., Chen, H., Chau, M. 2007. Automated criminal link analysis based on domain knowledge.Journal of the American Society for Information Science and Technology, 58(6), 842–855.
43. Twitter Sentiment Analysis Training Corpus. [Online]. Available from: http://thinknook.com/ Twitter-sentiment-analysis-training-corpus-dataset-2012-09-22/ (Accessed 2 March 2017).
44. Van Banerveld, M., Le-Khac, N., Kechadi, M.T., 2014. Performance Evaluation of a Natural Language Processing Approach Applied in White Collar Crime Investigation. In Proceedings of the Future Data and Security Engineering (FDSE), Ho Chi Minh City, Vietnam, 19–21 November 2014, 29–43.
45. Wan, C. H., Lee, L. H., Rajkumar, R., Isa, D., 2012. A Hybrid Text Classification Approach with Low Dependency on Parameter by Integrating K-Nearest Neighbour and Support Vector Machine. Expert Systems with Applications, 39(15), 11880–11888.
46. White, T., 2012. Hadoop: The Definitive Guide, 3rd ed.; O'Reilly Media/Yahoo Press: Sebastopol, CA, USA.
47. Yamamoto, Y., Kumamoto, T., Nadamoto, A., 2014. Role of Emoticons for Multidimensional Sentiment Analysis of Twitter. In Proceedings of the International Conference on Information Integration and Web-based Applications Services (iiWAS), Hanoi, Vietnam, 4–6 December 2014,107–115.