Event Detection And Violence Recognition In Concept-Drifting Time Constraints

Abstract :

An unprecedented way is accomplished by using concept words derived from statistical context analysis between sentences which is better than traditional methods that uses only keyword representation. Through scaling to a very large dataset we proposed an algorithm which discovers, and describes events with effective keyword networks, based on their coexisting peripheral co-occurrence. In our experiments, we used real-world news, and supervised them into paraphrases by weighting the all attempted events. We evaluated our scheme by a set of terms that maximally discriminated the percussion in news which also keep the evidences.

Chapter 1 : Introduction

This research was conducted in python language, and we used the random package from python library for random index technique. In this scheme, same index repetition was totally forbidden.

Chapter 2 : Background and Preliminary

Although Bengali is the fourth largest language of the world having over 200 million native speakers but still now Bengali language does not accomplished any grammar checker for a Bengali sentence. Parsing the meaning of Bengali sentences is in an inceptive stage still now. Very few research work have been driven to parsing the Bengali sentences rather than many research activities have been accomplished on the recognition of Bengali context-sensitivity. Phonological analysis of Bengali phrases is presented in several inquiries. As for building a keyword network we had to do the phonological analysis of Bengali phrases. In Bengali we do not have the concept of small or capital letters. Unlike English, every letter in Bengali word is capital only. For this reason we find difficulties in understanding whether a word is a proper noun or not. Bengali word also contains joint letters, and does not follow the regular grammar thus for subject-object-verb (S-O-V) structure of Bengali sentences does not fit precisely. For example *: “ami bhat khai”, and “ami khai bhat” is meaningful in Bengali, and in English both means the same : “I eat rice” but it is not a meaningful sentence in English when it is “I rice eat”.* In Bengali language often we do not use verb. For example, in the Bangla sentence “*Karim Valo Chele*” has no verb but in English *(Karim is a good boy)* at least one verb must be present in a sentence.

Like some other languages, there is also a very intense intimacy with phonology in Bengali language. In various tone level of a Bengali phrase it can mean different things which can also help us to classify the identical emotion. For example *: "আচ্ছা" or "Okay" both has the same meaning but in different languages which basically means agreement . This word is used for multiple purposes like if it is meant like a confusion or question then it might be spoken in different tone than the primary tone level. There are also some other tone levels which is for expressing diverse meanings.*

An associative intrusion detection in database is complex, and a dynamic process. It has been verified that effective attributes selection improves the detection for a network intrusion detection using decision tree-based attribute. In naive Bayesian tree, nodes contain, and split as regular decision-trees, but the leaves contain naive Bayesian classifier. This algorithm maximizes the precision, and recall accomplishment for most of the circumstances.

In predictive analytics and machine learning, the **concept drift** means the statistical properties of the target variable, which the model is trying to predict, change over time in unforeseen ways. The term *concept* refers to the quantity to be predicted. More generally, it can also refer to other phenomena of interest besides the target concept, such as an input, but, in the context of concept drift, the term commonly refers to the target variable.

In law, **time constraints** are placed on certain actions and fillings in the interest of speedy justice, and additionally to prevent the avoidance of the ends of justice by waiting until a matter is argued.

In this approach Decision Tree is used to detect multiple novel class. The basic decision tree algorithm ID3 (Iterative Dichotomiser) builds decision tree. In this technique, we built a decision tree from training data points and calculate the percentage of number of data points in each leaf node with respect to data points in training dataset. Now apply cluster on each leaf node of tree based on similarity. In real time classification novel class is arrived if number of data point in leaf node of tree is increase than percentage calculated before. The idea of detecting multiple novel class is to construct graph, novel class obtained is plotted on graph. After constructing the graph identify connected component, the number of connected component determines the number of novel class Procurement.[20]

Benefits : Law Enforcement, News Publications for better filtered news

Chapter 3 : Materials and Methods

Newspaper has always been a part of our daily life, and the best public accessibility to know what is happening around us. It has always been a favor for multifarious people in our society. It is the source of data, and information that helps to increase awareness, and knowledge of the citizens in a country. Here we focused on classifying the news which is mainly the evidence for violence so that we can provide a good assistance to our law enforcement teams.

Data Collection : In Bangladesh there is a daily newspaper named Prothom Alo which is the most renowned, and reliable source of news. We collected a huge amount of news in bengali font from this newspaper for testing our approach. All the news we collected are the real world events in various time divisions between 2016 to 2018, and this data is gathered through a portal. We have collected the news with their time of publishing and the title.

Algorithm 1 : Collection of the news
Input :

- start_date
- stop_date

Declare **News_array** as 2d array
for all news between start_date to stop_date
do
for each news of a day
do
news_array = parse the date, title, and the news
end for
end for
draw news_array into comma-separated values (*CSV*) file

Output : *Example of a data point after the collection of the news from star_date to stop_date*

date	tittle	news
2017-01-01	উখিয়ার অনিবন্ধিত শিবিরের পাশে নতুন বসতি	স্ত্রী নুর জাহান ও আট সন্তানকে নিয়ে দুই মাস ধরে রাখাইনের বুড়াশিকদার পাড়ায় খালের ওপর মাচা বেঁধে দিন যাপন করছিলেন মোহাম্মদ হোসেন। কিন্তু গত তিন-চার দিনে ওই এলাকায় মিয়ানমারের সশস্ত্র বাহিনী ও নাডালা বাহিনীর নতুন অভিযান শুরুর পর প্রাণ বাঁচাতে তাঁরা বাংলাদেশে এসেছেন।.....

Preprocessing : After collecting the news we had to preprocess our data. In preprocessing we removed all English symbolic letters, and some Bengali characters*(ex : ঁ, ং, ক্ষ etc),* and all the numbers. We also removed some intimations, and they are quotation, double quotation, exclamatory, question mark, colon, semicolon, comma, brackets, backslash, forward slash, percentage, equal and many more. We also removed bengali full stop(।) from the news.

Algorithm 2 : Preprocessing the data set
Input : data_set [*All news with the title and date*]
whitespace = *u"[\s\u0020\u00a0\u1680\u180e\u202f\u205f\u3000\u2000-\u200a]+"*
bengali_digits = *u"[\u09E6\u09E7\u09E8\u09E9\u09EA\u09EB\u09EC\u09ED\u09EE\u09EF]+"*

english_chars = *u"[a-zA-Z0-9]"*
punctuation = *u"[(),$%^&*+={}\[\]:\"|\'\~`<>/,□¦!?½£¶□¼©□□□⅓⅔⅕⅖⅗⅘⅙⅚⅛⅜⅝⅞⅟□¤¿°;-]+"*
bangla_fullstop = *u"\u0964"*
punctuation_sequence = *u"['\"“”‘’]+|[.?!,…]+|[:;]+"*
for each news in data_set
do
if news associate with whitespace **or** bengali_digits **or** english_chars
or punctuation **or** bangla_fullstop **or** punctuation_sequence
cleaned_text = remove these characters from news
end if
replace corresponding news from data_set by cleaned_text

Feature Procurement : So now our data set is ready we can accumulate feature from the news for the event classification. As we proposed the benefit is for the law enforcement teams, we settled our event on focusing only in violence so that we can appoint the most violent areas, time periods and crimes. At the origin of our approach we had to build a keyword network for balancing our features. Each feature is based on multiple keyword where every keyword illustrates their identification for the feature. A single keyword beneath a feature votes for their identification in a single news, and all the votes by each keyword turn into the weight for the feature. Likely we created six keyword networks by human expert assessment that represents the weight for individual feature which is relevant, and connected to a single type of violence. The keyword network representatives are murder, kidnap, hassle, protests, accident and terror. Here is the specimen of the keyword network named murder : *"খুন", "নিহত", "ঘাত", "আঘাত", "হত্যা", "গুলি", "চাকু", "বন্দুক", "পিস্তল", "আগ্নেয়াস্ত্র", "ছুড়ি", "অস্ত্র", "সশস্ত্র", "রক্তপাত", "মার", "মেরে", "লাশ", "মৃত", "ঘাতক", "পিটিয়ে"* **/**
We schemed our process as every news is an evidence for a event, and for the detection of the event location we had to create some keyword networks which measures the location for a single event. We declared eight keyword networks which is the all divisions of Bangladesh. Each of this keyword network holds districts of the corresponding division as keyword. Here is the specimen of the keyword network named sylhet : *"সিলেট", "মৌলভীবাজার", "হবিগঞ্জ", "সুনামগঞ্জ", "বালাগঞ্জ", "বিয়ানীবাজার", "বিশ্বনাথ", "কোম্পানীগঞ্জ", "ফেঞ্চুগঞ্জ", "গোলাপগঞ্জ", "গোয়াইনঘাট", "জৈন্তাপুর", "কানাইঘাট", "জকিগঞ্জ", "দক্ষিণ সুরমা", "ওসমানী", "বড়লেখা", "কমলগঞ্জ", "কুলাউড়া", "রাজনগর", "শ্রীমঙ্গল", "জুড়ী", "নবীগঞ্জ", "বাহুবল", "আজমিরীগঞ্জ", "বানিয়াচং", "লাখাই", "চুনারুঘাট", "মাধবপুর", "বিশ্বম্ভরপুর", "ছাতক", "জগন্নাথপুর", "দোয়ারাবাজার", "তাহিরপুর", "ধর্মপাশা", "জামালগঞ্জ", "শাল্লা", "দিরাই"।*

Algorithm 3 : Balancing the weight of a news for each keyword network
Input :
- data_set
- keyword_network

for each news in data_set
do
initiate with day, month, year, and news from data_set

for each word of a news
do
for each attribute of the keyword networks
do
if news word associate with any keyword
count the keyword attendance for the corresponding network
end if
accomplish the weight of the networks for a news
end for
end for
write values into a comma-separated values (*CSV*) file by the day, month, year and the accomplished weight of the networks for a news
end for

Output : *After procurement of the feature we isolated the dataset into two tables*

Table for time division : *1.0*

news_id	day	month	year	murder	kidnap	hassle	protests	accident	terror
0	8	5	2018	39	9	30	3	2	3
1	10	5	2018	7	1	22	3	5	1
2	24	5	2018	2	0	1	0	1	4

Table for area division : 1.1

news_id	barisal	chittagong	dhaka	khulna	rajshahi	rangpur	sylhet	mymensingh
0	0	8	6	0	0	0	0	0
1	0	0	0	0	0	1	0	0
2	0	1	6	2	0	1	0	0

Chapter 4 : Bottom Up Hierarchical (BUH) Classifier

In this experiment, our goal is to detect event with the recognition of exact violent news by their violence criterias. After using many classifier algorithms we couldn't reach at our expectations. So we constructed a classification algorithm which prevents concept-drifting problem with better tf–idf evidence and utilizes a supervised learning technique. Our principal goal is to giving the proper importance for individual keyword network so that we can classify the event and recognise their criteria of violence. We have got very surprising results from this BUH Classifier. In the beginning, we need to learn about the keyword networks. Here is our all keyword network with all keywords.

Murder	খুন	নিহত	ঘাত	আঘাত	হত্যা	গুলি	চাকু	বন্দুক	পিস্তল	আগ্নেয়াস্ত	ছুড়ি	অস্ত্র	সশস্ত্র	রক্তপাত	মার	মেরে	লাশ	মৃত	ঘাতক	পিটিয়ে
Kidnap	অপহরণ	হরণ	তুলে	গুম	ক্ষিপ্র	জোর	পাচার	দখল	পীড়ন	নিপীড়ন	শিকার	অত্যাচার	নির্যাতন	জুলুম	জবরদস্তি	বলাত্কার	নারীধর্ষণ	ধর্ষণ	ধর্ষিত	ধর্মনাশ
Hassle	মারামারি	হামলা	আহত	ধেয়ে	ধাওয়া	পাল্টা	আক্রমণ	হানা	ধোলাই	দ্বন্দ্ব	চক্রান্ত	ষড়যন্ত্র	ঝগড়া	সংঘর্ষ	দাঙ্গা	লড়াই	কলহ	সহিংস	সংঘাত	শত্রুতা
Protest	বিক্ষোভ	সভা	সমাবেশ	হরতাল	ছত্রভঙ্গ	হল্লা	বিবাদ	বিরোধ	গোলমাল	প্রতিদ্বন্দ্বিতা	সংগ্রাম্	আন্দোলন	মিছিল	অবরোধ	নাশকতা	ধর্মঘট	আলোড়ন	বিশৃঙ্খলা	ঝামেলা	প্রতিবাদ
Accident	দূর্ঘটনা	ভিড়	আকস্মিক	দুর্দশা	সংকট	বাগিয়ে	কবল	এলোপাতাড়ি	ক্ষত	ক্ষয়	বিপর্যয়	সর্বনাশ	আচমকা	হটাৎ	অপ্রত্যাশিত	দৃঢ়ভাবে	দুর্বিপাক	বিপাক	আপদ	আর্তনাদ
Terror	আতঙ্ক	বিস্তার	হৈচৈ	ভয়	উত্তেজনা	রেশ	ক্রুদ্ধ	ক্রোধ	বিপদ	অশান্তি	অস্থির	উপদ্রব	অসাধুতা	নিয়মভঙ্গ	উপদ্রব	প্রভাব	চক্র	চাঁচল্য	আশঙ্কা	নালিশ

So each keyword network holds twenty keywords. This keywords are trying to classify their corresponding networks which is helping to recognise the violence from the news data set by their criteria. We counted the presence of each keyword in two ways. One approach is binary appearance and other one is maximum appearance. So by the all consideration we should get three set appearance for an individual news.

By the previous method we get a single row for an identical news. We have counted each keyword network importance with total appearance Here is the example of a news which holds the id as "11" .

id	murder	kidnap	hassle	protest	accident	terror
11	3	33	1	3	0	0

Table D#1 : Weighted by total appearance of each keyword for their corresponding network.

In this system the weighting approach is not so strong for a classification algorithm. So we intend to give each keyword network a better rank by their corresponding keywords. In Details, for the requirement of the weight named as murder keyword network, we set the column murder as class variable and other networks as feature variables. Then we classified the characteristic for murder by a popular classification algorithm and set the training accuracy as the rank for the

murder network. Same process occurs for each network for their proper ranking. Other networks also become the class variable when the ranking is needed. For a individual news, class variable bends six times because we have six keyword networks.

In our system we ranked every attribute by two classification algorithm with two types of weighted keywords. We used Neural Network classifier which provides multilayer perceptron and we also used k-Nearest Neighbour classifier for both appearance approach of an individual news, which is the binary appearance and the maximum appearance.
In the following table D#2 we counted the keyword appearance as a boolean variable. Such as if "খুন" exist in the news then "খুন" keyword gets its importance 1 for murder network, If not then its importance is 0.

id	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
murder	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidnap	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0
hassle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
protest	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
accident	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
terror	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table D#2 : Binary presence of keywords for news id 11.

In the following table D#3 we counted the keyword appearance as much as it appeared in an individual news. Such as if "খুন" exist in the news then "খুন" keyword gets its importance the total amount it appeared in the news for murder network, If not then its importance is 0.

id	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11	11
murder	0	0	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
kidnap	0	0	0	0	0	0	0	0	0	0	5	0	19	0	0	0	0	9	0	0
hassle	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
protest	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0	0	0	1
accident	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
terror	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table D#3 : Maximum presence of keywords for news id 11.

Each news goes through this process and collects their rank for the corresponding keyword networks which exposes the importance of their violence criteria. After this process through two classifiers we achieve four types of importance for an individual news. Two types of importance

for Neural Network and K-Nearest Neighbour classifier attempts to find the impact of each keyword network on the news. Now we can use any other classifier to acknowledge the violence of the news. In table R#1 and R#2 we have accordingly binary and maximum presence by neural network classifier and in table R#3 and R#4 we have accordingly binary and maximum presence by K-NN classifier.

id	murder	kidnap	hassle	protest	accident	terror
11	1	1	0.05	1	1	0.1

Table R#1 : Weighted by neural network classifier with binary presence.

id	murder	kidnap	hassle	protest	accident	terror
11	1	1	0.95	1	1	1

Table R#2 : Weighted by neural network classifier with maximum presence.

id	murder	kidnap	hassle	protest	accident	terror
11	1	1	0.933	1	1	0.933

Table R#3 : Weighted by KNN classifier with binary presence.

id	murder	kidnap	hassle	protest	accident	terror
11	1	1	0.933	1	1	0.933

Table R#3 : Weighted by KNN classifier with maximum presence.

BUH Classifier Algorithm :
Input :

- data_set [Binary or Maximum appearance]

total_news = (data_set length)/20
for each news in the range of total_news
tracking_id = get the id of the news from the data_set
do
for each 20 rows from the data_set by the tracking_id
do
keyword_appearance_array = accumulate the appearance of the keyword for tracking_id news
end for
for each column in keyword_appearance_array
do
y_train = corresponding column as class variable
x_train = all columns except the class variable
clf = fit x_train and y_train into the classifier

accuracy = get training score of clf by x_train and y_train of tracking_id
network_ranked_array = accomplish the accuracy as the rank of the class variable keyword network of the tracking_id
end for
end for
write values of network_ranked_array into a comma-separated values (*CSV*) file

Output : Table R#1, R#2, R#3 and R#4 type ranked network for their corresponding data_set.

Now we can fit this datasets into a classifier and observe the result with comparing the previous approach. Here we used decision tree for the detection of the event and recognised the violence by neural network and K-NN classifier. In this system R#1, R#2, R#3 and R#4 dataset stands for the solution of concept-drifting error which supports the decision tree classifier to detect the event.

Chapter 5 : Results and Diagrams

Here we meet the problem of detecting events from multiple and heterogeneous news. This heterogeneity makes the event detection task more challenging, hence we accomplished a very potential approach. We are able to automatically detect, and measure the violence weight when a new real world event has occurred, and also able to classify and cluster the news by any kind of events. In this module the textual similarity between the event keyword network and news words is measured potentially. This module can classify and cluster by any potential topics but the concernment must be switched to the corresponding event. We can also classify events into categories such as plane crashes, economic collapses and natural disasters. Some classifiers algorithm has been tested for better accuracy such as Naive Bayes Classifiers, k-Nearest Neighbors (k-NN) and Decision Tree. There is several steps of refinement which increases the exactness of the result of the classifiers. We can reduce the number of **false positives** by using heterogeneous classifier and cluster algorithm which is entitled as ensemble learning.

As we refine and improve this module we need to revise how we are calculating the importance of a news. For example different production should have different weights such as if a news contains more than one area we can increase the importance. The objective of event detection is to detect episodic related stories from a massive news collection. In information retrieval, **tf–idf**, short for **term frequency–inverse document frequency**, is a numerical statistic that is intended to reflect how important a word is to a document in a collection. It is mostly used as a weighting factor in searches of information retrieval, text mining and user modeling. The tf-idf value increases proportionally to the number of times a word appears in the document, and is offset by the frequency of the word in the corpus, which helps to adjust for the fact that some words appear more frequently in general. Here, tf-idf was applied to observe words with the purpose of conducting event matching as violence recognition.

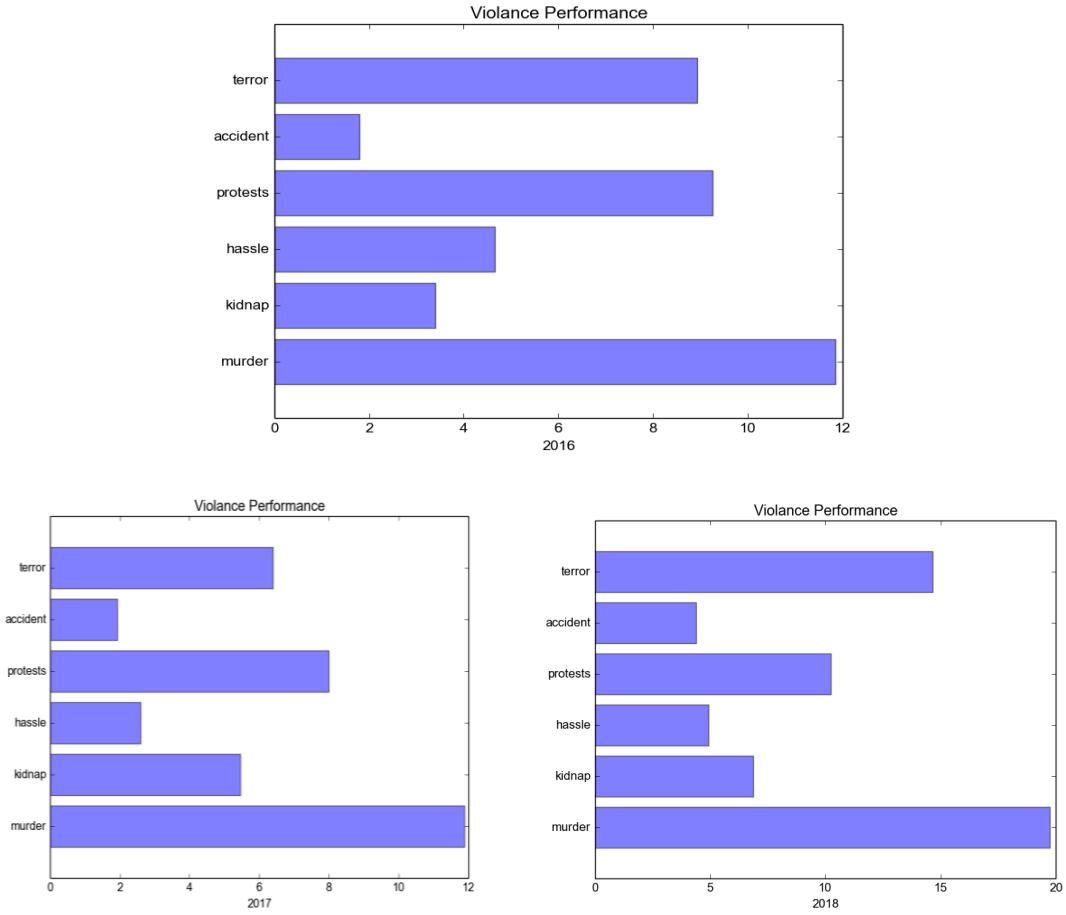
Potential concept terms are the key terms which is united by the news. Concept terms can be used for dealing with the problems of lexicon altering, and accordingly the idea of using

concepts has been applied for the quest of propagation. It combines the technique of global analysis and local feedback between a quest. A major problem is ignored by most of the classification techniques, which is concept-evolution, that means the appearance of a novel class. In case of intrusion detection, a new kind of intrusion might go undetected by traditional classifier, but our approach should not only be able to detect the intrusion, but also deduce that it is a new kind of intrusion. This scheme would lead to an intense analysis of the intrusion by human experts in order to understand its cause, find a cure, and make the scheme more assured. The detection process can be done in unsupervised way, but supervision is necessary for classification. Without external supervision, two separate clusters could be regarded as two different classes although they are not. Conversely, if more than one novel classes appear simultaneously, all of them could be regarded as a single novel class if the labels of those instances are never revealed. Furthermore, traditional novelty detection techniques simply identify data points as inconsistent that deviate from the normal class. But our scheme not only detects whether a single data point deviates from the existing classes, but also uncover whether a group of such outliers possess the potential of forming a new class by showing strong cohesion among themselves. Therefore, our scheme is a “multi-class” classification model and a novel class detection model.
[*ref :Classification and Novel Class Detection in Concept-Drifting*]

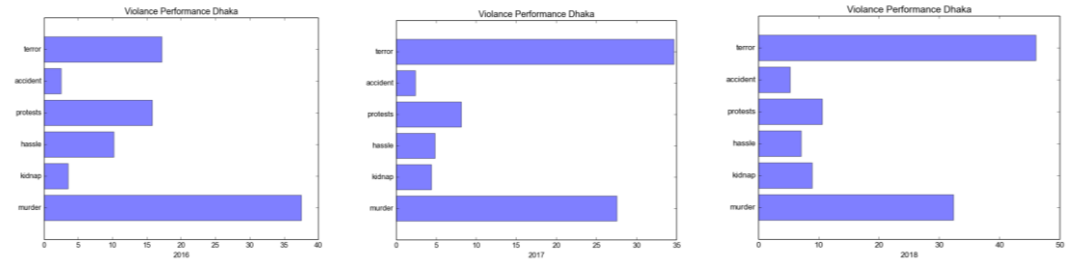
In table C#0 we can see the discrimination of all approaches with following BUH classifier, which is performed only in the news of 2018. We have 2844 numbers of news following by their id and six keyword networks.

Fact	Keyword network	Binary BUH of KNN	Maximum BUH of KNN	Binary BUH of neural network	Maximum BUH of neural network
Training set	0.542	0.653	0.619	0.635	0.638
Test set	0.485	0.661	0.238	0.621	0.613
Total number of misclassified news	418	275	619	308	496

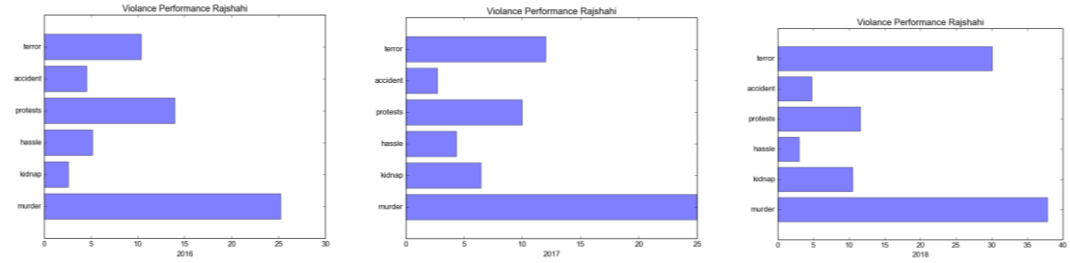
Table C#0 : The discrimination of all approaches



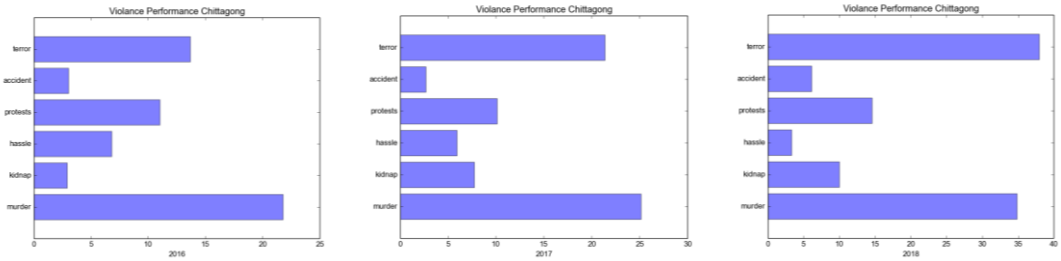
Graph 1.0 : All keyword networks frequency in a year



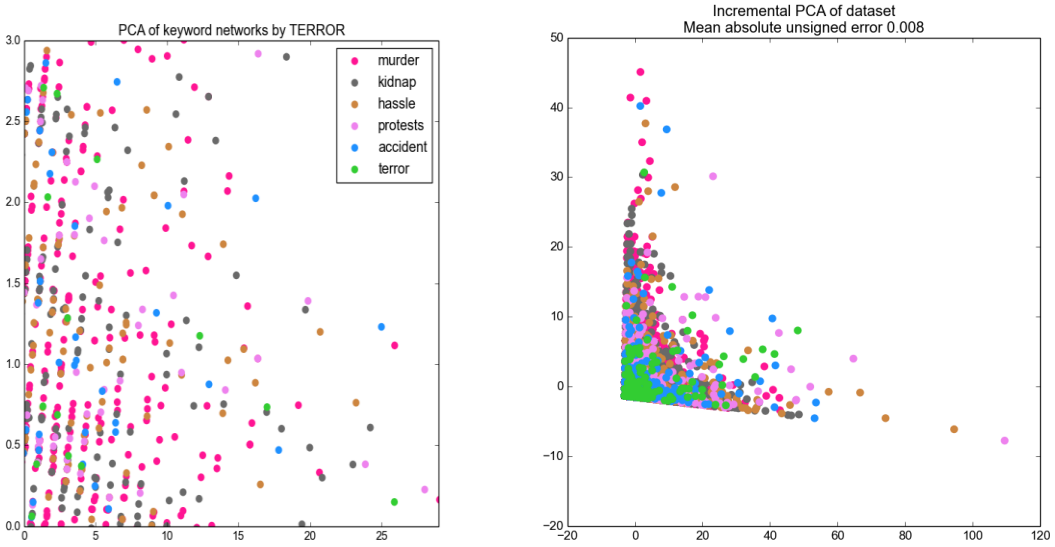
Graph 1.1 : All keyword networks frequency in Dhaka



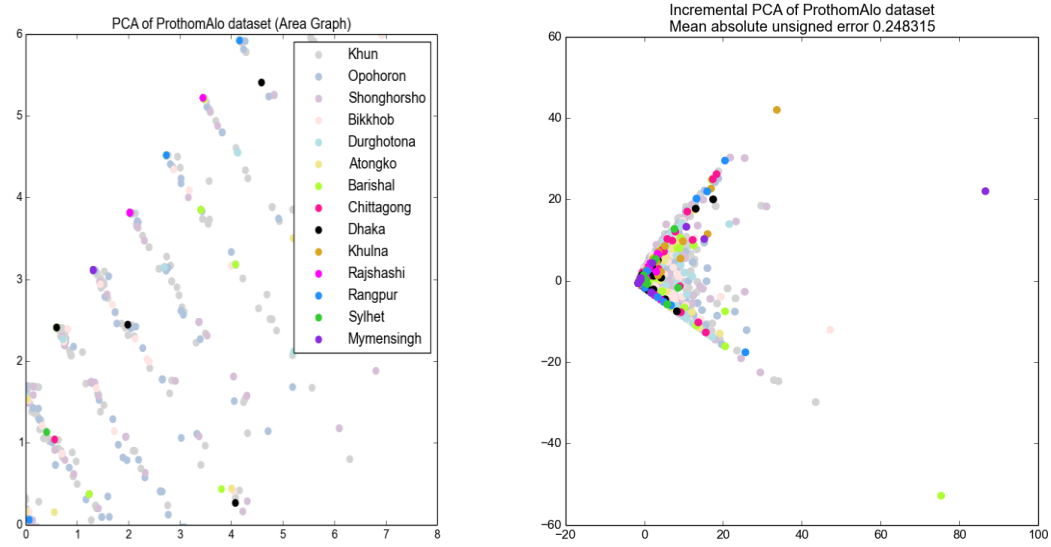
Graph 1.2 : All keyword networks frequency in Rajshahi



Graph 1.3 : All keyword networks frequency in Chittagong



Graph 2.0 : Principal component analysis (*PCA*) for Terror



Graph 2.1 : Principal component analysis (*PCA*) for Dhaka

Chapter 5 : Conclusion

This paper summarizes the current techniques in concept drift and novel class detection for event detection and violence recognition. First it introduce with the concept drift and novel class and its importance in today's world and in real application.

REFERENCE

[20] Amit Biswas, Dewan Md. Farid and Chowdhary Mofizur Rahman A New Decision Tree Learning Approach For novel Class Detection In Concept Drifting Data Stream Classification journal of computer science and engineering, volume 14, issue 1, july 2012