NEWS CLASSIFICATION AND CRIME INFORMATION EXTRACTION

A thesis submitted in the partial fulfillment of the requirement for the degree of

Master of Computer Science and Engineering

of

Jadavpur University

By

Sumanta Mukherjee

Registration Number: 140767 of 2017-18

Examination Roll Number: M4CSE19009

Under the guidance of

Dr. Kamal Sarkar

Department of Computer Science and Engineering

Jadavpur University, Kolkata-700032

India

2019

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JADAVPUR UNIVERSITY

TO WHOM IT MAY CONCERN

I hereby recommended that the thesis entitled "**News Classification and Crime Information Extraction**" prepared under my supervision by SUMANTA MUKHERJEE (Reg no. 140767 of 2018-19, Examination roll no. M4CSE19009 may be accepted in partial fulfillment for the degree of Master of Computer Science and Engineering in the Faculty of Engineering and Technology, Jadavpur University.

_____

Dr. Kamal Sarkar

Professor

Department of Computer Science and Engineering

Countersigned:

_____ _____

Dr. Mahan Tapas Kundu

Head
Department of Computer Science
and Engineering.
Jadavpur University,
Kolkata-700032

Prof. Chiranjib Bhattacharjee

Dean
Faculty of Engineering and
Technology
Jadavpur University,
Kolkata-700032

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JADAVPUR UNIVERSITY

DECLARATION OF ORIGINALITY AND COMPLIANCE OF ACADEMIC ETHICS

I hereby declare that this thesis contains a literature survey and original research work done by the undersigned candidate, as part of her MCSE studies.

All information in this document has been obtained and presented in accordance with academic rules and ethical conduct.

I also declare that, as required by these rules and conduct, I have fully cited and referenced all material results that are not original to this work.

Name: Sumanta Mukherjee

Exam Roll No.: M4CSE19009

Project Title: News Classification and Crime Information Extraction

---

Signature with date

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

JADAVPUR UNIVERSITY

CERTIFICATE OF APPROVAL

The foregoing thesis is hereby accepted as a credible study of an engineering subject carried out and presented in a manner satisfactory to warrant its acceptance as a prerequisite to the degree for which it has been submitted. It is understood that by this approval the undersigned do not necessarily endorse or approve any statement made, opinion expressed or conclusion drawn therein, but approve the thesis only for the purpose for which it is submitted.

FINAL EXAMINATION FOR 1. _____

EVALUATION OF THESIS: 2. _____

(Signature of Examiners)

ACKNOWLEDGEMENT

I express my honest and sincere thanks and humble gratitude to my respected teacher and guide **Dr. Kamal Sarkar**, for exclusive guidance and entire support in completing and producing this project successfully. I am very much indebted to him for the constant encouragement and continuous inspiration that he has given to me. And last but not the least, I also express my sincere gratitude towards Professor Mahan Tapas Kundu, Head of the Department, for providing me all the help needed by me. Finally, I convey my real sense of gratitude and thankfulness to all my friends and family members for their unconditional support without which I would hardly be capable of producing this huge work.

Sumanta Mukherjee

Class Roll No: 001710502029
Examination Roll No: M4CSE19009
Registration No: 140767 of 2017-18
Master of Computer Science and Engineering
Jadavpur University

List of Figures

List of Tables

Content

CHAPTER 1

Introduction

In any developing country crime rate analysis over different states is an important task and for different states, crime analysis for districts is also important. For having a rough view of crime rates newspapers are useful resources. As acquiring the crime reports from any local newspaper or a national newspaper is relatively easy and data extraction from them is also possible. But going through the huge collection of newspaper reports and getting the crime location and important data from them is not an easy task. This task can be roughly divided into two main parts first is spotting the crime news articles from the collection and second is extraction of data from them. The proposed system does this task automatically from newspapers of Bengali language. For getting the picture of crime scenarios in India based on different states and districts in India proposed system provides a map of crime in different states and cities of India. For getting the crime rate analysis, locations of crime, crime type, related personnel, etc. from newspaper reports, this system can be used as a tool. Such a tool can be helpful for many government agencies and personnel for different analyses of the crime rates in a very nominal time.

The proposed system is divided into six parts. 1st part is the Crime News Collection, 2nd Feature extraction 3rd Crime Information Tagging using CRF++ 4th Information Extraction 5th Rule-Based Ambiguity Removal and 6th Map Generation. The main two research areas of the system are news classification and information extraction. News classification has been very important research topic for many previous years. News classification systems have been built using Naïve Bayes, kNN, SVM and neural network-based architecture. In our system we have used a variant of Naïve Bayes algorithm for news classification. For data extraction we are tagging important word or word sequences from Bengali newspaper reports using

Condition Random Field. Here the tagging of the words is done with special crime tags described later in Table I. These crime tags have many similarities with named entity tags used for named entity recognition. But these tags concentrate more on crime specific information. At the end from the tagged data we are generating the crime map all over India. Also, for different states, district wise map is also generated by the proposed system.

1.1. Previous Work

The main concerned areas of our work are news classification, information extraction, and map generation. News classification with machine learning algorithms have been in the research domain for a long time and there have been different approaches to this idea like kNN, SVM, Naïve Bayes etc. KNN based news classification was one of them, proposed by Y. Zhou, Y. Li, and S. Xia (2009) [1] this classification algorithm chooses the class of the new or test data from measuring the distance from K nearest training samples. Naïve Bayes approach was also used by D. Lewis (1988) [2]. As the algorithm considers conditional independency of words, there were unsatisfactory results. Henceforth, some modification techniques were introduced by K. Schneider, (2005) [3]. Support vector machines were also used by Robert Cooley, (1999) [4] for this task. After that Chan, Chee-hong & Sun, Aixin & Lim, Ee-Peng. (2002) [5] also used SVM for the same but the system suffered from low recall score. Betterment techniques for SVM were introduced by J. Shanahan and N. Roma, (2003) [6] . For the information extraction, Conditional Random Field was first used for word tagging. This tool was used by many researchers for Parts of speech recognition, Named Entity Recognition like sequence tagging tasks. For this type of sequence tagging tasks HMM based systems are also very useful and were used by many. HMM or Hidden Markov Model is a type of probability based system which is proved to be a very useful model for sequence tagging task of languages with rich training data set, like English. But for Bengali like Indian languages where training data amount is not up to the mark HMM based systems were not very useful. Previous works on HMM based systems were done by Sarkar,(2013) [7] , Sandipan Dandapat, Sudeshna Sarkar,(2006) [8] for POS tagging showed the process for those systems. CRF based POS tagging systems were built by Asif Ekbal, Rejwanul Haque, S.bandyopadhay, (2007) [9]. POS tagging systems were also built using Support Vector Machines [10] and Entropy based methods were also built [11] . For social media languages POS tagging systems were built by Sarkar (2016) [12] using Conditional random Field. Other

than Part of speech tagging Name Entity Recognition task has a huge influence on our work. The implementation of our proposed tagging system is heavily followed by Named Entity Recognition works [13] [14] [15]. Combination based systems were developed by [14] and [15] where different machine learning algorithms like SVM, CRF and maximum entropy were combined. HMM based systems were developed by gayen and sarkar [13]. Though there are some influences of NER and POS tagging tasks on our work, proposed system differs from all of them. The proposed system is neither NER task nor POS tagging task completely. Our system is a crime tagging system. With some named entity tags there are some more crime specific tags (Table I) that are used in our system. For this feature, our system differs from all other systems mentioned above. In the system, many specific features were introduced for this kind of task.

1.2. Scope of the Work:

Our system introduces a new kind of tag set called as Crime tags. Previously in NLP there have been two different kinds of tags one is POS tags and another is Named Entity tags. These tags are useful for different types of information extraction tasks. Crime tags are another addition to the information extraction. For extracting important data from crime reports this tag set is very useful. This tag set can be used for tagging any kind of crime article. The tag set covers all related information for crime articles. In many types of text these tags can be used for getting important information.

In named entity recognition there has been tagging option for person name, locations. But this system more precisely assigns tags like official name, victim names, criminal names, crime locations etc. to news articles. Other than that they can find out other information like crime type, no of involved persons etc. This tagging can be used on different question answering platforms (like Quora, reddit etc.) to get useful information from crime-related posts and for building statistics from them.

1.3. Organization of Thesis:

The whole thesis comprises of 5 chapters. 1st chapter contains an introduction, 2nd chapter is description of news classification and sequence tagging task. 3rd is methodology with details

of every segment for the system. 4th chapter is results which contain all experimental results and evaluations in detail and 5th chapter is the conclusion which contains challenges and future scope.

CHAPTER 2

News Classification and Sequence Tagging

Our whole work mainly consists of two main research operations. One is news classification, and another is information extraction. Both topics have been in NLP research domain for a long time. News classification task is one of the most hugely used tasks for many researchers and organizations. News classification with machine learning tools has so many implementations. kNN, SVM, naïve Bayes and, deep learning tools have been used for this task. News can be classified in many topics like sport, business, political, crime and others. Classifying them to these classes depending on their content is the task. In case our work it is a binary class problem of classifying crime and non-crime articles. Information extraction from text is another most important topic for our work. Information extraction from text is also one of the most important tasks in NLP. It is the process of getting the important data and relations of the words from any text and to use them for a specific task. Word tagging or sequence tagging is the most important part of information extraction which is also a key component of the proposed system.

2.1. News Classification

For describing news classification, we are describing the necessary steps and then different implementation processes.

2.1.1. Steps of news classification:

News classification process can be roughly divided into three steps 1. Pre-processing 2. Feature extraction 3. News Classification.

1. Pre-processing: As news comes from different resources the data need to be cleaned before feature extraction and news classification. This pre- processing contains different special symbols removals and removal of words which appear customarily in text, known as stop words.
2. Feature Extraction: Text data cannot be supplied to a machine learning tool. Feature extraction from the data is very important for getting better accuracy. This type of feature extraction can be done in many ways like Boolean weighting based method, Information gain related method and frequency-based vectors. We have used frequency based vectors for feature extraction.
3. News classification: This stage is the final stage for news classification. Here feature extracted data is supplied to different machine learning tools for class prediction. Different algorithms for news classification are naïve Bayes, kNN, SVM, Neural Network etc. In case of our work we have used naïve Bayes method.

2.1.2. Algorithms for News Classification

1. Naïve Bayes: This is a probability-based algorithm which follows Bayes rule. This algorithm predicts the news class for an article depending on the given features for the article. Another variant of this algorithm for text is multinomial naïve Bayes. This algorithm does not consider the relative order of different words.
2. kNN or k Nearest Neighbour: This algorithm predicts the class for the articles depending on the distance measure of an article from the classes. This algorithm measures the distance of each article from its k neighbours and assigns the article to the neighbour group having the smallest distance.
3. SVM: Support Vector Machine is a mathematical tool which classifies the articles by creating the separating hyper planes. Support vector machines create hyper planes for every class. For differentiating the classes it maximizes the distance between the different hyper planes. As this is a mathematical model, this is the fastest model for classification of different articles.
4. Neural Network: Neural network is another algorithm for news classification this system works like human neuron. The vectors of the articles are given to the input nodes for classification. The output class is spotted on the output sections. Depending on the training data, the weights between different layers are adjusted using a training algorithm called back propagation. In this trained model the article vectors are supplied to get the output class.

2.2. Sequence Tagging

Information extraction is a very important task in NLP. Information extraction can be classified into two steps mainly one is word sequence tagging, second is getting the relations of tagged words for generating the important information. Our research work is more focused on the sequence tagging task. There have been two types of sequencing tagging tasks heavily used previously. One is POS tagging and another is Named Entity Recognition. Our work introduces crime tagging which is more useful for our crime information extraction task. All these three tagging systems have many procedural similarities with their uniqueness.

2.2.1. Stages of Sequence Tagging:

Sequence tagging task can be divided into three stages 1. Data Processing 2. Feature Extraction 3. Word tagging.

1. Data Processing: This is the first stage for any word tagging mechanism. In both the training and testing phase, this is important. This stage is for organising the data or text in a format, for being fed to a word tagging tool. This stage mainly organises the words, sentences, and files in a way so that they can be picked easily by the tagging tool for further processing. Like in CRF tool each of the word is kept in a newline and each sentence starts with one extra newline. In our process, we have used "END-OF-FILE" for spotting the end of a file which helps us in information extraction, map generation and evaluation process.
2. Feature Extraction: This is the most vital stage for any of sequence tagging task. The accuracy of any word tagging system depends upon the extracted features. This task is different for every tagging system. Specific feature extraction for every kind of tagging system is needed. In the case of named entity recognition length feature, context feature, vowel count plays a very important role and they enhance the accuracy very much. But for our crime tags, these features were not enough so, different special features like location feature, designation feature, colon feature were introduced to increase the efficiency of the word tagging system. These features play an important role in getting the crime location, related official, location stamp respectively.
3. Word Tagging: This is the final stage where the text after processing and feature extraction is given to a sequence tagging tool for final tagging. In the training phase this stage is for model building and in the testing this stage assigns the tag sequences

to the words. This stage can be implemented using different kinds of tools or algorithms like CRF, HMM, Entropy-Based system or a combined system. All of them accept the training file to build the model and assign tag sequences to words depending upon the trained model.

2.2.2. Sequence Tagging Tools:

Any kind of sequence tagging task depends on the chosen algorithm or tools. There are different kinds of sequence tagging tools to choose from. There are some probabilistic models like HMM(Hidden Markov Model), CRF(Conditional Random Field). Then there is entropy based model and combined models. Also, in deep learning RNN(Recursive Neural Network) is also another powerful tool for this task.

1. HMM: In simpler Markov models (like a Markov chain), the state is directly visible to the observer. The state transition probabilities are the only parameters for the system. But in case of hidden Markov model, the state is not directly visible, but the output (in the form of data or "token" in the following), dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. Therefore, the sequence of tokens generated by an HMM gives some information about the sequence of states; this is also known as pattern theory, a topic of grammar induction.
2. CRF: CRFs are a type of discriminative undirected probabilistic graphical model. CRF is very popular in NLP and computer vision research domain for its capability of labelling the sequence of words. In computation biology, it is also used for tagging DNA sequences and other sequence tagging tasks. Specifically, CRFs find applications in POS Tagging, shallow parsing, named entity recognition, gene finding and peptide critical functional region finding, among other tasks, proved as a very good alternative to the related hidden Markov models (HMMs). In computer vision, CRFs are often used for object recognition and image segmentation. We have used this tool and mathematical descriptions have been given later.
3. Entropy Base Model [11]: The Maximum Entropy model estimates the probabilities based on the imposed constraints. Such constraints are derived from the training data, maintaining some relationship between features and outcomes. The tagger uses a model of the form: $P(t|h) = \frac{1}{z(h)} \exp[\sum_{j=1}^{n} \lambda_j f_j(\mathrm{h}, \mathrm{t})]$ Where, t is tag, h is the context and the $f_j(h,t)$ are the features with associated weight λ_j. The problem of POS tagging can be formally stated as follows. Given a sequence of words $w_1 \ldots w_n$, we want to

find the corresponding sequence of tags $t_1 \ldots t_n$, drawn from a set of tags T, which satisfies:$P(t_1 \ldots t_n | w_1 \ldots w_n) = \prod_{i=1,n} P(t_i|h_i)$ Where, h_i is the context for word w_i. The Beam Search algorithm is used find the most probable sequence given the sentence. The features are binary valued functions which associate a tag with various elements of the context; for example: $f_j(h,t)$ is 1 if word(h)=Ami and t=PRP; else $f_j(h,t) = 0$.

4. Combined model: Combined model or system recombination is another way of sequence tagging. This method combines different types of machine learning algorithm to predict the output tag for the words than depending on a majority voting system it assigns the word with a tag. This type of system uses a base model like kNN or HMM or any other. For the words each tool predicts them. Then it outputs the tags with different sequence tagging systems and at last chooses the tag for a word which is given by maximum number of sequence tagging algorithm. One of them was used by K. Sarkar (2018) which uses 2 base models kNN and HMM. Then depending on the different algorithms it tags the words with named entity tags and chose the tag for a word depending on the voting mechanism.
5. RNN: Recurrent Neural Network is a very powerful tool coming from deep neural network domain. These neural networks are very useful for sequence tagging task. RNN have been used in many domains like speech recognition, text to speech synthesis etc. RNN is very useful in sequence tagging because of their power of inferring long relationships within a sentence. Basic RNNs are a network of neuron-like nodes organized into successive "layers", each node in a given layer is connected with a directed (one-way) connection to every other node in the next successive layer. Each node (neuron) has a time-varying real-valued activation. Each connection (synapse) has a modifiable real-valued weight. Nodes are either input nodes (receiving data from outside the network), output nodes (yielding results), or hidden nodes (that modify the data from input to output). LSTM or Long Short Term Memory is a type of RNN that has more significant applications in sequence tagging domain. LSTM uses memory to capture important previous word to enhance the output of the next words.

CHAPTER 3

Methodology

3.1. Overall Architecture

The whole system is divided into six parts we will describe each part separately. The six main parts are mainly 1. Crime data Collection. 2. Feature Extraction. 3. Word Tagging with crf++[1] 4. Information Extraction. 5. Rule Based Ambiguity Removal 6. Map generation. The basic architecture of the system is shown in Figure 1.

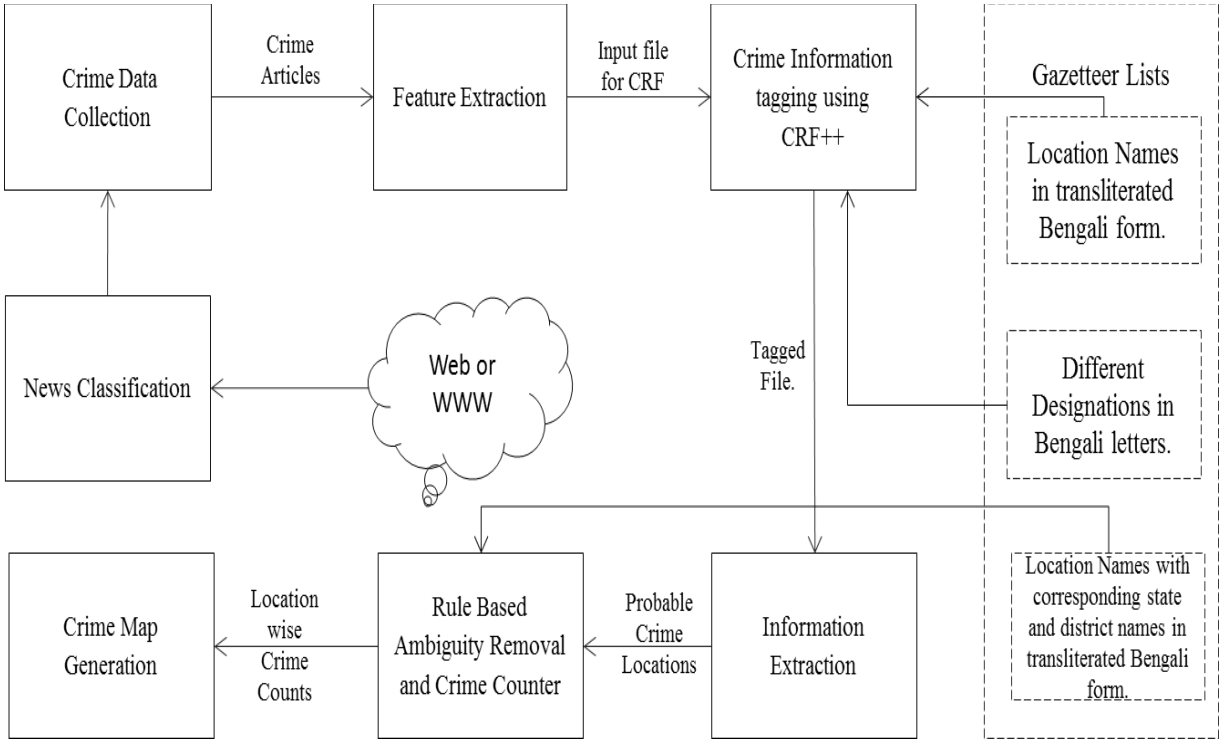


Figure 1. General Architecture of the System

CRF++:https://taku910.github.io/crfpp/

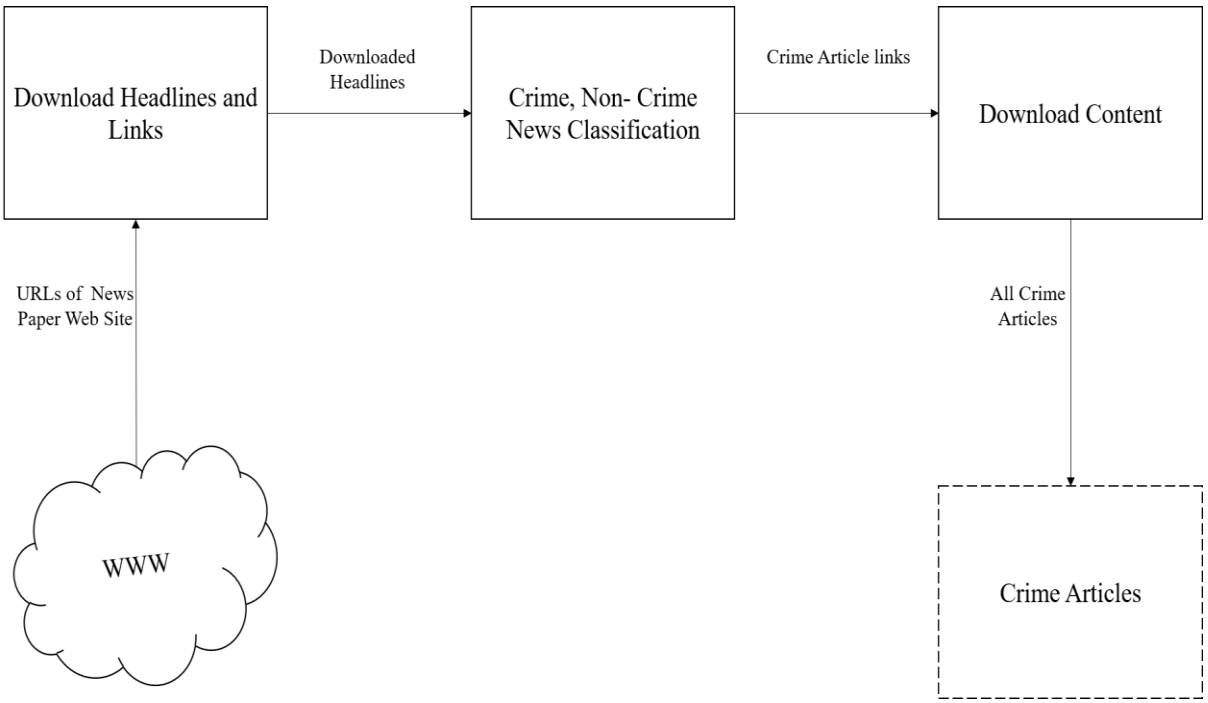


Figure 2. Crime Data Collection System

3.2 Crime Data Collection

This system is used for downloading the newspaper reports from the online websites. This section contains two types of web crawler, one is for headline and links collection and another is for content collection. As in our work non-crime articles are not used, we filtered them using news classification system and downloaded the contents of only crime articles. Therefore, links are downloaded from the web for all the articles but, the contents of all the articles were not downloaded at first. After the news classification of crime news and non-crime news the contents of crime articles are only downloaded. This process diagram is given in Figure 2.

3.2.1. Download Headlines and Links:

In the first segment of the system, headlines of different newspaper reports and links for the corresponding content from certain Bengali news websites (mainly www.anandabazar.com) are downloaded. Here scrappy python crawler with Xpath is used for crawling and parsing the web pages. The headings and links are stored to a text file. Multiple links or URLs were supplied for downloading different regional news headlines. Data for the training of news classification part was also gathered with this crawler but links for the contents were not

downloaded. Only headlines were necessary for this task. Some handpicked headlines were also added to the training file.

3.2.2. Crime and non-crime news classification:

This part is one of the most important segments for our system. There are several techniques for automatic news classification. Naive- Bayes, SVM are the most useful and easy classification methods for the binary type of classification. Here we have used a variation of Naïve-Bayes classifier called as Multinomial Naïve Bayes classifier. This type of naïve Bayes is most suitable for textual data where the frequencies of the words are very useful, not their positions. Chy, Abu & Seddiqui, Hanif & Das, Sowmitra. (2014) [16] has used this kind of system for their news classification task. This approach was followed with some necessary changes to the system.

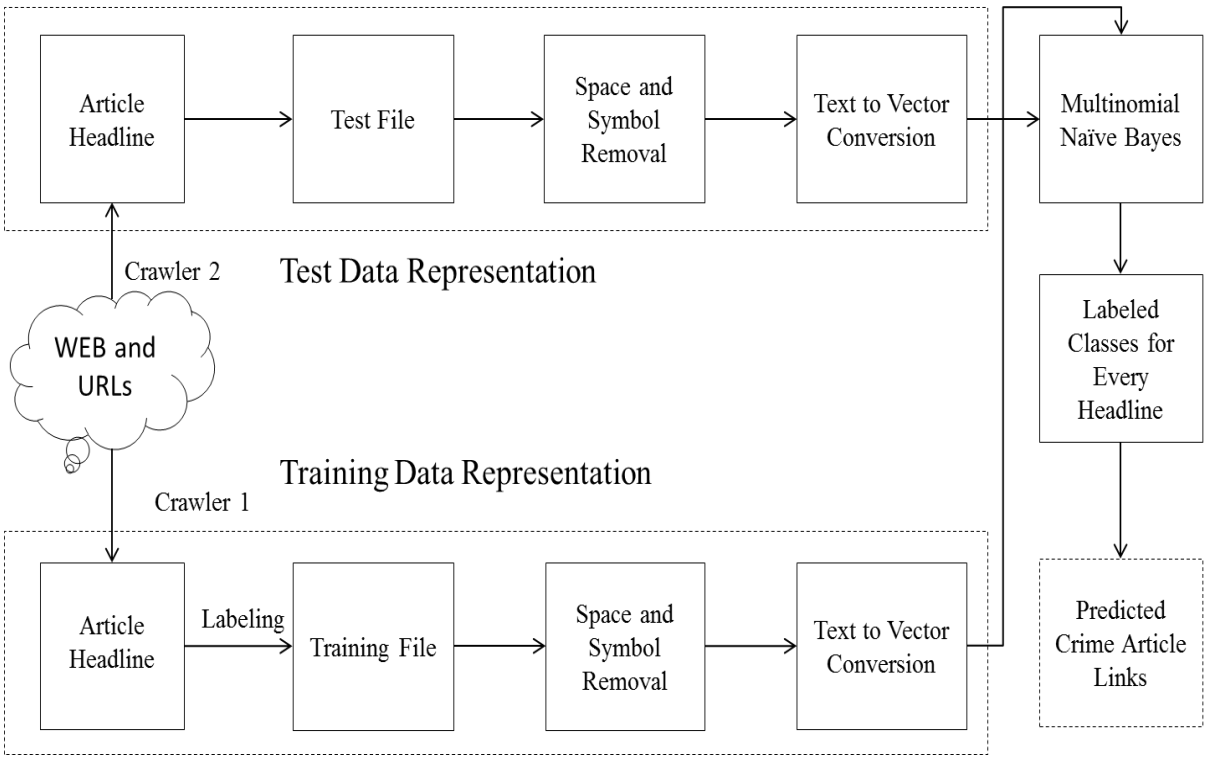


Figure 3. Crime, Non-Crime News Classification

Multinomial Naïve Bayes:

This is a probabilistic classification algorithm driven from the Bayes law with some constraints. Bayes law: P(C|A) = P(A|C)×P(C) P(A) In the context of text classification if A stands for the article and C stands for classes then the probability of the article belonging to a particular class can be given as the above. For multiple

classes and multiple features in an article, this probability calculation becomes too costly. Therefore, Naïve Bayes is used by assuming the relative condition Independence of the features. So if A has X= {x_1, x_2, x_3,..., x_n} features and C={c_1, c_2,...c_m} are all possible classes then by Naïve Bayes probability of a feature set belonging to a class is given by:

$$P(C|X) = argmax_{c \in C} \prod_{i=1}^{n} P(x_i|c) \times P(c)$$

Multinomial Naïve Bayes takes further constraints and it ignores the ordering of the words in an article, considering a bag of word approach it only takes the count of the words for a particular class. Multinomial Naïve Bayes assign the article to the class having maximum probability for the different features.

$$P(C|X) = argmax_{c_j \in C} P(c_j) \prod_{i=1}^{n} P(x_i|c_j)$$

In our system the problem is a two class problem (crime and not crime). The probability for the article headings are calculated for determining the class based on bag of word model. The article is assigned to the class having the highest probability.

News Classification Procedure:

Previously downloaded headlines are transferred to count vectors. These vectors are given to the Multinomial Naïve Bayes for classification where the vectors of pre-processed training files with correct classes are given. The process architecture is mentioned in Figure 3. For producing the training file the crawler was used for downloading random headlines with some handpicked headings. Then headings were labelled with the output class manually.

3.2.3. Downloading Content:

After the classification of crime and non-crime articles with the news classifier, the links for crime articles are supplied to this section. Henceforth, in this section, a crawler only

downloads the contents of the crime articles by parsing the web pages using the scrappy and XPath. All the contents are kept in a text file with "END-OF-FILE" mark at the end of each content.

3.3. Feature extraction

In the last section the contents were downloaded. But for supplying them to CRF they need to be processed further which happens in this stage. In order to get the better result, we extracted different features from the data for this specific task. Total 12 different features were used for identifying different type of information and to assign different tags easily.

3.3.1. Input data structure for CRF++:

After the feature extraction, the data with features are fed to the CRF++ tool. Before feeding the data to the CRF++ tool the data is pre-processed. CRF++ takes textual data in a column format where every column specifies a feature. After downloading the crime data from the web, it first executes one program to organize every word in a new line with another column containing a default tag 'O'. Each of the document was separated by a "END-OF-FILE" word tagged as 'O'. For training the CRF++ tool the words were tagged manually after that. The tagged data were then passed to the feature extraction code for generating a formatted data with each feature as a column.

3.3.2. Format of tagging:

For sequence tagging in the data, we have used I-O-B format tagging. Here the beginning of the sequence is tagged with "tag name"_B and intermediate words are tagged with "tag name_I". Other non-important data or words were tagged with O. For the task our tag set has a total 43 tags. All possible tags are mentioned in Table I.

TABLE I DIFFERENT CRIME TAGS

ATTRIBUTE	**VALUES**	**TAG**
Location		LOC
Crime Type	A. Personal Crime 1. Assault 2. Battery 3. Kidnapping	 CrmPerAst CrmPerBat CrmPerKdn

	4. Homicide 5. Suicide 6. Sexual Assault	CrmPerHom CrmPerSui CrmPerSex
	B. Property Crime 1. Larceny 2. Burglary 3. Robbery 4. Arson 5. Embezzlement 6. Forgery 7. False pretences	 CrmProLrc CrmProBrg CrmProRob CrmProArs CrmProEmb CrmProFrg CrmProFprt
	C. Inchoate Crime 1. Attempt 2. Solicitation 3. Conspiracy	 CrmIncAtm CrmIncSol CrmIncCon
	D. Statutory 1. Drunk driving 2. Selling alcohol to a minor	 CrmStaDrn CrmStaSel
Casualty	A. Wounded	CasW
	B. Death	CasD
	C. Kidnapped	CasK
Weapon	A. Handgun	WepHndGun
	B. Rifles	WepRif
	C. Shotgun	WepShtGun
	D. Fire Arms (type unknown)	WepFarm
	E. Knives or cutting instruments	WepKnCut
	F. Hands, fists, feet etc.	WepBdy
	G. Explosives	WepExp
	H. Other weapons	WepOth
PERSON	A. Victim	PerVic
	B. Criminal	PerCrm
	C. Officials	PerOff
	D. Suspects	PerSus
	E. Gang	PerGng
EVENT TIME	A. Date	EvnDt
	B. Day	EvnDy
	C. Time	EvnTm
CARDINALITY	A. Death	CardD
	B. Wounded	CardW
	C. Kidnapped	CardK
	D. Criminals	CardC

	E. Arrested	CardA

3.3.3. Feature Details:

a) Context feature: We used the context of the words as a feature. For the current word previous word and the next word is used as a feature. Also, we have used a conjugate feature of < previous word, current word, next word>.

Ex:

ফের O
আত্মহত্যার CrmPerSui_B
ঘটনা O
রাজস্থানের LOC_B
কোটায় LOC_I
। O

So, in the piece of text for the word 'আত্মহত্যার' previous word 'ফের' and next word 'ঘটনা' is taken as a feature.

b) Number feature: This feature is used to recognize numbers. This is a Boolean type feature which is set to 1 if there is a number in a word or the word is a number.

Ex:

এবার	0	O	
আত্মঘাতী		0	CrmPerSui_B
১৭	1	O	
বছরের	0	O	
এক	0	O	
ছাত্র	0	O	
।	0	O	

c) Colon Identifier: This feature is also a Boolean type of variable it is one if the word is a colon symbol. Colon symbol was differently tagged than other special symbols as they play an important role identifying the time and location stamps used in newspaper articles.

Ex:

NONE	0	O
:	1	O

২৬	0	EvnDt_B
ডিসেম্বর	0	EvnDt_I
,	0	EvnDt_I
২০১৮	0	EvnDt_I
, 0	O	
১৮	0	O
:	1	O
৩০ 0	O	
: 1	O	
০৯ 0	O	
:	1	O

d) Length feature: This is a string feature. Here it can have four values {L1,L2,L3,L4} where if the length of the current token or word is one then it is set to value L1, if the length of word is two then it is set to L2 and respectively as L3 and L4 for word length of 3 and 4. The words with length more than four are also set with value L4.

Ex:

ফের	L3	O
আত্মহত্যার	L4	CrmPerSui_B
ঘটনা	L4	O
রাজস্থানের	L4	LOC_B
কোটায়	L4	LOC_I
।	L1	O

e) Designation feature: This feature is also a string feature. This is set to p_off if the current word is any designation and it set to p_off for the next words till the word is not any special symbol or any end line or end of file. For identifying designation a file is used which contains different designations. If the word is not any designation then the value is set to “none”.

Ex:

জানিয়েছেন	none	O
সেক্টর	none	O
৪ ৯ -এর	none	O
স্টেশন	none	O
হাউস	none	O
অফিসার	p_off	O
অজয়	p_off	PerOff_B

আগরওয়াল p_off PerOff_I
। p_off O

In this example, the feature is turned on to p_off after encountering a designation officer and it is on till finding a special symbol which in this case is an end of a sentence.

f) Vowel count: This is a counting feature which is the number of vowel count in the word. As the words are in Bengali the vowel or "swaroborno" s in the words are counted.

Ex:

কোটায় 2 LOC_B
এই 2 O

g) Prefix feature: This feature is a string feature and divided in four parts. If the word length, Wlength >=2 than it is P1 = prefix after deleting the last character else P1 = word. If Wlength >= 3, P2 = prefix after deleting the last two characters else P2 = word. If Wlength >= 4, P3 = prefix after deleting the last three characters else P3 = word. If Wlength >= 5, P4 = prefix after deleting the last four characters else P4 = word.

Ex:

ফের ফে ফ ফের ফের O
আত্মহত্যার আত্মহত্যা আত্মহত্য আত্মহত্ আত্মহত CrmPerSui_B
ঘটনা ঘটন ঘট ঘ ঘটনা O
রাজস্থানের রাজস্থানে রাজস্থানরাজস্থা রাজস্থ LOC_B
কোটায় কোটা কোট কো ক LOC_I

h) Suffix feature: Suffix feature acts like the prefix feature but only the suffixes of the words are used. If the length of the word, Wlength >= 2, value is set to S1 = suffix containing the last character else S1 = word. If Wlength >= 3, it is set to S2 = suffix containing the last two characters else S2 = word. If Wlength >= 4, S3 is set to suffix containing the last three characters else S3 = word. If Wlength >= 5, S4 = suffix containing the last four characters else S4 = word.

Ex:

ফের র ের ফের ফের O
আত্মহত্যার র ার যার ্যার CrmPerSui_B
ঘটনা া না টনা ঘটনা O
রাজস্থানের র ের নের ানের LOC_B

কোটায় য় ায় টায় োটায় LOC_I

i) Location feature: This feature is also a string feature and it has three possible values. This is a gazetteer feature. The default value used is “no_name” i.e. if the word is not a complete or partial match to any location name mentioned in a separate location name text file[2].

The text file for location names used here is transliterated to Bengali with google translate and manual intervention after downloading. This value is set to l_name if the word is a complete match with any name in location name file. This value is set to l_p_name if it is a partial match with any location name mentioned in location name file. Here partial match means if the (length of current word –length of any word in location name file is <=2 and the non-matched words belong to a set of common prefixes used with the location names or noun forms in Bengali language. In location name file with the location names corresponding state and district names were also transliterated to Bengali for further use.

Ex:

ঘটনা no_name O
রাজস্থানের l_p_name LOC_B
কোটায় l_p_name LOC_I

In this example l_p_name is added for word ‘রাজস্থানের’ and ‘কোটায়’ as these are not perfect match to the location name ‘রাজস্থান’ and ‘কোটা’ respectively from gazetteer list of location names. But if last two characters are removed then those are location names therefor l_p_name is added in this tag.

জানতে no_name O
পেরেছে no_name O
রাজস্থান l_name LOC_B
পুলিশ no_name O

But in this case ‘রাজস্থান’ is perfect match with a location name from the gezetter list for location names. Therefore l_name is tagged for the word.

j) Crime location and time identifier:

This is a string feature and it has four set of values and the values are {day, time, prob_loc_set, loc_set}. This feature based on some very common practices of Bengali newspaper articles. In newspapers “ghotona”,”ghoteche” are some words which are written in most of the contexts of crime locations. We have used this common practice for this feature. In the sentences where both of this words or there other forms like ‘ghotonati’, ‘ghotonay’ ‘ghote’ are present this feature is used. Words after this both words are tagged with “loc_set”. If this words or any of the words are present then previous words before them in the sentence are tagged with prob_loc_set. In case where this words are at last and where if any of words are name of days then they are tagged ”day” and word presenting any part of the day like morning, noon, afternoon etc. are tagged with “time”. This is a very useful feature for this kind of system where finding out the crime location from so many locations in a file is important.

Ex:

বুধবার	day	EvnDy_B	
রাতে	time	EvnDy_I	
ডানকুনিতে	prob_loc_set	LOC_B	
দিল্লি	prob_loc_set	LOC_I	
রোডের	prob_loc_set	LOC_I	
ধারে	reset	O	
একটি	reset	O	
লজের	reset	O	
ঘটনা	reset	O	
।	reset	O	

In the example ‘ঘটনা’ is present in the sentence at last. Before the word when some days of week or time occur it is tagged with day and time which in this case happens for ‘বুধবার’ and ‘রাতে’. Then the next section before ‘ঘটনা’ is tagged prob_loc_set and it keep tagging before some words like road, village, city, block, apartment comes.

ঘটনাটি	reset	O	
ঘটে	reset	0	O
উত্তরপ্রদেশের	loc_set	LOC_B	
মুজফ্ফরনগরের	loc_set	LOC_I	

কাকরোলি loc_set LOC_I
এলাকায় loc_set LOC_I
। reset O

In this case ‘ghotonati’ and ‘ghote’ are present so the next words are tagged ‘loc_set’.

k) Special symbol feature:

This is a Boolean feature. This is set to one if the current word is a special symbol else, it is set to zero.

Ex:

থাকতো ০ O
, 1 O
সেখানেই ০ O
ঝুলন্ত ০ O
অবস্থায় 0 O
পাওয়া 0 O
গিয়েছে 0 O
তার 0 O
মৃতদেহ 0 CasD_B
। 1 O

So it is visible from the above example that this feature is one when word is a special symbol, else 0.

l) Dynamic Feature:

The predicted tag of previous word is also given as a feature for current word.

Ex:

ফের O
আত্মহত্যার CrmPerSui_B
ঘটনা O
রাজস্থানের LOC_B
কোটায় LOC_I
। O

In this case if LOC_B is a predicted tag for word “রাজস্থানের” then for predicting LOC_I in next word, LOC_B is used as a feature.

3.4. Crime Information Tagging

For getting the useful information from newspaper data CRF is used for tagging. Important words or word sequences are tagged in this section. Sequence tagging is done by the CRF++ tool where the different crime tags were used to capture the information. The set of tags are mentioned in Table I.

3.4.1. Conditional Random Field:

Conditional random Field is a probabilistic model for sequence tagging or labelling sequence data proposed by Lafferty ET. Al., (2001) [17] . Conditional Random field works with the help of feature functions. Designing the feature function depending on the work is very important. Different feature functions are built and weights are given to different feature functions. The mathematical representation of the CRF is shown below. Each feature function f_j is assigned a weight λ_j . Given a sentence s, we can now score a labelling l of s by adding up the weighted features over all words in the sentence:

$$score(l|s) = \sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i, l_{i-1})$$

(The first sum runs over each feature function f_j, and the inner sum runs over each position i of the sentence.) Finally, we can transform these scores into probabilities P (l|s) between 0 and 1 by exponentiation and normalizing or soft max normalization:

$$P(l|s) = \frac{\exp\big(score(l|s)\big)}{\sum_{l'} \exp\big(score(l'|s)\big)}$$

$$P(l|s) = \frac{exp[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l_i l_{i-1})]}{\sum_{l'} \exp\big[\sum_{j=1}^{m} \sum_{i=1}^{n} \lambda_j f_j(s, i, l'_i l'_{i-1})\big]}$$

3.5. Information Extraction

The tagged words of CRF are extracted from the text for each article. Currently, for the location mapping all the location tagged words were fetched. In news articles there are some small locations like city names, village names, street names etc. are tagged. But for the final

map generation, the corresponding state name and district names are selected in this part. A hierarchically structured data[2] for different locations, districts and states are given to the system. This data was transliterated to Bengali after downloading with the help of “Google translate” and manual intervention. From the tagged words this segment finds all possible district and state names.

3.6. Rule Based Ambiguity Removal

In CRF part there are still some ambiguities for choosing the crime location. Multiple locations were tagged for single file. This part refines them and finds the crime location. Based on some rule based operations the final locations of crimes are selected. From all the locations that are tagged for a file district names and state names are fetched using the matching algorithm than for all these names a priority list is made. The locations are prioritised depending on their context. Some words like "ghotona" and "ghoteche" are given the highest precedence. These words occur in context with crime location for most of the reports. Then the location where words like police station (’thana’) or court (’adalat’) are present are given higher precedence then the locations where the location signifies someone’s residence. This priority based mechanism increases the chance of spotting the crime location. But, for many locations that occur in multiple places over India are not eradicated by this method. For locations where multiple states and districts are present for one location. Tagged word sequences and location stamp of the report plays an important role. When location sequences are tagged, the ordering of the locations is used to find the correct state or district. The location stamp is also used to remove duplicate locations. In cases where the state of tagged locations are different from the state of location stamp of the reports is removed when multiple locations for the same word come up.

3.6.1. Matching Algorithm:

In matching algorithm, the location list with district names and state names are given with the word sequence tag loc.

1. First, it matches individual words with location names in location list.
    1.1. First, the whole word is searched if it encounters match then it stores the list of state and district in a list and move to next word in sequence. If the word is a state name then it returns state name only.

1.2. If no match is found with the whole word then the last character is removed from the word. If the removed character belongs to common prefixes used for nouns in the Bengali language it checks the new word with all location names and returns the district and state names for matched cases and move to next word. If the word is a state name then it returns state name only.

1.3. If still no location is found then it removes last two characters form the word. If the removed character belongs to common prefixes used for nouns in Bengali language it checks the new word with all location names and returns the district and state names for matched cases and move to the next word. If the word is a state name then it returns state name only.

2. If step 1 return null list of states and districts then from sequence consecutive two words are concatenated and it is set as the new word. Then step1 occurs again. This happens for every two consecutive word.
3. If step 2 return null list of states and districts then from sequence consecutive three words are concatenated are taken and it is set as the new word. Then step1 occurs again. This happens for every three consecutive words in the sequence.

3.7. Map Generation

From the final locations, the crime rates of a district and states are counted. Then counts are plotted as map for the data. At the last section the visual output of the map is generated for both states and districts by using GUI based system.

CHAPTER 4

Results

In our system, the performance depends on news classification, crime information tagging and, map generation parts. Each of the parts has a different evaluation process and metrics. For better evaluation of the system, we used different testing methods. The news classification part uses precision, recall and f-measure for the evaluation. The results of these measures for news classification part are given in Table II. The crime information tagging section was evaluated using 5-fold cross validation and the results are mentioned in Table III, IV & VI. As the final result is produced after the map generation, for every document in the test file we have manually noted the resulting states and districts of the crime and then compared it with the results. The confusion matrix for this task is given for state level in Table V.

4.1. Experimental Data:

In this whole process, we have used the web news downloaded from anandabazar patrika for both training and testing section. In the training section of news classification we have used 14244 news headlines which were tagged manually. 18250 news headings were used for testing the same system. In word tagging system near about 200 files were tagged manually word by word. The training file consists of 4000 lines .For testing 1550 documents were used (This are crime articles classified by text classification system). In testing file, 42500 sentences were there.

4.2. News classification:

For calculating the accuracy of news classification we are using different metrics. These metrics are described in Table II.

Evaluation metric	Score
Precision	0.858
Recall	0.866
F score	0.856
Accuracy	0.866

TABLE II MULTINOMIAL NAÏVE BAYES CLASSIFIER ACCURACY MEASURES

4.3. Crime information tagging with CRF++:

For checking the accuracy of the tagging by CRF we used 5 fold cross validation method with the training data. The Scores are mentioned in Table III & IV, where Table III is the table for all tags accuracy including 'O' tags and Table IV is for only crime tags excluding 'O' tags.

FOLD NO.	TRAIN FOLD	TEST FOLD	ACCURACY
1	1,2,3,4	5	90.44
2	1,2,3,5	4	90.39
3	1,2,4,5	3	91.26
4	1,3,4,5	2	89.91
5	2,3,4,5	1	91.47
Average Accuracy			90.69

TABLE III ACCURACY MEASURE FOR CRIME INFORMATION TAGGING (INCLUDING 'O' TAGS)

FOLD NO.	TRAIN FOLD	TEST FOLD	ACCURACY
1	1,2,3,4	5	42
2	1,2,3,5	4	46
3	1,2,4,5	3	43
4	1,3,4,5	2	42
5	2,3,4,5	1	41
Average Accuracy			42.8

TABLE IV ACCURACY MEASURE FOR CRIME INFORMATION TAGGING (EXCLUDING 'O' TAGS)

4.4. Final Data Plotting:

Though, we have designed a rich set of crime tags and used CRF for tagging with all tags, for present map generation task we have only used location information because our primary

objective is to generate location wise crime map. For location tagging our CRF's performance is given in Table VI.

This system gave 82% accuracy. The evaluation process has neglected out of India data. The measure for this section is given by the following confusion matrix on state level. The matrix is given in Table V. The confusion matrix gives three columns for each of the states. True Positive, False Positive and False Negative. True negative is omitted in this case as it is not useful in this context. Though the location tag accuracy is 50 % this section accuracy comes to 82 % because for every file there are many words tagged are loc but spotting the location one time correctly in file increases the accuracy for final plotting.

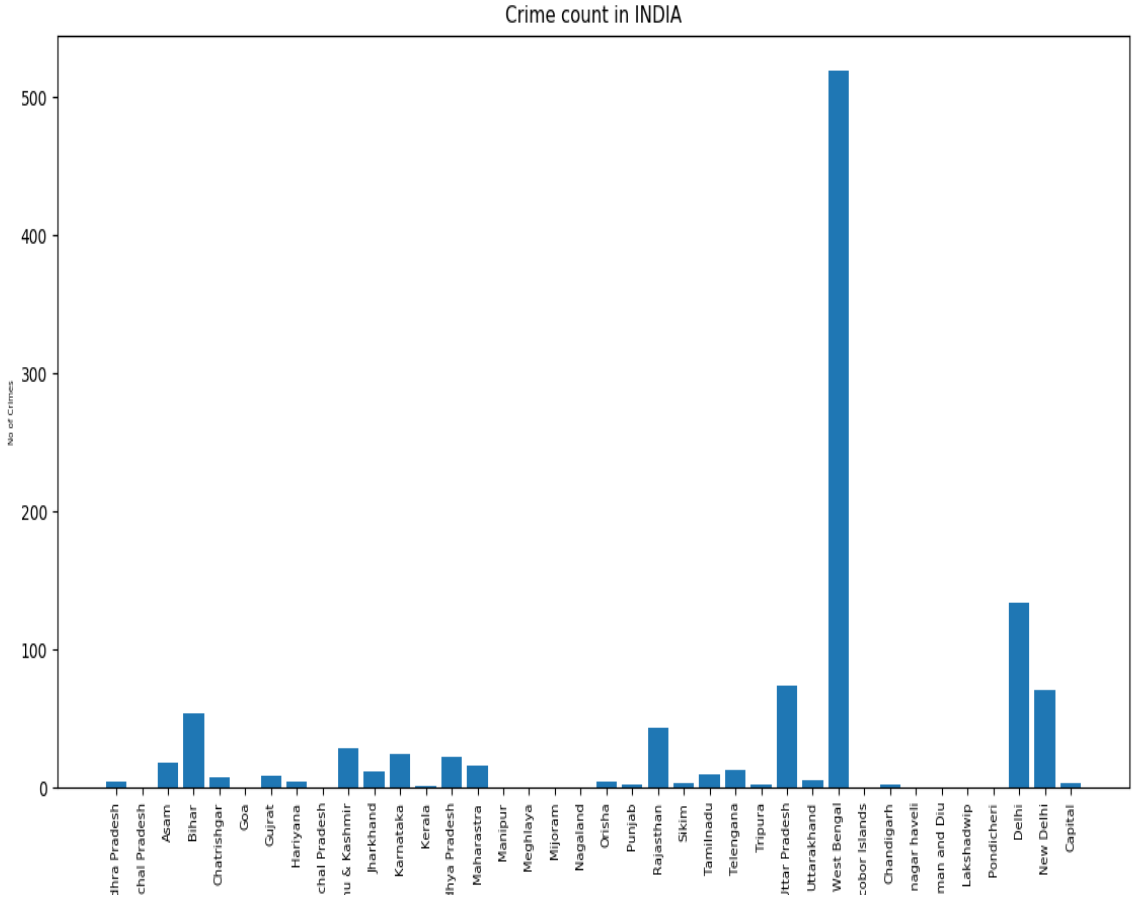


Figure 4. State wise Crime Map

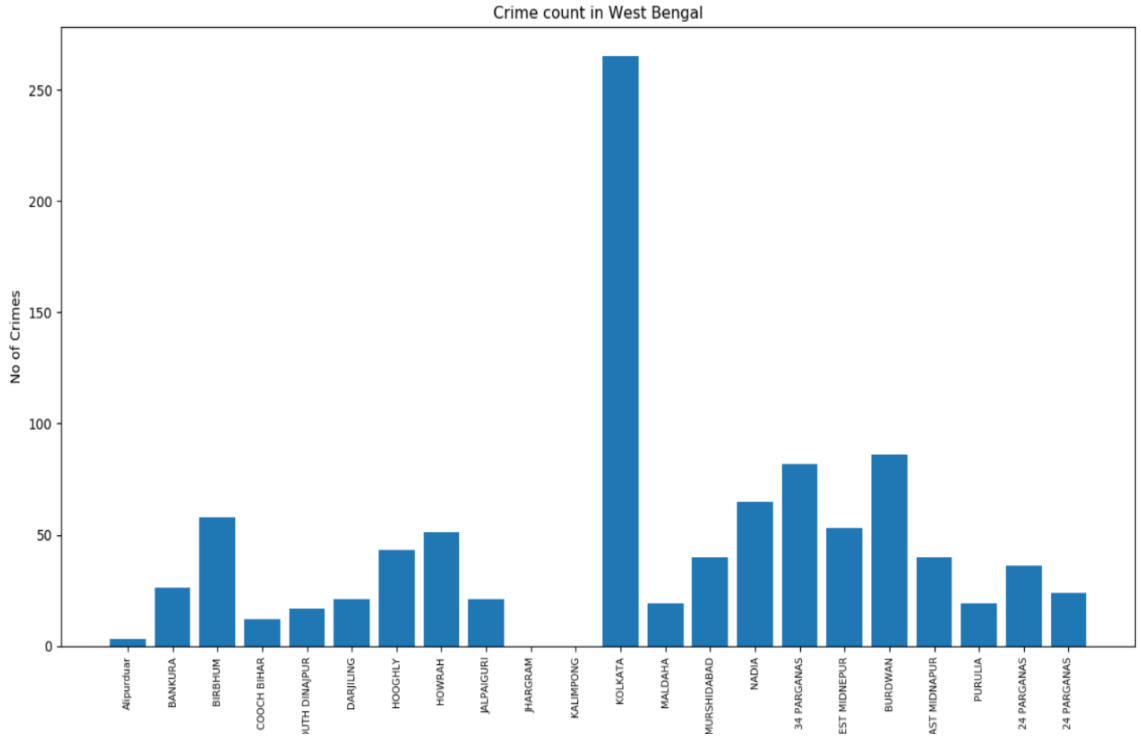


Figure 5. District Wise Crime Map

State Name	True Positive	False Positive	False Negative
Andhra Pradesh	8	0	1
Arunachal Pradesh	0	0	0
Asam	8	0	1
Bihar	11	8	5
Chatrishgarh	2	2	0
Goa	1	0	0
Gujrat	9	1	1
Hariyana	7	1	1
Himachal Pradesh	1	1	1
Jammu & Kashmir	54	0	23
Jharkhand	1	2	1
Karnatak	12	5	2
Kerala	2	0	1

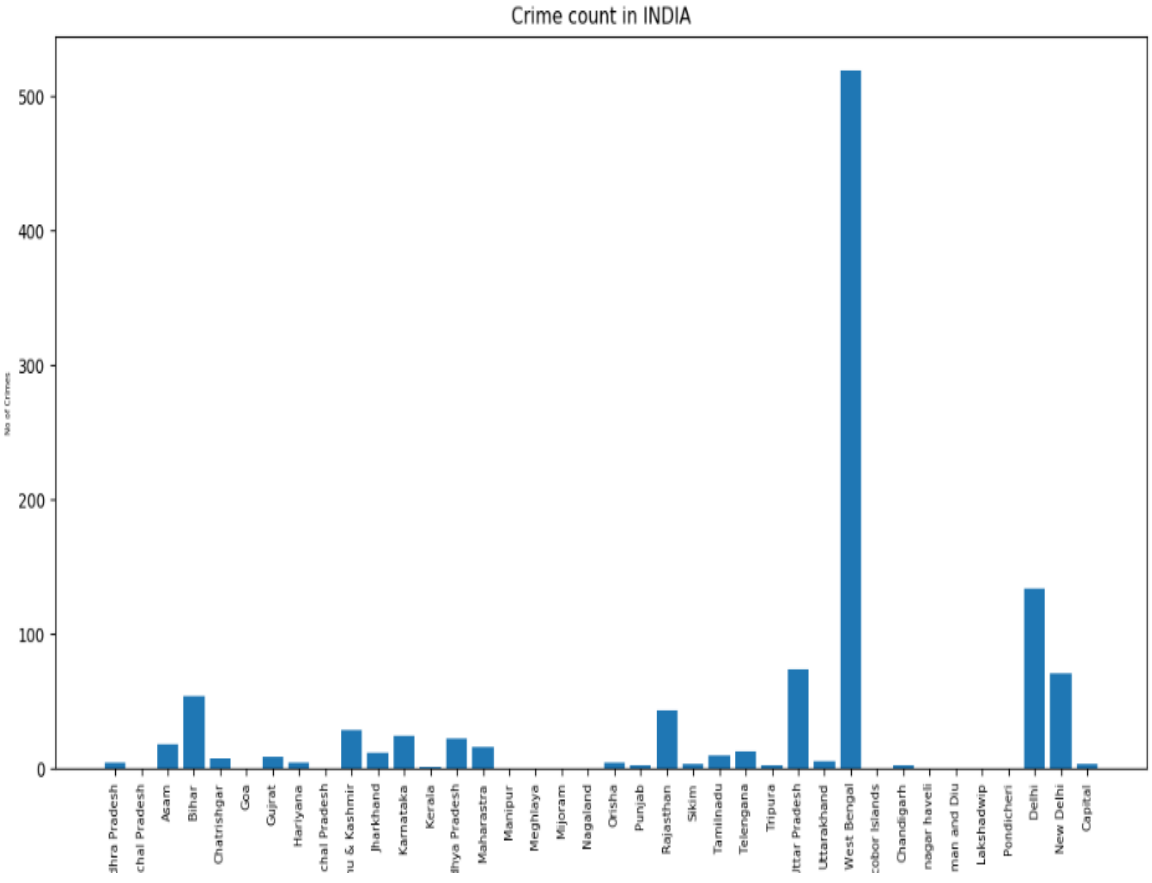
Madhya Pradesh	15	2	0
Maharastra	13	6	4
Manipur	1	3	0
Meghalya	8	0	0
Mijoram	8	0	0
Nagaland	1	0	0
Odisha	4	3	0
Punjab	6	0	3
Rajasthan	22	5	2
Sikim	1	0	0
Tamilnadu	8	2	3
Telengalna	10	12	4
Uttar Pradesh	55	4	22
Uttarakhand	1	0	3
Tripura	2	0	0
West Bengal	613	24	110
Delhi	126	6	50
Pondichery	1	0	0
Chandigarh	8	0	0
Dadra and Nagar Haveli	0	0	0

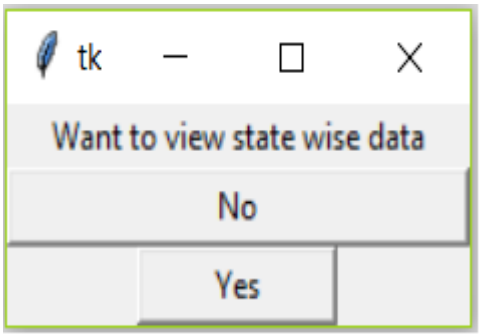
TABLE V STATE WISE CONFUSION MATRIX

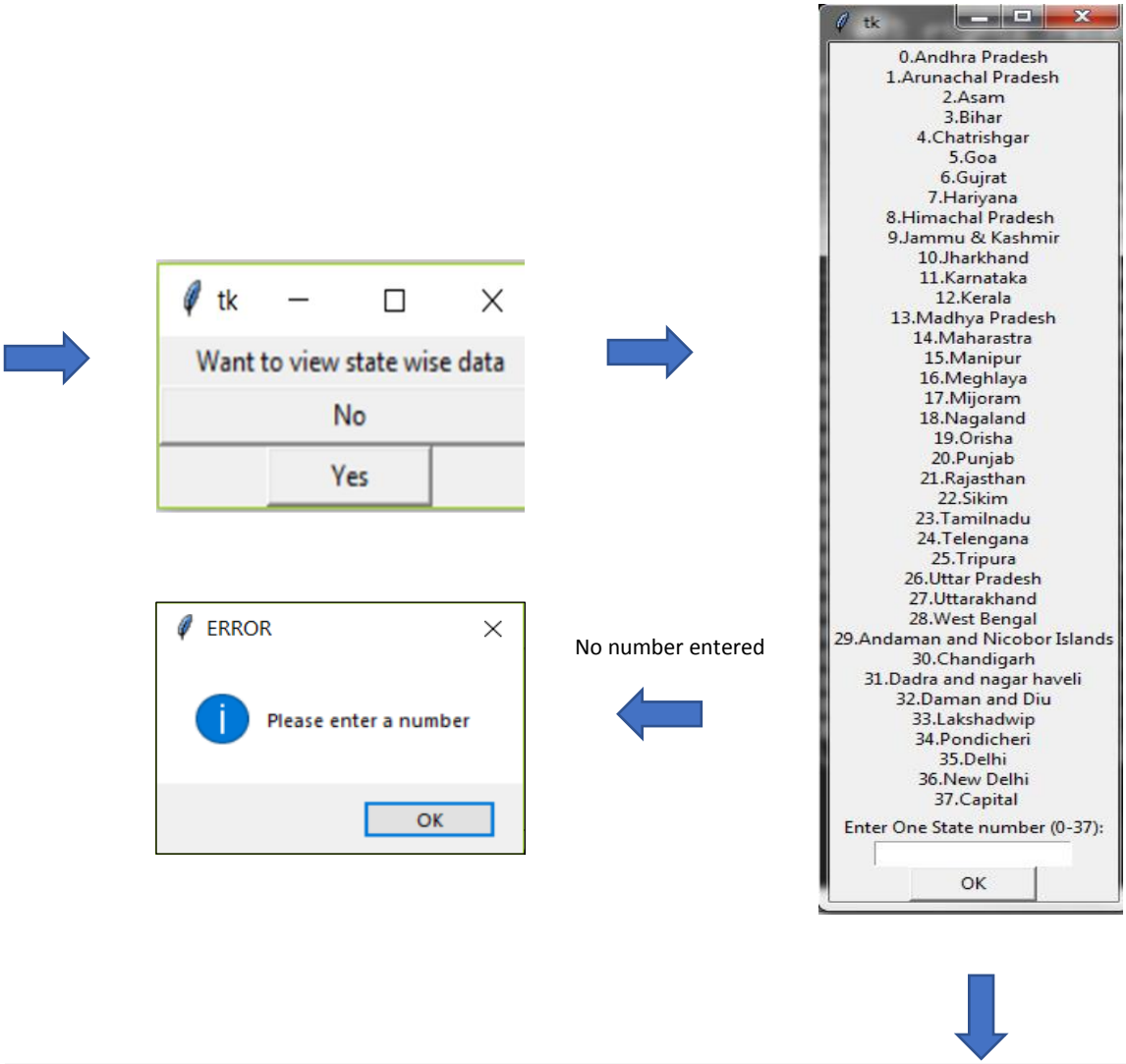
FOLD NO.	**TRAIN FOLD**	**TEST FOLD**	**ACCURACY**
1	1,2,3,4	5	52
2	1,2,3,5	4	56
3	1,2,4,5	3	46
4	1,3,4,5	2	50
5	2,3,4,5	1	50
Average Accuracy			50

TABLE VI ACCURACY FOR LOCATION TAGS

4.5. Application Screenshots







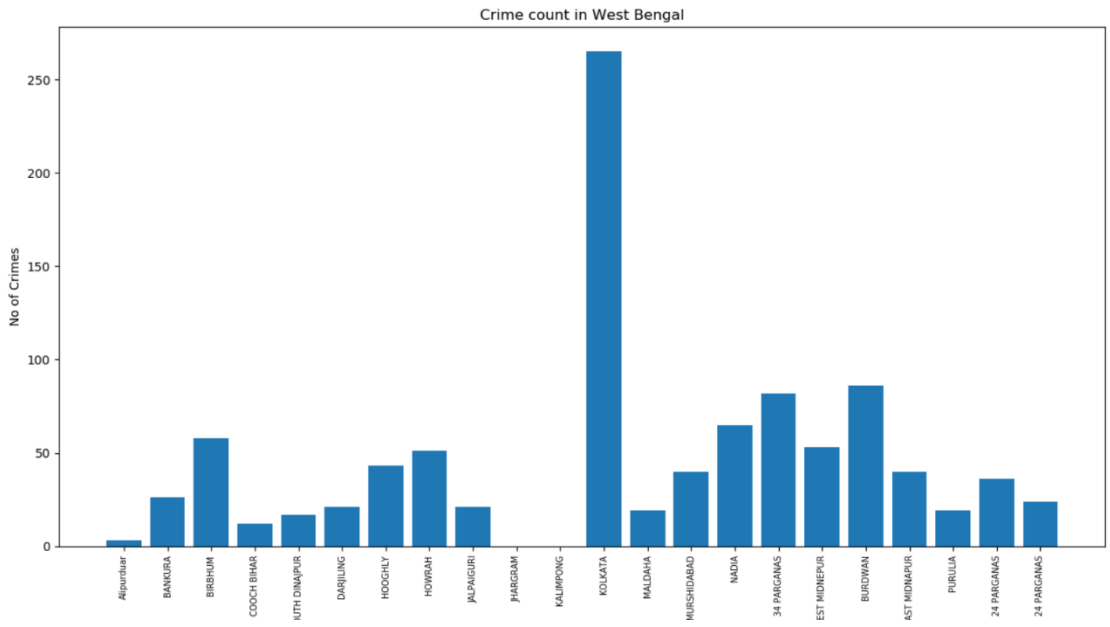


Fig 6. GUI For Map Generation

CHAPTER 5

Conclusion

The whole system is focused on the users with access to huge collection of newspaper articles within a country for faster visualization of real-time crime scenarios of the country or a state within it. The whole system is an automated process for getting the data within a timespan and retrieving the crime rates from them. More training data for the word tagging with CRF, the addition of Bengali stemmer for naïve classifier, using the advance tools like Recurrent Neural Network for sequence tagging can produce better accuracy. We are focusing to gather and tag more data for the system and to migrate to deep neural networks like RNN for better results.

5.1. Challenges:

As discussed earlier the proposed system introduces a new type of tagging system called crime tags. Name Entity Recognition tags are the closest set of tagging systems to crime tags. Implementing a whole new tagging system and defining the tag set for the work was the first challenge. As the work focuses on crime data extraction, tagging the correct set of words to a tag was a challenge in crime data collection part. Because there are so many possible sets of words which may signify one single tag, therefore, tagging them properly was one of the most important and challenging task. In crime data collection another problem is getting huge regional news. Bengali newspapers cover a huge amount of regional news but getting the news of national level from different states was difficult. Our crawling system tries to get news from different states in Bengali as much as possible. In the case of News Classification Multinomial Naïve Bayes is a great tool to work with but capturing the rare crime headlines was a difficult task therefore so many headlines were added to the training file of news

classification system by hand. Short, very complicated or sarcastic headlines spotting were also a difficult task in news classification.

After the data collection, spotting the best features from the data set to correctly tag the words was the next challenge. In these types of tagging systems POS and Named Entity tagging were previously there but works for crime tags were not available. Therefore, research was done on the dataset to identify the best features from them. Though some features were used from the Named Entity recognition tasks and POS tagging tasks, many new features were introduced in our task.

The last task of our system was feature extraction and map generation. For getting the state and district names from many small names was a challenging task. For identifying the states and district names from the tagged location names we used a hierarchically structured data. But, this data[2] was not available in Bengali. So these names were manually transliterated to the Bengali font using Google translate. Still google translate is not accurate so manual checking of them and making the necessary corrections took a huge effort.

Other than this duplicate location removal (i.e. same location name occurring in many states), data processing for CRF++ were other necessary challenges of the work.

5.2. Future Scope:

In the current system, there are also some options for the betterment of accuracy. Using Recurrent Neural Network after collecting more training data can make the system better. The news Classification module's performance can be enhanced by developing a good Bengali Stemmer.

Being a newly introduced tagging system, this system has many possibilities in the different applications. In our system, we have generated crime maps as final output. But, other than this proposed system can be used for extracting crime type, related personnel, the weapon used etc. In the case of official agencies (like police, administration) it can be used for summarizing the crime rates, crime type, over the reports with some necessary changes. For different languages, this system can also be reconstructed with some modification to the feature set. This can be done with help of different lingual researchers. If different languages are used in this system, then the crime analysis can be more accurate over the national level.

By adding different regional newspaper reports and necessary feature extraction for different language this system can act as a versatile tool for accurate crime profiling.

References

[1] Y. Zhou, Y. Li, and S. Xia, "An improved knn text classification algorithm based on clustering,," *Journal of Computers*, vol. 4, no. 3, pp. 230–237, 2009.

[2] D. Lewis, Naive (bayes) at forty: The independence assumption in information retrieval, 1998.

[3] K. Schneider, "Techniques for improving the performance of naïve bayes for text classification," *in Computational Linguistics and Intelligent Text Processing*, 2005, pp. 682–693.

[4] R. Cooley, "Classification of News Stories Using Support Vector Machines," 2000.

[5] C. hong Chan, A. Sun, and E.-P. Lim, "Automated Online News Classification with Personalization," 2002.

[6] J. S. N. Roma, Improving svm text classification performance through threshold adjustment,, 2003.

[7] S. K. and G. V, "A Trigram HMM-Based POS Tagger for Indian Languages," S. S., U. S., and B. B, Eds., vol. vol. Berlin, Heidelberg: Springer, 2013.

[8] S. Dandapat and S. Sarkar, "Part of Speech Tagging for Bengali with Hidden Markov Model," 2006.

[9] A. Ekbal and R. H. S. Bandyopadhyay.2007, "Bengali part of speech tagging using conditional random field," pp. 131–136.

[10] A. E. S. Bandyopadhyay, "Part of speech tagging in Bengali using support vector machine," 2008, pp. 106–111.

[11] S. Dandapat and SPSAL, "Part Of Specch Tagging and Chunking with Maximum Entropy Model." *IJCAI, 2007*, pp. 29–32.

[12] K.Sarkar,"A CRF based POS tagger for code-mixed Indian social media text, arXiv preprint *arXiv:1612.07956*," 2016.

[13] V. Gayen and K. Sarkar, "An hmm based named entity recognition system for indian languages: The ju system at icon 2013," *CoRR,* vol. abs/1405.7397, 2014.

[14] K. Sarkar, "Hindi named entity recognition using system combination*," International Journal of Applied Pattern Recognition*, vol. 5, p. 11, 01 2018.

[15] A. Ekbal and S. Bandyopadhyay, "Bengali named entity recognition using classifier combination," *in 2009 Seventh International Conference on Advances in Pattern Recognition*, Feb 2009, pp. 259–262.

[16] A. Chy, H. Seddiqui, and S. Das, Bangla news classification using naive Bayes classifier, 2014, 10.1109*/ ICCITechn.2014.6997369.*

[17] "Introduction to Conditional Random Fields," 2012. [Online]. Available: http://blog.echen.me/2012/01/03/introduction-to-conditionalrandom-fields/