RESEARCH ARTICLE OPEN ACCESS

News document analysis by using a proficient algorithm

K.Meena[*], R.Lawrance[**]
**(Research Scholar, Research & Development Centre, Bharathiar University, Coimbatore, India)*
*** (Director, Department of Computer Applications, Ayya Nadar Janaki Ammal College, Sivakasi, India)*

ABSTRACT
News articles analyzing is one of the emerging research topic in the past few years. News paper discusses various types (political, education, employment, sports, agriculture, crime, medicine, business, etc) of news in different levels such as International, National, state and district level. In this news articles, crime discussion plays a major role because one crime leads to a many other crimes and also affect many other lives. In India, Madurai is one of the important places which have many historical monuments. Madurai is a sensitive place. This paper analyzes the crimes which occur in the year 2015 in and around Madurai. This analysis helps to police department to reduce the occurrence of crime in the future. This proposed system used Support Vector Machine (SVM) for effectively classify the document. News documents are preprocessed using pruning and stemming. From the stemmed words, the informative words are selected and weighted using feature selection methods such as Term-Frequency and Inverse Document Frequency (TF-IDF) and Chi-square. It returns the high dimensional vector space. It is reduced to low dimension using Latent Semantic Analysis (LSA) method. Compute the cosine similarity between the key document and news documents. Based on the value, the news documents are labeled as crime and non-crime. Some of the documents are used to train the SVM classifier. Some of the documents are used to test the performance of developed system. From the comparative study, it is identified that the performance of the proposed approach improves the classification accuracy.

Keywords: News document, Chi-square, Latent Semantic Analysis, Cosine-similarity, Support Vector Machine

I. INTRODUCTION

News articles are extremely occasion sensitive by nature. News articles are used to know the events which are happen in the state, center and international levels. It is used to express the thoughts of people, president of people, ministers, chief ministers and prime minister. It also makes awareness about the schemes of government, criticism of schemes and status of state, center, country and inter-national level. News articles consist of medical tips, business news and new methodologies of agriculture used to people. It also consist of sports news, cinema news which are attracted by the people those who are interested in sports and cini field. Even though nowadays various mediums are available to spread the news, news articles are usually read by the all people to know the incidents of previous day. Employment opportunity column in news paper is used for the people those who are search for the jobs. Jeweler and textile agencies used this media for advertise their products. Two wheelers and four wheelers Company also use this media to publish their new products. General knowledge, political news, new inventions, court judgments, crime details and police department actions are also available in the news paper.

The task of analyzing the news items is both interesting and challenging. Past research has dealt with forecasting the popularity of news items on web. Crime analysis is a key step in the sequence of activities aimed at conceiving, implementing, and evaluating measures to prevent crime. Crime Analysis is the qualitative and quantitative study of crime and police related information in combination with socio-demographic and spatial factors to apprehend criminals, prevent crime, reduce disorder and evaluate organizational procedures.

This proposed system analyzes the crime on Madurai region from the January 2015 to December 2015. Police departments face a number of crime problems because Madurai region is one of the sensitive areas in Tamilnadu. The distinctive crimes meet by the public, such as murder, theft, alcohol-related problems such as underage drinking and drunk driving, as well as assaults, sexual assaults, and rape. Such problems often consume many of their resources. This system can help to police departments in preventing such crimes in or around Madurai through the analysis of crime occur in previous year. The previous year analysis is used to determine why a problem was occurring, who was responsible, who was affected, where the problem was located, and what form the problem takes. The implementation of an adapted set of actions that address the most important findings of an analysis phase. Responses typically focus on at least two of the following: (1) preventing future occurrences by deflecting offenders; (2) protecting likely victims; or (3) making crime locations less conducive to problem behaviors. The focus of this system reflects the fact that this type of analysis is used for police departments and that there is an increasing demand for the application of crime analysis techniques that utilize data available from new papers.

The news articles for this study collected from English news papers in and around Madurai region. Multi dimensional features are selected from these articles. This paper uses TF-IDF, chi-square and Latent Semantic Analysis (LSA) to select the informative features. The primary objectives of the proposed system are 1) to select enlightening features related to crime 2) to train and test the SVM classifier model using the selected features with different kernel settings 3) to provide information to the police department about the analysis of crime in and around Madurai. This paper is arranged as follows. Section 2 provides details of related works. Section 3 illustrate about the crime analysis of proposed system. Section 4 discusses the results and discussions. Section 5 concludes the proposed work.

II. RELATED WORKS

Roja bandari et al.[1] analyze the news articles and predict the online popularity of that articles using support vector machine. Steven Pires and Ronald Belance[2] discusses the crime problems faced by the police departments in and around the college campus. Geographic Information System is used to identify the spots for criminal behavior has been mostly useful for police to reduce problem.

Masoumeh Zareapoor and Seeja K. R.[3] proposed various feature selection and feature extraction method used for email classification. It also shows that Latent Semantic Analysis(LSA) outperforms other methods used for feature extraction. Monica Rogati and Yiming Yang [4] performed an wide study of the performance of over 100 variants of 5 filter feature selection methods using two benchmark collections (Reuters 21578 and part of RCV1) and four classifiers (Naive Bayes, Rocchio, K-Nearest Neighbor and Support Vector Machines). The result of the study shows that the methods which include chi-square attain high performance in classification. Tehseen Zia, Qaiser Abbas and Muhammad Pervez Akhtar [5] conducted a study to analyze the performance of five feature selection method such as information gain, gain ratio, Chi statistics, symmetric uncertain and OneR using six classifiers (naive Bayes, KNN, support vector machine with linear, polynomial and radial basis kernels and decision tree) on two Urdu test collections: naive collection and EMILLE collection. It shows that Linear SVM with feature selection methods IG and Chi is outperformed other combinations of classifiers.

III. CRIME ANALYSIS

3.1 Preprocessing

Methods used for preprocessing is used to reduce the size of the file.

3.1.1 Pruning

This process removes the stop words used such as 'a', 'an', 'and' etc. A Newspaper article has various stop words. Pruning used to increase the predictive accuracy of the articles by removing stop words.

3.1.2 Stemming

This process produces the stem of the word. In this news articles, a word is occur in various formats such as word rob is appear as ‘robbery’ in one document, ‘robbed’ in another document. Stemming is used to avoid counting 'robbery' and ' robbed ' as two separate words. Stemdocument function in r is used to remove the suffixes attached to the words.

3.2 Term-document matrix

Construct term-document matrix after preprocessing is completed. This matrix is constructed with the term frequency only. In this matrix, row represents the words and column represents the articles. Each cell contains number which tells number of times that word appears in the particular article.

3.3 Feature selection

Feature selection methods used to remove the features that are irrelevant and also used to strengthen the weight of the relevant words. Methods used for feature selection use an estimation function applied to a single word [6]. Various feature selection methods such as term frequency, document frequency, information gain, inverse document frequency, probability ratio and gain ratio etc. are available [7]. These methods commonly used in information retrieval to find the individual words and their occurrence or co-occurrence.

3.3.1 Term Frequency

Number of occurrences of a term in a document is called term frequency. TF is calculated by add 1 with the value of taking log for the frequency of term t occurs in the document cd.

$$tf(t, cd) = 1 + \log f_{t,cd}$$

3.3.2 TF-IDF (Term Frequency – Inverse Document Frequency)

IDF is calculated by dividing the total number of answer by the number of document containing the term, and then taking the logarithm of that quotient. Here N represents the total number of documents, CD represents the documents and t is the term in the document.

$$idf = log \frac{N}{|\{cd \in CD: t \in cd\}|}$$

These two methods are used to improve the strength of words and it given as input to the feature transformation

3.3.3 Chi-square

Chi-square chooses words that have the more distribution across categories. If a term has a higher chi-squared score, that term is more useful [3]. This method used for measuring the independence between occurrences of two words. It is calculated by using
dchisq(list, df)
where list means weight vector values and df represents the number of values that are free to vary.

3.4 Feature transformation

Methods used for feature transformation is used to reduce the feature set size. There are various methods like Hyperspace Analog to Language (HAL), Principle Component Analysis (PCA) and Latent Semantic Analysis(LSA) for feature transformation.

3.4.1 Latent Semantic Analysis

Latent Semantic Analysis is a natural language processing technique used for analyzing the relationship between the set of documents and the terms. LSA literally means analyzing documents to find the underlying meaning or concepts of those documents [8]. It is a fully automatic mathematical technique for extracting and inferring relations of expected contextual usage of words in the passages of discourse which returns a matrix. In this matrix, rows represent the unique words and columns represent each paragraph. The mathematical technique of Single Value Decomposition (SVD) is used to reduce the matrix. Latent Semantic Analysis constructs a rectangular matrix of words by passages, with each cell containing a transform of the number of times that a given word appears in a given passage. The matrix is then decomposed in such a way that every passage is represented as a vector whose value is the sum of vectors standing for its component words. Similarities between words and words, passages and words are then computed using cosine similarity function. LSA works well on dataset with diverse topics. This method is reliable and faster than other dimensionality reduction methods. The weighted matrix(m) is converted into semantic space using the function lsa

$$mspace = lsa(m)$$

The format of semantic space is converted into an answer term matrix using the function

$$as.textmatrix(mspace)$$

3.4.2 Principal component analysis

Principal component analysis is a method used for extracting vital variable from a large set of variables from the data set. It is used to capture essential information and also used to reduce the size of the data set. PCA is performed on a symmetric correlation matrix. It should be numeric and have standardized data. In PCA, first principle components are identified. A principle component is a standardized linear combination of the original predictors in a data set. Executing PCA on un-normalized data will lead to large loadings for variables with high variance. Before finding the principle component, data should be normalized. For given m*n dimensional data, min (m-1, n) principle components can be constructed where m represents the number of documents n represents the number of terms. The correlation between these components should be zero. It is an unsupervised approach because the response variable is not used to determine the direction of these components. prcomp(matrix) used to compute the principle component analysis on the given input and return a list. In that list, the value of rotated data is used for calculation.

Using these procedures, construct the high dimensional vector space for the articles and the key factors. The key factor consists of words which are considered as crime. With the articles vector and the key factor vector, compute the cosine similarities. Resultant score is stored in a separate file. According to the score, label (Crime, Non crime) is given to the article. This resultant article name, score and label is given as input to the classification process.

3.5 Classification

Many text classifications have been proposed in data mining. Some of them are decision trees, naïve bayes, rule induction, neural network and Support Vector Machine (SVM). Naive bayes is often used in text classification because of its simplicity and effectiveness [9].

3.5.1 Support Vector Machine

Support Vector Machine is the supervised learning methods used for classification and regression tasks that are originated from the statistical learning theory. It is used in many fields such as bio informatics, text mining and image recognition. SVM is a machine learning technique used as to minimize the error based on the Structural Risk Minimization (SRM) principal [10].

SVM provides only global minima by using the concept of convex quadratic programming [11]. SVM was developed for binary classification. The fundamental feature of SVM is the operating maximum margin hyper plane whose position is determined by maximizing distance the distance from support vectors. The training point which is closer to the optimal hyper plane is called support vector. If the decision surface is obtained, it can be used for classifying new data. With the help of training data, SVM maps the input space into a high dimensional feature space. Consider a training data set of label pairs (p_i,q_i) where $p_i \varepsilon R$, $q_i \varepsilon \{1,0\}$ and i=1,2…. m. In the data set, p is an input vector space that contains the important words. The class label 1 denotes the crime document and the label 0 denotes the non-crime document. SVM classification finds an optimal separating hyper plane that can distinguish the two classes from the training data set. The SVM contains many arguments like formula, data type, kernel, scale degree, gamma and cost. The kernel attribute is used in training and predicting such as linear, sigmoid and radial. SVM is used in this proposed system due to its high performance, capacity to deal with the high dimensional data.

3.6 Flow graph and algorithm of proposed approach

The proposed document based SVM classification system takes the documents and generates a model that classifies the new documents into different predefined class. Preprocessing work is used to reduce the size of the file. Feature selection and feature transformation is important because it is used to improve the performance of the classification process. Figure 1 shows the flow process of the proposed approach. The documents are split into two group: Training documents, Test documents. Training documents are used to train the classifier. Test documents ate used to evaluate the action of the developed system.

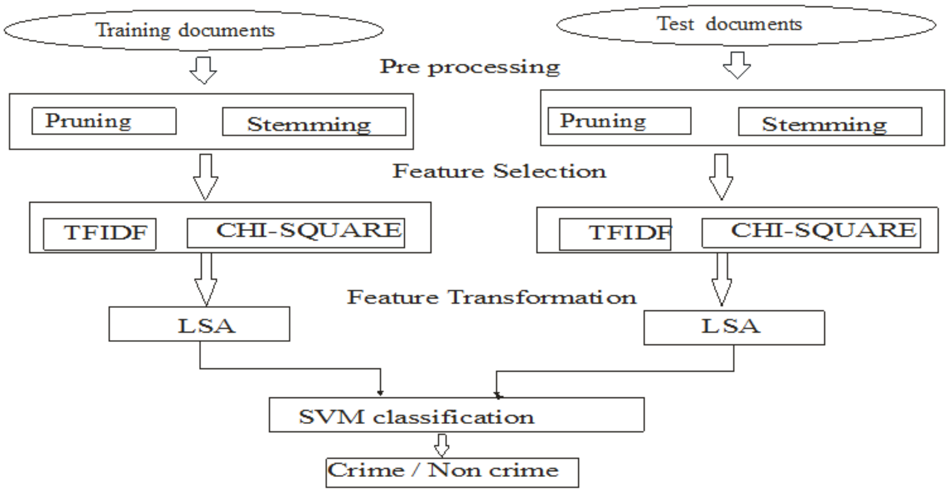

Fig. 1 Flow graph of proposed work

Algorithm:

```
Objective:  To classify the documents
Input      : Documents and Key document
Output: Documents are classified
Procedure:
    1.  Prune the documents
    2.  Stem the documents
    3.  Term-document matrix is constructed
    4.  TI-IDF and chi-square weighting scheme is used for feature selection
    5.  LSA is used for feature transformation
    6.  Compute cosine similarity is between the document vector returned from the LSA and key
        document vector
    7.  According to the cosine similarity values label is given to the documents
    8.  Some of the documents are considered for training and some of the documents used for testing
    9.  SVM classification is used to classify the documents as crime and non-crime.
```

Algorithm I CCLS

IV. RESULTS AND DISCUSSION

This section presents the result of the proposed approach using news articles in the year 2015. It is implemented in R software. Collected articles are stored as separate documents and keywords related to the crime are stored in a separate document named as key document. Documents are classified into two types: crime document and non-crime. The documents and key document are preprocessed using pruning and stemming process.

Feature selection is the important process used to select small subset of features that diminish redundancy and make the most of relevance to the target [12]. In this proposed system, some of the feature selection techniques such as term frequency, TF-IDF and chi-square are used to select the features that are highly discriminative. It leads to enhanced learning performance of classification, lower computational cost and leads to develop improved model. Table I shows the classification result after using feature selection method.

Table I Classification rate after using feature selection methods

Feature selection method	**% of crime documents correctly identified**	**% of crime documents misidentified**
TF	63	37
TF-IDF	74	23
Chi-square	72	24
TF-IDF + chi-square	77	23

The stemmed words are weighted using TF-IDF and chi-square methods. After feature selection, input space contains some set of features. If the number of data in feature selection is too large, input space will be transformed into a reduced set of features. So subset of features selected from feature selection method is given as input to feature transformation techniques such as PCA and LSA. It maps the original feature space into a new feature space with lower dimensions. In PCA, best eigenvectors are selected as new features and rests are discarded [3]. LSA is a technique mainly used for text classification. It examines an association between a term and concepts in a free format text. It also returns the features which are mostly related to terms and documents in a lower dimension. Table III shows the classification result after using feature selection and feature transformation methods.

Table III Classification rate after using feature selection and feature transformation methods

Feature selection + Feature transformation method	**% of crime documents correctly identified**	**% of crime documents misidentified**
TF-IDF + PCA	49	51
TF-IDF + LSA	88	12
TF-IDF + PCA+ LSA	58	42
TF-IDF + chi-square+ LSA	**96**	**4**
TF-IDF + chi-square+ PCA	64	46

Compute the cosine similarity between the vector for the document and vector for key document. Based on the value, the documents are labeled as crime and non-crime. These documents are given as input to the SVM classifier. Some of documents are used to train the classifier. The performance of the SVM classification is tested with test documents. The test set can be classified in one of four ways: True positives (the number of crime documents classified correctly), True negatives (the number of non-crime documents classified correctly), False positives (the number non-crime documents misclassified) and False negatives (the number of crime documents misclassified). Table III represents the results of classification using proposed algorithm CCLS with other methods. Figure 2 shows the classification accuracy of proposed method and various other methods.

This crime data set is also analyzed in different ways which are useful to the crime department. Figure 3 shows the number of crime occurred in the months of 2015 in and around Madurai. From that figure, crime department able to analyze the crime rate of months. June month has the highest crime rate. March month has the lowest crime rate.

Table III Comparison of proposed algorithm with other methods

Method	**Classification rate (%)**	**Error rate (%)**
TF-IDF + chi-square+ LSA +Naïve bayes	91	9
TF-IDF + chi-square+ LSA +SVM radial	87	13
TF-IDF + chi-square+ LSA +SVM sigmoid	82	18
CCLS	**100**	**0**

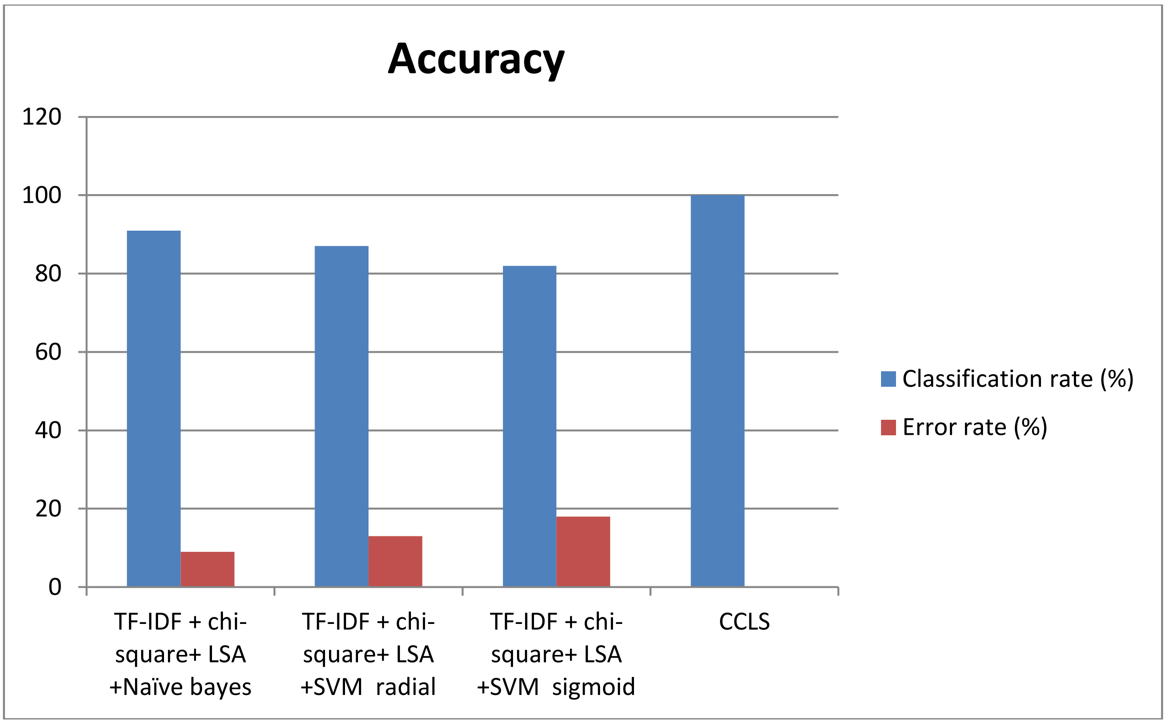

Fig. 2. Comparison of proposed with other methods

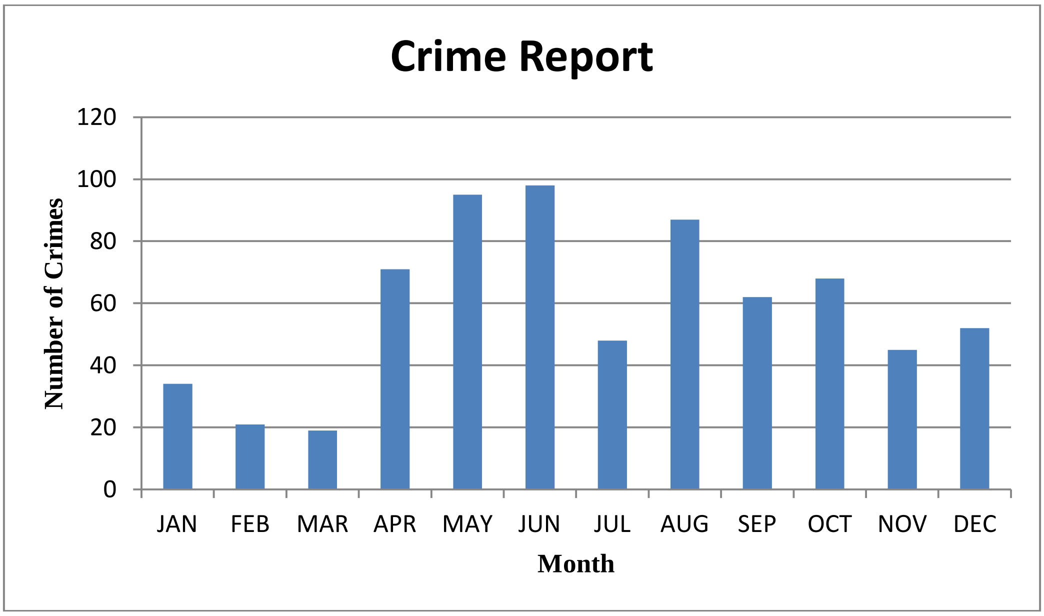

Fig. 3. Crime rate of months in the year 2015 in and around Madurai

V. CONCLUSION

News document has been classified by using SVM. Sizes of the documents are reduced using pruning and stemming. TF-IDF and Chi-square are used to select the didactic words from the document. The high dimensional vector space is reduced using LSA. These methods improve the classification accuracy of the SVM. The proposed algorithm produces better classification result when compared with other methods. This analysis is mostly helpful to the law enforcement department to categorize the crime and also used to reduce the crime rate of the month.

REFERENCES

[1]. Bandari, Roja, Sitaram Asur, and Bernardo A. Huberman. "The pulse of news in social media: Forecasting popularity." arXiv preprint arXiv:1202.0332 (2012).

[2]. Steven Pires and Ronald Belance. “Mapping Smoking Violations on a College Campus: Implications for Prevention”. Crime Mapping & Analysis News, Issue 4, 2015

[3]. Zareapoor, Masoumeh, and K. R. Seeja. "Feature Extraction or Feature Selection for Text Classification: A Case Study on Phishing Email Detection." International Journal of Information Engineering and Electronic Business 7.2 (2015): 60.

[4]. Rogati, Monica, and Yiming Yang. "High-performing feature selection for text classification." Proceedings of the eleventh international conference on Information and knowledge management. ACM, 2002.

[5]. Zia, Tehseen, Qaiser Abbas, and Muhammad Pervez Akhtar. "Evaluation of Feature Selection Approaches for Urdu Text Categorization." International Journal of Intelligent Systems and Applications 7.6 (2015): 33.

[6]. Soucy, Pascal, and Guy W. Mineau. "Feature selection strategies for text categorization." Conference of the Canadian Society for Computational Studies of Intelligence. Springer Berlin Heidelberg, 2003.

[7]. Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text classification using machine learning techniques." WSEAS transactions on computers 4.8 (2005): 966-974.

[8]. Landauer, Thomas K., Peter W. Foltz, and Darrell Laham. "An introduction to latent semantic analysis." Discourse processes 25.2-3 (1998): 259-284.

[9]. Kim, Sang-Bum, et al. "Effective methods for improving naive bayes text classifiers." Pacific Rim International Conference on Artificial Intelligence. Springer Berlin Heidelberg, 2002.

[10]. Vapnik, Vladimir. The nature of statistical learning theory. Springer science & business media, 2013.

[11]. Chen, Hui-Ling, et al. "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis." Expert Systems with Applications 38.7 (2011): 9014-9022.

[12]. Tang, Jiliang, Salem Alelyani, and Huan Liu. "Feature selection for classification: A review." Data Classification: Algorithms and Applications (2014): 37.