

Pattern Mining From Unlabeled News Article Dataset Using Semi-Supervised Learning



Iftakhar Ali Khandokar

MSCSE-Id: 012-212-023

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of
MSc in Computer Science & Engineering

January 2023

Abstract

Text classification is one of the prominent tasks in the field of Natural language Processing as day by day the amount of textual data is growing rapidly, Therefore it is an emergent demand to build some kind of knowledge model out of this growing data to extract the internal information out of the data samples due to the limitation of memory and computational power. One such example of this kind of rapidly growing data is the news articles produced daily by the vast amount of news publication platforms. Therefore in this work, we would like to introduce an automated approach to detect target events from these textual news articles and the type of events that are related to violent incidents where multi events labels will be detected within the news articles and extract several types of information from the news articles. To ensure the authenticity of the classification and event pattern analysis we adopted a semi-supervised approach where a small volume data predictive model is used to amplify the dataset eventually gathering enough data to feed it into the Deep net. To categorize the type of events are been detected are Murder, Rape, Kidnap, Clash, Suicide, and Teen Suicide where we experimented with multiple feature extraction techniques like N-gram, TF-IDF, Word-2-Vec, Fasttext, and BERT from which the BERT-based classifier achieved the highest accuracy of 92% in the provided test set. Utilizing the best-performing model we conducted Trend and Pattern analysis experiment on a five-year period of time series data which reveals some exciting insight information that is related to or affected by these violent events concerning various units of time.

Acknowledgements

I would like to thank my advisor Dr. Swakkhar Shatabda. His guidance and direction throughout the thesis have been invaluable, and I greatly appreciate the new perspective he provided. His energized hints, visionaries, attentive remarks and proposals, and exceptional help at each phase of my MSc were acknowledged and fundamental. His capacity to muddle me enough to at last answer my very own inquiry effectively is something profitable that I have realized, and I would endeavor to imitate it if at any time I get the chance.

I also thank Dr. Nurul Huda, MSCSE Director, for his constructive feedback on coursework and for being an affable minor advisor. I want to thank Chowdhury Mofizur Rahman, Vice Chancellor, for giving me the opportunity. And, finally, I want to acknowledge my friends and family for bearing with me and having a sympathetic ear whenever I faced struggles.

Contents

List of Figures	vi
List of Tables	vii
1 Introduction	1
1.1 Motivation	1
1.2 Objectives of the Thesis	2
1.3 Thesis Contributions	2
1.4 Organization of the Thesis	2
2 Literature Review	4
2.1 Semi-Supervised Based Approach	4
2.2 Active Learning Based Approach	5
2.3 Deep Learning Based Approach	5
2.4 Unsupervised and HMM Based Approach	7
2.5 Event and Knowledge Extraction from Text	7
3 Methodology	9
3.1 Training Data Collection	9
3.2 Pre Processing Collected Data	10
3.3 Target Data points Extraction	10
3.4 Predictive Model Training Using Different Features	11
3.5 Large Data set Construction & Experimenting With different features	11
3.6 Test Data Collection & Comparing Model of different Volumes	12
3.7 Fetching Time Series Data & Mining Meaningful Knowledge	12
4 Performance Analysis	13
4.1 Performance Evaluation of the Small Volume Classifiers	13
4.2 Performance Evaluation of the Large Volume Classifiers	14
4.3 Performance Comparison between the Large and Small Volume Classifiers	15
4.4 Pattern and Trend Analysis	16
4.4.1 Event frequency Histogram Analysis	17
4.4.1.1 Hypothesis Test - Most Frequently Occurring Violent Event	17
4.4.2 Yearly Violent and Non-Violent Event Article Comparison	18

4.4.2.1	Hypothesis Test - Most Event Impact full Year	19
4.4.3	Yearly Violent and Non-Violent Event Article's Occurrence Ratio Comparison	20
4.4.4	Violent Event Occurrence Comparison Over the Demi-Decade	20
4.4.5	Probability Distribution of Events on Different Day Intervals	21
4.4.6	Area Histogram of Different Events Over the Demi-Decade	22
4.4.7	Violent Event Dominance Comparison Over Different Months	23
4.4.7.1	Hypothesis Test - Half Yearly Period Event Ratio	23
4.4.8	Active Age Group Analysis in Violent Events	24
4.4.8.1	Hypothesis Test - Event Associated Most Active Age Group	24
4.4.9	Monthly Heat-Map of Violent Events Over the Years	25
4.4.10	Weekly Heat-Map of Violent Events Over the Years	26
4.4.10.1	Hypothesis Test - Event Distribution Ratio on Different Weekdays . .	27
4.4.11	Geographical Area Analysis Based on Violence Intensity	27
4.4.11.1	Hypothesis Test - On Divisions Violence Event Ratio	28
4.4.11.2	Co-relation test between Divisions population count and Violence Ratio	29
4.4.11.3	Co-relation test between Divisions population count per Km and Violence Ratio	29
4.4.11.4	Co-relation test between Divisions Area Km square and Violence Ratio	30
4.4.12	Co-Relation Between Target Violence Events	30
5	Conclusions	32
5.1	Summary	32
5.2	Future Work	32
	Bibliography	33

List of Figures

3.1	Full System Workflow Diagram.	9
4.1	Performance Comparison of Different Feature-Based Classifiers on Small Volume Set.	14
4.2	Performance Comparison of Different Feature-Based Classifiers on Large Volume Set.	15
4.3	Performance Comparison of Small and Large Volume Set Classifiers Based on the Trained Set.	16
4.4	Performance Comparison of Small and Large Volume Set Classifiers Based on the Test Set.	16
4.5	All Events Frequency Histogram.	17
4.6	Yearly Violent & Non-Violent Event Article Comparison.	19
4.7	Yearly Violent and Non-Violent Event Article's Occurrence Ratio Comparison.	20
4.8	Violent Event Occurrence Comparison Over the Demi-Decade.	21
4.9	Probability Distribution of Events on Different Day Interval.	21
4.10	Area Histogram of Different Events.	22
4.11	Violent Event Dominance Comparison Over Different Months.	23
4.12	Active Age Group Analysis in Violent Events.	24
4.13	Monthly Heat-Map of Violent Events Over the Years.	26
4.14	Weekly Heat-Map of Violent Events Over the Years.	26
4.15	Geographical Area Analysis Based on Violence Intensity.	28
4.16	Pearson Relation Between Violence Events.	31

List of Tables

4.1	Performance Comparison of different feature classifiers using small volume data set on train set and test set.	14
4.2	Performance Comparison of different feature classifiers using Large volume data set on train set and test set.	15
4.3	P_{value} of Wilcoxon Signed-Rank Test of Different Events.	18
4.4	Two-Tail Wilcoxon P-Value table of Different Years Violence Record Frequency. . . .	20
4.5	Two-Tail Wilcoxon test P-value comparison table of different Age groups.	25
4.6	Events Wilcoxon Signed-Rank Test Pvalues Comparison Table of Weekdays.	27
4.7	Wilcoxon Signed-Rank Two-Tailed test P-value Comparison table of Divisions.	29
4.8	Pearson Relation Value Table Between Violence Events.	31

Chapter 1

Introduction

The following work is based on the Natural Language Processing domain and aims with reasoning with textual data. The form of textual data used in this study is news articles which are rapidly growing with every unit of time. The high velocity and massive volume of the news data make it very optimal to conduct any kind of affiliated pattern and trend analysis which we also intend to accomplish in this work. The necessity for such a type of analysis is that the increasing amount of data needs to be put through some method or compression mechanism which will portray the full insight information of the whole data despite preserving the entire dataset. This will not only dig out inner aspects but also will represent the data in a memory-efficient manner. In this chapter, we would highlight the problem we aim to deal with and also the proposed method of doing it. In the first section 1.1, we have shed light on the background of these tasks like the current status of the task what factors are missing, and the challenges that follow. In the second section 1.2, we present the objectives of the work and our proposed solution for overcoming the rising challenges. In the next section 1.3, we have stated the contribution of the work done, and the organization of the rest of the articles is articulated in the final section 1.4.

1.1 Motivation

The advancement of the Natural Language Processing domain has paved the way for developing some significant applications, which were considered very hard-labored [1]. One of the most exciting and useful tasks in the NLP domain is text classification and clustering. The following work also focuses on a similar aspect which is the detection of Violence events from textual news articles. In previous studies, several works have been found like news classification or clustering where news belonging to a similar domain is grouped together with specific labels, [2]. Although there hasn't been any such work that focuses on detecting any specific event from the contents of the news articles. To fill the void we propose a method for detecting occurred events from the textual news articles and the event that has been considered for the study is a violent event like Murder, Rape, Kidnap, etc. Inspiration for preferring such events is the negatory impact caused by these events on society, hence this is an attempt to dig down on the various related aspects to these events. Another aspect of concession regarding the study is analyzing the vast volume of data from the past period archive will help us reason with

the underlying related factors, which we intend to conduct in the time-series data experiment of the article [3].

1.2 Objectives of the Thesis

As we discussed earlier the target of this Thesis is to build an event detection model from textual Bangla news articles to classify news articles that confirm the occurrence of a certain types of violent events. Due to the absence of such a news article dataset, we adopted a web crawling technique to collect our Bangla news article dataset and started with identifying such online news portals that have a rich archive of reported news articles. The volume of the collected dataset is very heavy due to the high number of different categories and topics. We adopted a semi-supervised approach to label the dataset, which is incomprehensible to accomplish with a supervised and unsupervised approach. Using the stated approach we wanted to utilize and extract different text processing features from the collected dataset such as TF-IDF, Word 2 Vec, N-Gram, etc [4]. Each of the features will provide a different detection model having the ability of its unique kind from those we want to determine the most prominent classifier. Exploiting the best Detection model we want to conduct a time-series study on unlabeled demi-decade news data regarding the selected violent events to analyze various insight patterns.

1.3 Thesis Contributions

Stating the problem that we're addressing in this work and discussing the challenges that are to be overcome here we would like to list the contributions of this paper in a listing tier:

1. A semi-Supervised approach to labeling a large unlabeled textual news article dataset where we utilized a small volume dataset trained news article classifier [5] to overcome the manual labeling labor.
2. A fully labeled news article dataset where we intervened with finalizing the labels and addressed the problem as a multi-class classification as many new articles contains which isn't considered in [5].
3. Conducting Pattern and Trend information extraction experiment on a time series of unlabeled textual news data of five years (2014-2018) collected from the Prothom Alo online news archive.
4. Applying the statistical tests to analyze clear-sightedness the results of the time series of unlabeled textual news data along with investigating the relationship among various entities from the established facts.

1.4 Organization of the Thesis

The organization of the Thesis is in the following order, in the Second Chapter 2 some of the Related works to the Thesis domain are highlighted which motivated and assisted the target process, The

methodology of the Thesis is presented in the following Chapter 3, moving on to Chapter 4 where the performance analysis of the proposed methods are discussed and depicted along with some of the pattern analysis experiments related to the target events are claimed and analyzed with statistical significance, Finally, in Chapter 5 the conclusion and potential future works of the Thesis is discussed.

Chapter 2

Literature Review

In this section, we want to elaborate on some of the most fascinating works done in the field of Natural Language Processing and Knowledge mining from textual data in the last decade. This chapter is divided into four sections and each section contains ceratin domain type text processing work. In the first section, we reviewed semi-supervised text classification tasks which means working with labeling or providing knowledge on a certain subspace of the whole dataset later on classifying the whole dataset. The second section contains some active-learning-based approaches where the performance is enhanced with human intervention in the classification loop and the third section describes the deep learning-based approach which includes complex networks and high-volume data for the conclusive approach we have also gone through some unsupervised and HMM-based approaches given in the fourth section. Finally, in the fifth section, we have reviewed some prominent works on event and pattern extraction works from textual data content which is also the focused target problem of this work:

2.1 Semi-Supervised Based Approach

Throughout time various text classification works have been carried out by numerous researchers in [6] the authors conducted a study on such types of work from which the authors concluded that the nature of the data plays an important role rather than the selection of classification algorithms. In [7] the authors pointed out the tasks related to supervised text classification like data cleaning, labeling pre-processing and training, and highlighted some of the least visited areas like feature construction, textual feature selection which the authors of [8] addressed discussing along with other problematic solution like dimensionality reduction. Words are the prime targets in text classification therefore identifying these important terms can be time-consuming to discourse this obstacle weighting mechanism can be applied described in [9] along with the phenomena of single vs multi labels which have paved the path to solving many extraordinary problems like [10] using Bi-LSTM based semi-supervised text classification. In [11], the authors proposed a heterogeneous graph attention network that can learn the importance of adjacent words treating them as attention-based neighboring graph nodes they used for classifying short texts on Twitter and other social media one such example of semi-supervised learning data scarcity problem is where the authors used Tmix an augmentation approach to generate a large amount of interpolated texts which made their trained model acquire better performance

than state-of-the-art approaches. In [12] the authors used GAN-BERT architecture which reduced the requirement of annotated samples to 50 to 100 as the scarcity of data. The authors of [13] came up with a cross-entropy training scheme using a supervised and unsupervised training approach which enhanced the performance of simple models on complex language tasks. Another similar approach to semi-supervised is the weakly supervised technique stated in [14] and [15] where they designed a weakly supervised training architecture using two modules one to generate training documents and the other is self retraining module which reduced the lacking of training data and classifying text according to their hierarchy which very simple supervision like partially labeling or keywords but the accuracy these weakly supervised models is very impressive.

2.2 Active Learning Based Approach

The addition of active learning in the knowledge mining task has significantly reduced human labor to mention in [16], the authors carried out the task of extracting information from legal license text which will allow identifying the text into predefined classes. In [17], they set up another system with an active learning mechanism using Multi-nominal Naive Bayes (MNB) in a one-vs-all setting with a separated feature selection technique for each MNB classifier. In [18] The active learning setup works through training a classifier on an annotated data set and applying it to an unannotated data set and instances consisting in the data set are ranked accordingly to the validity of the model for labeling then instances with the highest certainty are passed to the ORACLE for annotation hence the system is retrained with newly annotated enhanced examples eventually the active learning system outperformed the SVM system by acquiring 83% accuracy. Another such example is [19], where they proposed two distinctive approaches for relation extraction one from web documents and another from knowledge graphs. For relation extraction from web documents, the authors built a system with a human in the loop where the system generates a subset of candidates by calculating distances between entities and presents them to the human expert then the accepted candidates by the human expert are used to train the relation classification deep network by extracting the lexical and sentence level features achieving F-1 score of 0.0321. Also in [20], the authors demonstrated the advantages of using sequence representations constructed on word embeddings to appraise the seed selection and sample selection methods in an active learning channel for clinical knowledge extraction. They trained an initial model on a small labeled data set and used this seed selection technique to choose the most informative samples in the unlabeled data making the model better at every iteration which vastly reduces the effort of manual annotation, their proposed seed selection method gave a very satisfying performance acquiring F1-measure of 0.8309.

2.3 Deep Learning Based Approach

In [21] the authors presented 150 research articles findings based on deep learning and reviewed more than 40 datasets and some of their studied works are [] where the authors carried out multi and single-label Arabic text classification using deep learning-based architectures among them convolutional-GRU and attention-GRU achieved 91.18% and 96.94% accuracy on the target dataset and in [22] the

authors used BERT to automatically classify near-miss reports on construction site and tested with a collected near-miss database. In [23], the authors carried out clinical text classification using Active and deep learning techniques, and in [24] the authors implemented a triggering word malicious content detection system to prevent backdoor attacks using LSTM architecture. In [25], the authors used a graph convolutional network to learn co-occurrence and document word relations for text classification, and in [26] the authors adapted an ensemble-based text classification model using SVM, naïve Bayes, and random forest and a deep learning conventional network. The incredibly increasing amount of textual data paved the opportunity of using deep learning techniques in textual Natural Language processing so, below some of the honorable works are mentioned to start with Name entity recognition is one of the vital tasks in the field of Natural Language Processing. In [27], the authors carried out this work in three subtasks named entity recognition, entity classification, and relation extraction from medical texts. The authors used two distinctive deep networks one for name entity recognition and classification and another for relation extraction made of BI-LSTM and Conditional Random Field (CRF) layers. They tested their approach on two individual DDI corpus and the eHealth-KD dataset acquiring an F score of 87.19% and 87.2% in name entity recognition, 76.21%, and 85% in entity classification, 66.78%, and 54.23% in relation extraction respectively which outperformed previous works. In [28], the authors came up with an approach to extract knowledge from existing rule-based chatbots to build a neural network conversational agent-based Chabot as constructing a rule-based chatbot needs too much time and effort. They used online rule-based available chatbots conversation as their dataset and appointed 70% to the train set rest was given to the test set. They built the Neural Conversational Agent with Seq2Seq LSTM deep network and to evaluate the proposed system the Rule-Based and the NCA chatbot compared with the same question where BLEU score was used to measure the similarity which works by counting the similar number of n-grams according to which the NCA chatbot managed to extract 75% knowledge of the Rule-based chatbots. The authors of [29], proposed a set of enhancements to previously proposed methods of automatically identifying and extracting algorithmic pseudo-codes along with the sentences that convey algorithmic metadata. The authors presented 60 novel features consisting of content-based, font style-based, and structure-based which were used to train the deep network formed of a character convolutional neural network and two Bi-Directional LSTM architectures. The authors claimed their architecture achieved an F1-score of 93.32% outperforming all previous work by 28%. In [30], the authors used a complementary convolution neural network (com-CNN) with attentional multiple instance learning (MIL) which is a compound mixture of an embedding layer, two separate CNN layers, a fusion layer, attention module, and a fully connected layer to obtain highly comprehensive features for extracting domain relation from noisy Chinese texts. The authors mentioned their proposed architecture surpassed state-of-the-art approaches with an accuracy of 97.61% and 95.49 on two different datasets wiki and Baike. In [31], the authors used four distinctive deep network architectures to extract interaction between drugs from biomedical texts, where they categorized the interactions into five classes. The four deep architectures are CNN, B-LSTM, AB-LSTM, and joint AB-LSTM where, these are the F1-Score 63.40, 68.66, 67.43, and 69.39 accordingly. In [32], the authors proposed a method to automate the process of verifying banking transactions (e.g., money transfers, tax payments, salary payments, credit applications) which goes through several verifying processes and demands lots of human labor. After

extracting text from documents through OCR, they used character-level BiLSTM-CRF to detect the Name entity and afterward verified, the transaction through training a Fasttext model which, overall achieved an F1-score of 90.17 verified through human expertise.

2.4 Unsupervised and HMM Based Approach

In [33] the authors proposed a novel unsupervised text classification approach and discovered that the unsupervised approach is less driven by the selected algorithms than the supervised approach though the supervised approach performs slightly better than the unsupervised text mining approach. Also in [34], the authors used a bottom-up unsupervised technique to mine emotion-based social media content to extract user emotions about a specific topic. In [35], the authors carried out the task of mining health-related noisy texts using an unsupervised approach that targeted mining noisy spelled texts with closely matched fine texts. In [36] the authors mined user opinions using the hidden Markov model approach which mined texts based on the sequence of texts to learn textual patterns also in [37] the authors used HMM word sequence model to detect spam SMS. Apart from the methods discussed above, there are many other eminent techniques of text classification and concept extraction one such work is [38], and its authors introduced a new unsupervised system for extracting concepts from the clinical text where the system creates a representation of concepts by combining natural language descriptors with word representation composed into higher-order concept vectors which are responsible for assigning labels to candidate phrases which use syntactic chunker for extraction. Their system achieved an F-score of 0.32 which outperformed other unsupervised methods in the relative task and in [39], the authors conducted the task of complex terminology extraction from textual web data by integrating two modules one module is based on Hidden Markov Model (HMM) which combines linguistic and statistical knowledge for complex term extraction another module is based on Markov Chain for refining terms given by the HMM module. Their proposed approach was tested on four separate data sets with an F-measure score of 0.424, 0.263, 0.279, and 0.282.

2.5 Event and Knowledge Extraction from Text

In [40], the authors built a Named Entity Recognition and Relation Classification Module using two separate deep learning models to extract events from clinical textual data, and in [41] the authors explained various approaches to automatically constructing ontology in any kind of textual data. One such interesting work on trend analysis is [42] in which the authors collected blog online posts of a specific time series and used text mining clustering techniques to extract trendy keywords to analyze the fashion trend of that particular period on the other hand in [43] the authors conducted Spatio-temporal trend analysis on Twitter data to analyze the voter's trend for the upcoming election. in [44] the authors studied the trendy information on human resource management using text clustering techniques on the human resource management textual dataset. Various approaches methods and tools have been introduced for knowledge extraction from texts like in [45], the authors elaborated various types of event extraction methods like Data-driven event extraction which completely relies on quantitative methods to identify relations, knowledge-driven event extraction which works with

predefined knowledge pattern modules for event extraction and, Hybrid event extraction methods which work with combining pattern-based algorithms with bootstrapping or initial clustering for event detection in decision support applications like In [46], the authors used text exemplars that were generated by human annotators for assigning classes to unseen text from a specific domain ontology. overlapping n-grams and a bag of bi-characters were used as the prime feature for their work and compared their approach with human annotations through 5*5 cross-validation. Their Exemplar-based algorithm managed to reach pretty near the human-generated annotations. In [47], the OpenIE approach was used to generate a binary relation knowledge graph from texts as the OpenIE method is capable of dealing with complex sentences and their proposed approach surpassed the previous state-of-the-art methods with an F1-score of 0.8827. In [48], the authors proposed a new method and implemented their system in java to extract rules from scientific texts specifically gynecology-related cases using WordNet dictionary, Gate platform, ProtegeOWL API to work with the OWL ontologies and SWRL rules, and the Jena API with excellent performance F-measure of 95.83%. In [49] the authors analyzed offline-to-offline commerce texts using bilingual text mining between tweets of America and China to understand key entities and further understand relationships among those entities.

Methodology

The below figure-1 is portrayed the workflow diagram of our system where all of the methods and phases of our work are shown in a flowing sequence from data collection, preprocessing, and training classifier to analyzing different time-series data through a diverse number of tools and approaches. Therefore in this chapter, we would like to present the insight details of our work which will facilitate getting a clear understanding of our methodology. Below are given all of the phases of our adopted process in a sequence of their preface priority of commencement:

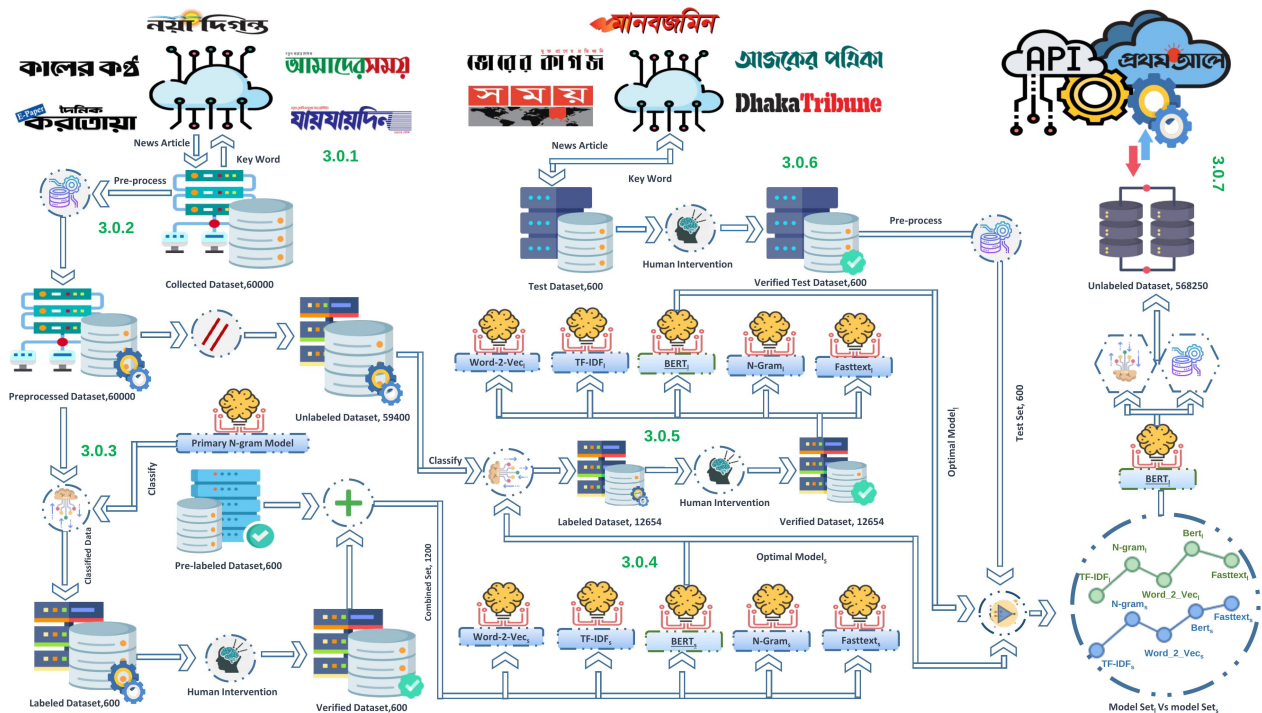


Figure 3.1: Full System Workflow Diagram.

3.1 Training Data Collection

During the initial phase of our work beforehand, we began with collecting textual news article data for the sole purpose of constructing the training set. As our source, we selected the online news platforms, [50] which are the online representative counterpart of the actual physical newspaper and consist of a massive collection of news articles of daily events occurring in both present and past periods. The

reason for selecting these specific newspapers is that these news portals maintain a large archive that has all the records of news articles from previous years which is not found in other online news portals. The news article data was collected with the web crawling technique using the library BeautifulSoup [51]. The procedure of fetching the target event embedded news articles works through a query along with a provided key (K_p) parameter which is a set of words related to the representation of the target events hence making a query in the news repository will draw out the news articles matched with the provided key (K_p) which were stored in the data set (D_i). For each event a unique key set was used acquired from [5] and to mention some of the online sources which were Kaler Kontho [52], JaiJai din [53], Amader shomoy [54], Ajker Patrika [55], Bhorer Kagoj [56], etc each has its online news portal besides actual publication. To ensure the uniqueness of the data set (D_i) we placed a constraint of partially matching the incoming article with already included data articles in D_i and the size of the data set made up to 60000.

3.2 Pre Processing Collected Data

After collecting data and building the primary training data set (D_i) it was essential that each of the textual news articles is pre-processed to enhance the performance of the predictive classifier for more precise classification. For the beginning, we started with tokenizing the text articles with the assistance of the BNLP (Bengali Natural Language Processing) [57], NLTK, [58] tool developed in python after that each of the tokens was inflicted with a sequence of validation phases like if any of the tokens is a superset of punctuation mark character set or stop-word else subset of irrelevant words all accumulated from [5] then the token will be elicited from the token list of the text article afterward if the token is passed to the next phase then the token is pre-processed with techniques like Stemming and Lemmatization using the `bangla_stemmer`, [59] library which is also developed in python. After successfully passing through all of the pre-processing phases the tokens of the updated token list were once again combined together to backtrack to the textual format and were stored in a new separate dataset D_p derived from the primary data set D_i which will be used for training the classifiers. The dataset was processed and stored in Dataframe provided by the Pandas library in Python.

3.3 Target Data points Extraction

Although data set D_p originated from D_i procured through the matching parameter key set K_p it doesn't guarantee the fact all of the collected data points are related to a violent event or strung any indication of an occurrence of the target events. The possibility of extracting this kind of data points is very high as many articles could include target word without having any base of occurrence so to deal with this uncertainty we applied the N-gram feature-based logistic regression predictive model which was trained in [5] to extract violence event occurrence related data through classification for each respective violence class with a uniform distribution note that the non-violence class wasn't considered in this trial as we have a separate collection of news article set which doesn't include any violence event from every different genre. After collecting 600 samples equally distributed respectively among the six violence event classes a new data set D_{pl} was formed having a size of 600 samples which

is a subset of the original data set D_p where $(D_{pl} \in D_p)$ and rest of the portion were kept in a new separate unlabeled data set of 59,400 samples D_{ul} which is also a subset of D_p where $(D_{ul} \in D_p)$.

3.4 Predictive Model Training Using Different Features

After partially labeling the dataset D_{pl} we conducted a verification study of the labeled data points to ensure the labeling authenticity and to avoid any kind of misconception in the posterior predictive models and the task was performed by going through the content of the textual news articles to verify if the model has done a rational job if not relabel the data sample with appropriate labels via human intervention. After conducting the verification study we achieved the creditable data set D_{fl} from D_{pl} where $(D_{fl} \supset D_{pl})$. For amplifying the volume of data we combined dataset D_{fl} with the primary data set D_{ps} of 600 samples constructed and labeled by human annotators in [5] which makes our new combined dataset D_{csd} volume 1200 excluding the non-violence related data points which will be added later maintaining uniformity of the dataset. The data points of the Dataset were balanced with the Undersampling technique concerning each of the individual classes to ignore any Biasness in the models. Afterward using this combined set D_{csd} we trained four distinctive predictive classifiers with four unique features from the set of { Word_2_Vec, TF-IDF, Fasttext, N-Gram, BERT } and the classifiers derived from this feature set is identified as $BERT_s$, $Tf-IDF_s$, $Word_2_Vec_s$, $N\text{-gram}_s$ and $Fasttext_s$ in Figure-1 system workflow diagram. The { Word_2_Vec, TF-IDF, N-Gram } features were used to train several classifiers with different Algorithms provided in the scikit-learn library, [60] in all of the experimental studies of the thesis. The list of the selected algorithms is as follows Logistic Regression, Decision Tree, Random Forest, SVM, and Naive Bayes. But in all of the experiments, the highest performance metric was achieved by Random forest hence Random Forest was the utmost competitor with the other two classification approach which are BERT, [61] and Fasttext, [62].

3.5 Large Data set Construction & Experimenting With different features

After experimenting with the small data set D_{csd} we wanted to utilize the unlabeled data set D_{ul} as this set comprises a large portion of the whole collected data to do so we adopted a similar approach as discussed above where we used the $BERT_s$ predictive model derived from the small data set D_{csd} as it had to most optimum performance. Initially, all of the data samples were classified with $BERT_s$ where 12,654 data samples denoted violence-related events, and the rest of the 46,746 samples were dumped as it would not be a fair contribution to the predictive model training task. Through human intervention, these violent event-related 12,654 samples were put through the verification test for eradicating any uncertainty from the classified data and were eventually used to construct the new larger volume dataset D_{cld} where $(D_{cld} \in D_{ul})$. Just like the previous experiment, the data points of the Dataset were balanced with the Undersampling technique and we used this verified large dataset D_{cld} to train new predictive classifiers with the same feature set of { Word_2_Vec, TF-IDF, Fasttext, N-Gram, BERT } and derived four new predictive classifiers $BERT_l$, $Tf-IDF_l$, $Word_2_Vec_l$, $N\text{-gram}_l$ and $Fasttext_l$ where BERT again proved to be the most prominent feature.

3.6 Test Data Collection & Comparing Model of different Volumes

After training completion of the predictive models in both large and small volume data sets, it was required to test the predictive model on unseen data to evaluate the performance of the classifiers consisting of both sets. To do so we needed to collect and build a testing dataset D_t where ($D_t \notin D_p$) which means it should have no similarity with the original training dataset D_p . In order to maintain this constraint, we collected data from different online sources other than the training data collection source which are Shomoy news, Manob jomin, Ajker potrika, Bhorer kagoj, Dhaka Tribune, etc. These data samples were also collected through the previously used key parameter K_p and also pre-processed following the same steps mentioned before and were verified through human intervention. The volume of the test set D_t is 1000 samples in total where each unique class label has almost the same amount of balanced samples. Now, coming to the observation we discovered that the classifier set of the large volume $\{ \text{BERT}_l, \text{Tf-IDF}_l, \text{Word_2_Vec}_l, \text{N-gram}_l \text{ and } \text{Fasttext}_l \}$ of all unique features achieved higher accuracy than the small volume classifier set $\{ \text{BERT}_s, \text{Tf-IDF}_s, \text{Word_2_Vec}_s, \text{N-gram}_s \text{ and } \text{Fasttext}_s \}$ seemingly with the addition of more data exalted the performance accuracy of the classifiers of same feature extraction techniques.

3.7 Fetching Time Series Data & Mining Meaningful Knowledge

The conducted experiment done on the comparison of the two classifier sets disclosed evident reasons for selecting the most optimal predictive classifier which is the BERT_l derived from the large volume dataset D_{old} which we will utilize in our further experimental studies. For the next experiment, we proposed to investigate the data of different timelines to observe what kind of knowledge pattern we can extract using the BERT_l classifier. We started with collecting textual news data from different years and for the source, we targeted the Prothom Alo [63] online news archive as they amass the most massive amount of textual news data from the present day to the previous decade. Prothom Alo [63] online news platform provides an open-source API for fetching data from their repository hence we used the API link and extracted a voluminous amount of data over several years by developing a python API interactor script. Each of the fetched data objects contained the following attributes which are the Date of the article, Title of the article, Author of the article, Section whether foreign or local news article, Related tags to the article, Main content of the article, and id of the news article. The order of acknowledging of the articles was random so we had to categorize these data objects concerning their occurring year, month, and day to transparently observe the sequential trend of the flowing time series. Insights discovered from this time series study will be presented in the upcoming sections of the article.

Chapter 4

Performance Analysis

In this section, we want to shed light on the performance of the classifiers trained on data sets of different volumes and elect the optimal model for extracting knowledge from the time series data. The model derived in [5] had an accuracy of 83.4% which was a Logistic Regression Based classifier and it is the primary model exploited to build the small volume data set. For investigation, we evaluated this primary logistic regression model with the recently constructed test data set (D_t) of 1000 samples, where the model managed to acquire 61% accuracy. The most preeminent reason for the model's performance downfall is that in this work we defined the problem as a multi-classification problem as many of the news articles might contain several related violent events occurrence like rape with murder and kidnap with rape etc. In [5] the model was trained with respect to one violent event per article format which was a major drawback of the proposed method so in this work we trained the classifiers considering the existence of multi-class. The evaluation metric used in the experiment is accuracy meaning the total number of correctly classified violent events divided by the total volume of the dataset. Below this section, the detailed performance analysis of the predictive classifiers is illustrated:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.1 Performance Evaluation of the Small Volume Classifiers

After construction of the small volume dataset, (D_{csd}) of 1200 data samples consisting of six different violent events afterward 200 non-violence news article data samples were injected into the set. $BERT_s$, $Tf-IDF_s$, $Word_2_Vec_s$, $N\text{-}gram_s$ and $Fasttext_s$ are the five different predictive classifiers that are trained on the five diverse features BERT, TF-IDF, Word_2_Vec, N-Gram and Fasttext respectively on the final composition of these data samples. The performance of these classifiers are given in the table below in both the train and test set:

Below is a visualization graph of the different feature-based classifiers in both the test and train set derived from the small volume dataset. The horizontal coordinate denotes the performance accuracy band along with different feature classifiers which are placed in the vertical coordinate and the performance results of the training and test set are separated in two distinctive colors placed in each of the respective classifier's coordinates.

Table 4.1: Performance Comparison of different feature classifiers using small volume data set on train set and test set.

Base Feature Classifier	Performance Accuracy	
	Train Set	Test Set
Word_2_Vec _s	68%	60%
Tf-IDF _s	79%	74%
Fasttext _s	87%	82%
N-gram _s	95%	86%
BERT _s	95%	92%

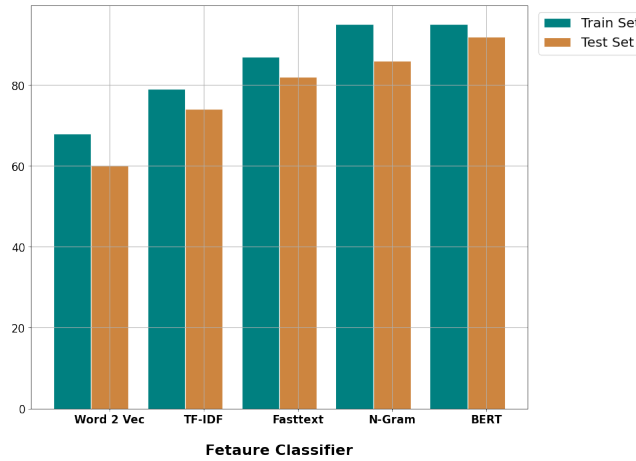


Figure 4.1: Performance Comparison of Different Feature-Based Classifiers on Small Volume Set.

From the above graph, we can see that the BERT and N-gram-based classifiers acquired the best performance result but N-gram performance was poor in the test set compared to the BERT-based classifier rest of the performance of the feature-based classifiers are in the following order Fasttext, TF-IDF, Word2Vec.

4.2 Performance Evaluation of the Large Volume Classifiers

Similar to the small volume data set we injected 2000 non-violence data samples in the large volume data set (D_{cld}) of 12,654 samples and using the same unique feature set { Word_2_Vec, TF-IDF, Fasttext, N-Gram, BERT } produced five different predictive classifiers Word_2_Vec_{*l*}, Tf-IDF_{*l*}, Fasttext_{*l*}, N-gram_{*l*} and BERT_{*l*} respectively where the predictive classifier BERT_{*l*} performed prominently in both train and test set (D_t):

Below figure-3 is a performance visualization graph of the different feature-based classifiers in both the test and train set which are derived from the large volume training dataset. In this figure, also the horizontal coordinate denotes the performance accuracy band along with different feature classifiers which are placed in the vertical coordinate and the performance results of the training and test set are separated in two distinctive colors placed in each of the respective classifiers coordinates just like the small volume dataset classifiers comparison graph.

Table 4.2: Performance Comparison of different feature classifiers using Large volume data set on train set and test set.

Base Feature Classifier	Performance Accuracy	
	Train Set	Test Set
Word_2_Vec _l	77%	72%
Tf-IDF _l	84%	80%
Fasttext _l	96%	91%
N-gram _l	99%	92%
BERT _l	98%	95%

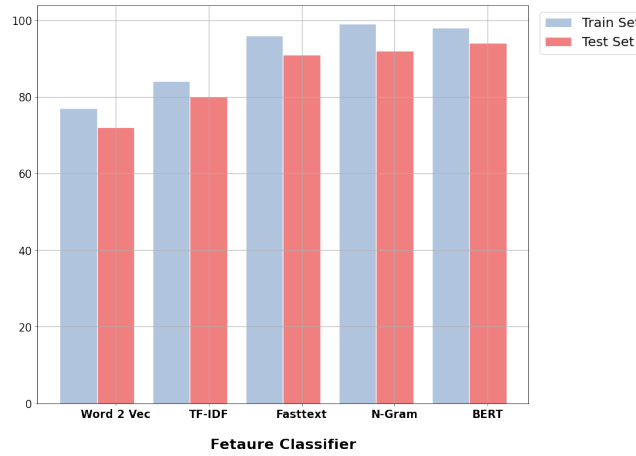


Figure 4.2: Performance Comparison of Different Feature-Based Classifiers on Large Volume Set.

In the large volume dataset trial, the BERT and N-gram-based classifiers managed to preserve their top position just like the small volume dataset experiment, and the rest of the classifier's performance order is exactly the same as the prior small volume results. But each of the distinctive feature-based classifiers trained on the large volume dataset had a drastic increase in their respective performance accuracy than the small volume trained classifiers.

4.3 Performance Comparison between the Large and Small Volume Classifiers

After successfully constructing the predictive classifiers from both the small volume D_{csd} and large volume dataset D_{cld} we compared the classifiers based on their performance accuracy. At first, each of the predictive classifiers was compared on the performance acquired on the respective training set of both volumes where D_{cld} contains all of the data points of D_{csd} which means $(D_{csd} \subset D_{cld})$. All the models derived from the large volume classifier D_{cld} performed better than the classifiers of the small volume set D_{csd} where BERT_l acquired the maximum accuracy among the other contemporary models shown in figure-4.

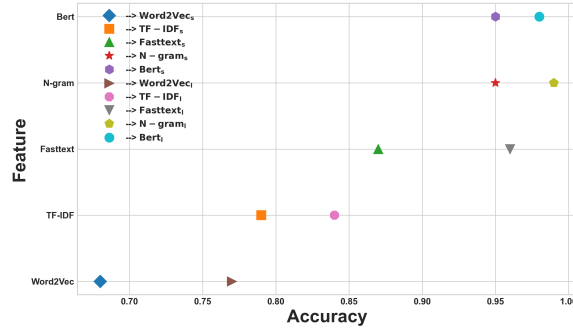


Figure 4.3: Performance Comparison of Small and Large Volume Set Classifiers Based on the Trained Set.

Below is the performance comparison graph of the classifiers of the small & large volume set where each of the predictive classifiers were judged concerning the constructed test set D_t where $(D_t \not\subseteq D_{cld})$ & $(D_t \not\subseteq D_{csd})$. Just like the training set experiment all of the classifiers maintained their prior respective rank and BERT_{*l*} managed to cut the highest amount of accuracy shown in figure-5.

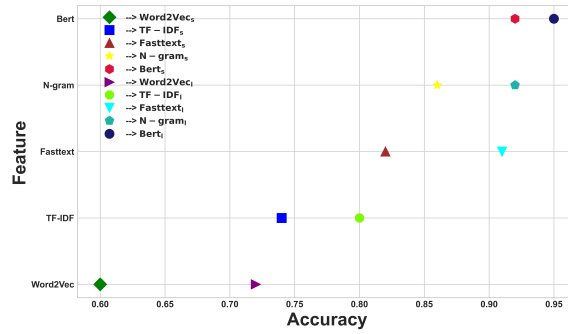


Figure 4.4: Performance Comparison of Small and Large Volume Set Classifiers Based on the Test Set.

Based on the results of the trained and test set experiment BERT_{*l*} is the elected model for mining patterns and extracting trends from the high volume unlabeled data of the last demi-decade.

4.4 Pattern and Trend Analysis

In this section, we are going to represent the trend and analytical experiment result on the large volume unlabeled news article dataset. We collected the unlabeled data from Prothom Alo, [63] online news archive via the free API portal provided by Prothom Alo, [63] and each of the fetched data objects contained the following attributes which are the Date of the article, Title of the article, Author of the article, Section whether foreign or local news article, Related tags to the article, Main content of the article and id of the news article also mentioned in the previous section 3.7. The unlabeled dataset contains all the produced news articles by Prothom Alo, [63] online news platform for the following 5 years 2014, 2015, 2016, 2017 & 2018. As shown in the previous experiment result of 4.3 BERT_{*l*} derived from the dataset D_{cld} achieved the highest performance accuracy to rely on and the experimental study shown in this section is all instituted upon this model.

4.4.1 Event frequency Histogram Analysis

Below we can see the event frequency histogram which was derived from the D_{cld} dataset. So, first, we split all the year's data by months and applied the $BERT_l$ model on all the chunks of data. Hence we stored the number of all the unique events that were found in the monthly chunks of data which we used to plot the histogram.

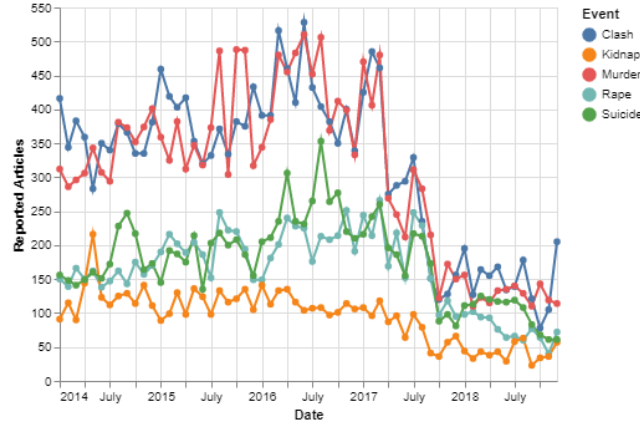


Figure 4.5: All Events Frequency Histogram.

From the histogram above we can see the most frequent violent event is the Clash event which has the highest peak among all the other events. Analyzing the news article's political and religious outrage is the prime triggering factor of these events. In the second position is the Murder event which is pretty close to the Clash events frequency and the rest of the frequency of the events are in this respective order Suicide, Rape, and Kidnapping.

4.4.1.1 Hypothesis Test - Most Frequently Occurring Violent Event

From the frequency histogram, we see that the highest occurring events are Clash and Murder which are seen throughout the demi-decade data histogram, and the rest of the three events Rape, Kidnap Suicide histogram falls below the histogram of Murder and Clash event which makes this three events the least occurring event in the demi-decade news article data. So from the described fact, we can claim the hypothesis that Murder Clash is the most occurring violent event. To analyze the stated hypothesis we conducted the Wilcoxon Signed-Rank Test on the frequency of the event calculated from the same time intervals. The formula is given below:

$$W = \sum_{i=1}^{N_r} [\text{sgn}(x_{2,i} - x_{1,i}) * R_i]$$

where W , is the test statistic, N_r is the sample size, (sgn) is the sign function, $x_{2,i}$, $x_{1,i}$ are the two distributions for ranked pairs and R_i is the rank of i . In our sample, we calculated the frequency of different events within the same timeline and generated samples of frequency for each of the different events.

$$W_p = \begin{cases} \prod_i \prod_j \sum_{t=1}^{N_t} [\text{sgn}(E_{if,t} - E_{jf,t}) * R_t] \\ \text{where } j \in \{ \text{Suicide, Rape, Kidnap} \} \\ \text{where } i \in \{ \text{Murder, Clash} \} \\ \text{and } t \in \{1,5,10,20,49,79...\} \end{cases} \quad (4.1)$$

Based on the above equation these are the following hypothesis that can be claimed from the results seen in fig-6. The null hypothesis claims that Murder and Clash are the most frequently occurring violent event on the other hand the alternative hypothesis states that Murder and Clash are not the most frequently occurring violent event compared to Suicide, Kidnap, and Rape.

H_0 : Murder and Clash Event is the most highest occurring event than Kidnap, Rape & Suicide.

H_1 : Murder and Clash Event is not the most highest occurring event than Kidnap, Rape & Suicide.

Table 4.3: P_{value} of Wilcoxon Signed-Rank Test of Different Events.

Event	Clash		Murder	
	statistic	pvalue	statistic	pvalue
Murder	0.16012	0.872780	–	–
Clash	–	–	0.16012815	0.87278
Suicide	1.44115	0.149541	1.6012815	0.109314
Rape	1.601281	0.1093145	1.60128	0.1093145
Kidnap	2.401922	0.01630	2.401922	0.016309

Analyzing the results from the P_{value} Table we can see that the P_{value} between all the other events with Murder and Clash event is less than the threshold ($P_{value} < 0.05$) hence the population medians are significantly different from each other but on compared to each other Murder and Clash events P_{value} is greater than the threshold ($P_{MCvalue} > 0.05$) so we can reject the alternative hypothesis and accept the null hypothesis that is Murder and Clash events are the most frequently occurring violent events.

4.4.2 Yearly Violent and Non-Violent Event Article Comparison

In this experiment, we wanted to find which year among the demi-decade has the record of the highest number of violent events hence we split the demi-decade unlabelled data by year and classified each of the articles with the $BERT_l$ predictive model. In this case, we summed up the number of articles that were classified as violent consisting of event articles irrespective of their class value and on the other hand the non-violent articles were summed up.

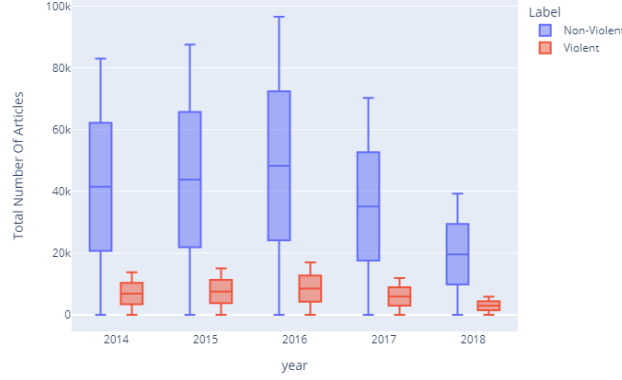


Figure 4.6: Yearly Violent & Non-Violent Event Article Comparison.

From the figure above we can see that the highest number of non-violent and violent event articles was found to be recorded in 2016 also it has the highest number of news articles apart from the other year's data chunks. another observation we can see the percentage of violent articles decreasing from 2016 to 2018 also the number of total articles has decreased and the same pattern is observed if we move left from 2016 to 2014.

4.4.2.1 Hypothesis Test - Most Event Impact full Year

To verify what is depicted in fig-7 we conducted a Wilcoxon two-tail test among the five years of data. Initially, we divided all the data samples by their reporting year and again separated each of the chunks according to their predicted label which we considered non-violent and violent if any of the violent categories is predicted. Hence we got the ratio of violence and non-violence ratio of each year of the demi-decade. Among the non-violent violent our target is the violent samples. Forming the data distribution we would like to state the following hypothesis:

$$W_p = \begin{cases} \prod_{ty} \prod_{sy} [\text{sgn}(V_f, ty - V_f, sy) * R_{sy}] \\ \text{where, } ty \rightarrow 2016 \\ \text{where } sy \in \{ 2014, 2015, 2017, 2018 \} \end{cases} \quad (4.2)$$

H₀: 2016 doesn't contain the highest violence article reports.

H₁: 2016 has the highest violence article reports.

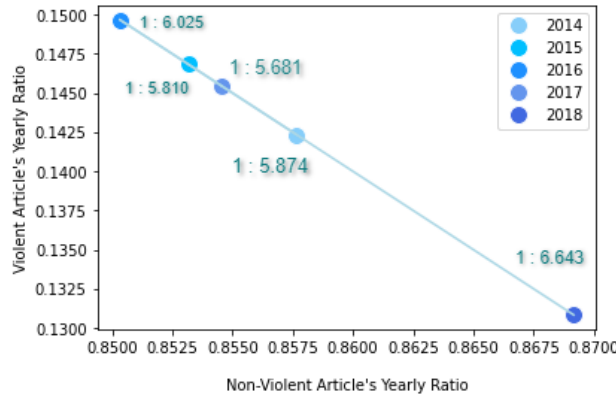
As we can see from the table each of the year's P-values is less than the threshold of 0.05 ($P_{value} < 0.05$), so so we have sufficient evidence to reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1) which clearly denotes that the population medians of 2016 and the other years are significantly different from each other.

Table 4.4: Two-Tail Wilcoxon P-Value table of Different Years Violence Record Frequency.

Years	Year 2016	
	statistic	pvalue
2014	3.752776	0.000174
2015	2.5403411	0.0110744
2017	2.713546	0.0066567
2018	2.9403411	0.0009765
2019	2.3403411	0.00048828

4.4.3 Yearly Violent and Non-Violent Event Article's Occurrence Ratio Comparison

This experiment is related to the above-shown experiment where we compared the number of violent non-violent articles numbers of every year from the demi-decade news article set. In this section, we calculated the Non-Violent: Violent article frequency ratio for all the five-year data chunks.

**Figure 4.7:** Yearly Violent and Non-Violent Event Article's Occurrence Ratio Comparison.

From the above ratio plot, we can see that almost all the year's Non-Violent: Violent ratio value is almost equivalent to 1:6. Here we found a similar pattern in all of the year's ratio which is very similar to each other($R \equiv 6.0$). We wonder if the news article publishing organization is balancing the number of violent occurring news articles in everyday newspaper articles from all the distinctive categories.

4.4.4 Violent Event Occurrence Comparison Over the Demi-Decade

In this experiment, we have compiled all the detected violent event articles in the D_{cld} dataset and summed up the total number of violent event articles with respect to their class label value. Below we can see a violent event class number frequency plot and a total number of reported cases comparison over the demi-decade's news article's based upon the summed-up value of all the individual violent class labels.

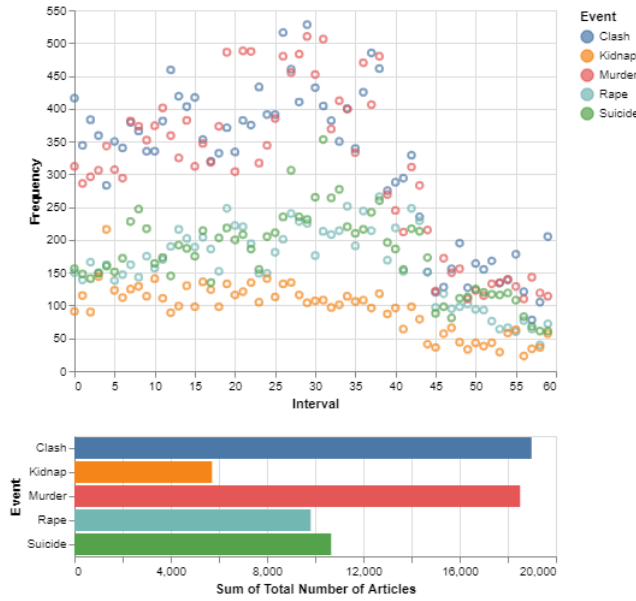


Figure 4.8: Violent Event Occurrence Comparison Over the Demi-Decade.

From the results we can clearly see that the most reported type of violent news article is the Clash Violent event in the last demi-decade and following this in the second position is the Murder violent event case which is the most dominant group after the Clash event. The other events have article numbers in the following descending order Suicide event, Rape event, and lastly the Kidnap event from the last demi-decade.

4.4.5 Probability Distribution of Events on Different Day Intervals

The following experiment was executed to analyze the probability distribution of different events at different day intervals. From our earlier study, we found the mostly occurring violent events over the demi-decade's news article data but in this experiment, we want to find out which violent event has the highest probability of occurring over different day intervals.

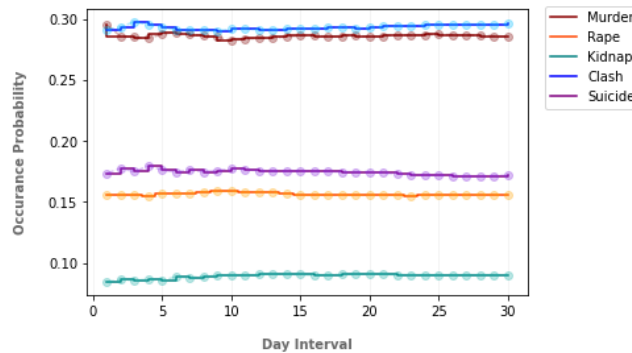


Figure 4.9: Probability Distribution of Events on Different Day Interval.

So in the above figure, we can see the probability distribution of the five different events on increasing day intervals ranging from 1 to 30. From the result, it is visible that the Clash event has the highest probability of occurrence and it is the mostly found event in the whole unlabeled dataset in the second place having the highest occurring probability is the Murder event and the rest of the

order of the events are as exactly as the highest number of occurrence report which is Suicide, Rape and the Kidnap event.

4.4.6 Area Histogram of Different Events Over the Demi-Decade

In this experiment, we separated the labeled data which is derived from classifying the unlabeled demi-decade news article set by the different types of events. So the separate set would contain all the occurring dates counted for each violent event. With this newly labeled dataset, we plotted an area histogram for all of the events which we differentiated by distinctive colors in the figure below. The area histogram is within the (2014-2018) yearly boundary.

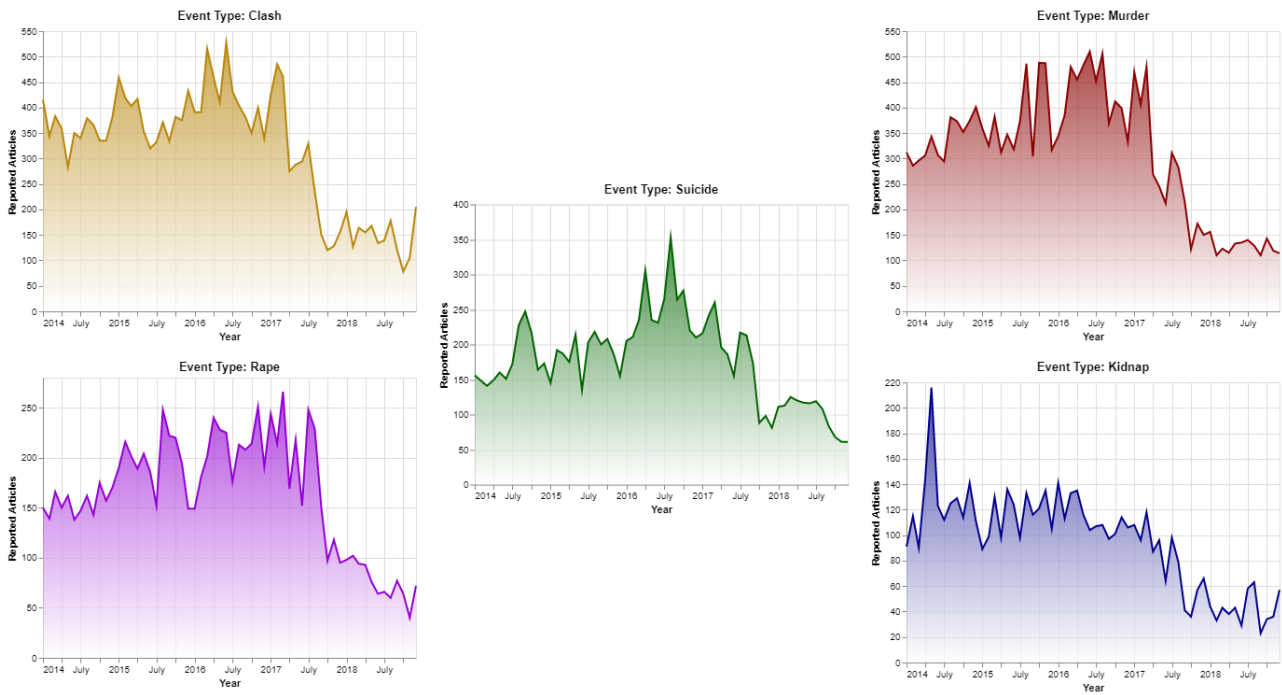


Figure 4.10: Area Histogram of Different Events.

If we analyze each of the area histograms of the respective events we can see that the Clash event's global maximum point of occurrence report is near July (2016) and the global minimum point of occurrence report is near July (2018) the occurrence rate is in decreasing order since 2014. The Murder event's global maximum occurrence point is also near July (2016) and the minimum point is after August (2018) in decreasing order. The Rape event's global maximum occurrence point is near June (2017) and the minimum point is after August (2018) in decreasing order. The Suicide event's global maximum occurrence point is near August (2016) and the minimum point is after September (2018) in decreasing order. The Kidnap event's global maximum occurrence point is near July (2014) and the minimum point is between July-August (2018) in decreasing order. All of the event's occurrence rates have shown to be decreasing since 2014 but in the midpoint, it has increased peak points.

4.4.7 Violent Event Dominance Comparison Over Different Months

This experiment focuses on the number of total violent cases reported in each of the separate months of the year for this study, we have considered all of the five years of data of the unlabeled D_{old} dataset. So at first, we separated the news article data by the reporting year after that we further split the news article data by reporting months and classified each article as every month of the respective year. Below we can see a time series area graph for each of the years every month shown in numerical order 1 to 12.

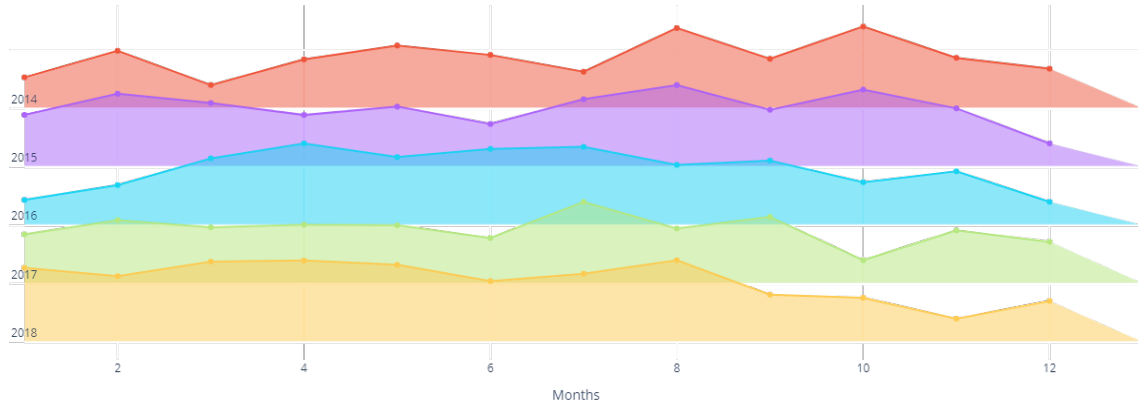


Figure 4.11: Violent Event Dominance Comparison Over Different Months.

Now, if we analyze the result for each of the years we can observe that in the year 2014, we have the two maximum violent case-reporting months that is August and October, on the other hand, the lowest number of cases reported in the month of March. For 2015 the maximum number of case reporting months remain the same in August and October but the lowest number of cases was found in December. For 2016 maximum point is April and we have two of the most minimum points month which are January and December. Lastly, for the years 2016 and 2017, the maximum and the minimum number of reported cases months are the following (May - October) (June - November).

4.4.7.1 Hypothesis Test - Half Yearly Period Event Ratio

From the above figure, we could see some of the pick values maintaining a certain side of the area graph which lead us to test if there is any specific part of the year that contains more violent article reports than the other one. Hence, we divided the dataset based on their reporting year and again separated the data dividing the year into two better halves each containing the data samples of six distinctive months. As our target samples are the violent ones so filtered out the non-violent samples from each half of every pair of the year. Using these data pairs for each year we would like to validate the under-stated hypothesis:

$$W_p = \begin{cases} \sum_{y=2014}^{2018} \prod_{p1} \prod_{p2} [\text{sgn}(V_f, p1 - V_f, p2) * R_y] \\ \text{where } p1 \in \{ \text{January : June} \} \\ \text{where } p1 \in \{ \text{July : December} \} \end{cases} \quad (4.3)$$

H_0 : There is no difference in violence records between the half-yearly period of the year.

H_1 : There is a difference in violence records between the half-yearly period of the year.

From the Wilcoxon statistical test, we got the P_{val} of 0.3125 which is less than the threshold ($\alpha = 0.5$) hence ($P_{val} < \alpha$) means the half-yearly violence ratio samples is of the same distribution and the distributions are equal if in the case if the ($P_{val} > \alpha$) than it would be opposite. So, the null hypothesis (H_0) is true and the alternative hypothesis (H_1) is insignificant.

4.4.8 Active Age Group Analysis in Violent Events

In this experiment, we wanted to excavate the age group that is related to this kind of violent event. For this, we have uniformly distributed the age group by a window of 20 which will be the range of each of the age groups. The news articles that we are dealing with mention a person in the textual content in a certain pattern which is mentioning the name of that person and the age of that person is put beside the name in brackets regardless of the age. Therefore, after classifying all the news articles in the unlabeled dataset we picked out the violent detected articles and applied Bangla Name Entity Recognition tool on the selected article to detect the mentioned person's name position so we can extract the age of that person using the B-NER python tool.

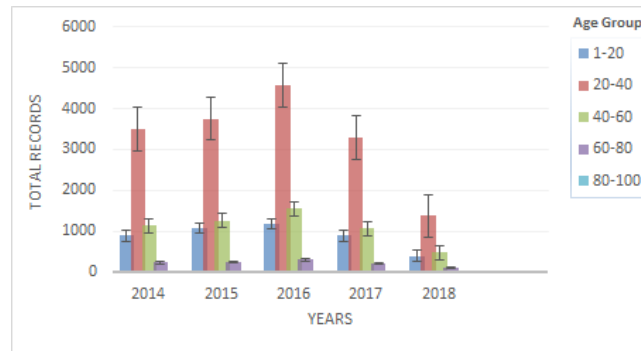


Figure 4.12: Active Age Group Analysis in Violent Events.

After detecting any person's name we stored the age number in a list before we proceed it is worth mentioning that we have generalized all the violent events as one so there is no distinct set based on the event type and as both victim and culprit are mentioned we have considered both and haven't used any filter to separate the person objects. As we can see in the above figure that in all of the year's news article data, the most active age group related to violent events is the (20-40) age group after that the most active group is (the 40-60) age on third the teenage group (1-20) is the most active. Although the (60-80) is slightly noticeable the (80-100) has a very insignificant presence possible reason could be the higher age and lacking people belonging to this age group.

4.4.8.1 Hypothesis Test - Event Associated Most Active Age Group

Above from the violent active age group comparison fig-13, we can see that (the 20-40) age group people are mostly related to violent events. To prove this assumption we collected violence frequency

distribution of these age groups over different periods all over the demi-decade dataset. Hence established the below-stated hypothesis to test our claim by applying a two-tailed Wilcoxon statistical test on the constructed distribution dataset:

$$W_p = \begin{cases} \sum_t \prod_{tg} \prod_{sg} [\text{sgn}(V_f, tg_i - V_f, sg_i) * R_i] \\ \text{where, } tg \rightarrow \{20 - 40\} \\ \text{where } sg \in \{ \{1-20\}, \{40-60\}, \{60-80\}, \{80-100\} \} \\ \text{where, } i \rightarrow \{1, 20, 45, 67, 17, \dots\} \end{cases} \quad (4.4)$$

H_0 : The age group 20-40 is not the highest violent active group.

H_1 : The age group 20-40 is the highest violent active group.

Table 4.5: Two-Tail Wilcoxon test P-value comparison table of different Age groups.

Age Groups	Age Group: 20-40	
	statistic	pvalue
1-20	4.084563445043468	4.41597598995539e-05
40-60	3.624711931363078	0.00028928371617217385
60-80	5.38296771896457	7.32676819947719e-08
80-100	5.410017808004594	6.301848221392269e-08

The P_{val} we got in all of the cases shown in the table is less than 0.5 ($P_{val} < \alpha$). That means that the distribution of the age group (20-40) is different compared to other contemporary age groups. Therefore we reject the null hypothesis (H_0) and accept the alternative hypothesis (H_1).

4.4.9 Monthly Heat-Map of Violent Events Over the Years

In this section, we plotted a heat map for all of the target violent events over the five years of news data. This heat map is the representation of the number of cases found over a month's news article data related to each of the violent events respectively. From this, we can identify the most dominant months of the violent event occurring within. If we look at the figure we can see six distinct heat maps for each of the violent clashes in the y-axis the years are placed and the x-axis is referring to the 12 months of the year.

To list some of the most highly occurring density-emitting months of the respective violent events starting with the event of Murder we can see that August, March, October, and November are the most heated months for this event if we average the value of all five years and most heated year is 2016. For the Adult Suicide event April, August, September, March, and October are the most heated and in this case, the year remains the same which is 2016. In the Clash event January, February, March, May, and June are the top event record-holder months and in the Rape event August, November, May, March, and April whereas for the Kidnap event May, March, April, and January is the highly heated months having reported for this kind of news articles. In terms of years, Rape seems to be a very prominent event in the years 2014,2015 2016 on the other hand Kidnap is very noticeable in 2014.

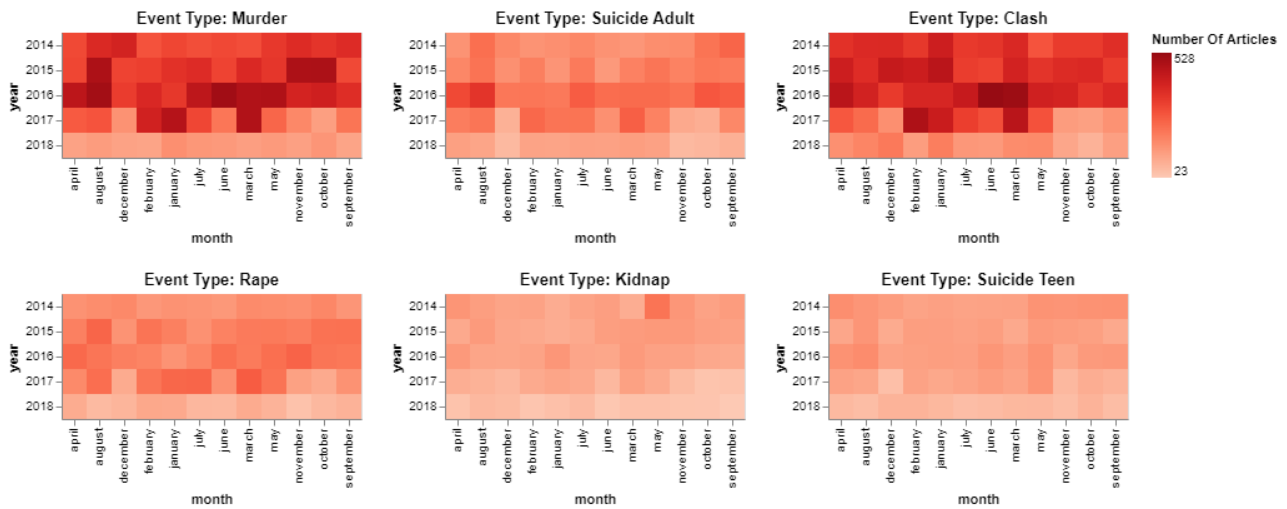


Figure 4.13: Monthly Heat-Map of Violent Events Over the Years.

4.4.10 Weekly Heat-Map of Violent Events Over the Years

In this section, we plotted a heat map for all of the target violent events over the five years of news data. This heat map is the representation of the number of cases found over a month's news article data related to each of the violent events respectively. From this, we can identify the most dominant months of the violent event occurring within. If we look at the figure we can see six distinct heat maps for each of the violent classes in the y-axis the years are placed and the x-axis is referring to the 12 months of the year.

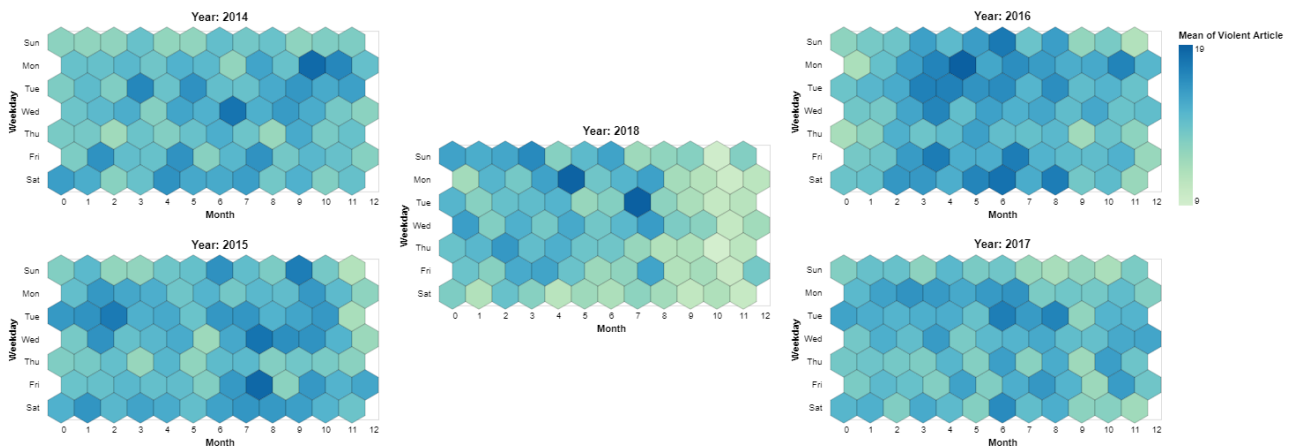


Figure 4.14: Weekly Heat-Map of Violent Events Over the Years.

After detecting any person's name we stored the age number in a list before we proceed it is worth mentioning that we have generalized all the violent events as one therefore there is no distinct set based on the event type and as both victim and culprit are mentioned we have considered both and haven't used any filter to separate the person objects. As we can see in the above figure that in all of the year's news article data, the most active age group related to violent events is the (20-40) age group after that the most active group is (the 40-60) age on third the teenage group (1-20) is the most active. Although the (60-80) is slightly noticeable the (80-100) has a very insignificant presence possible reason could be the higher age and lacking people belonging to this age group.

4.4.10.1 Hypothesis Test - Event Distribution Ratio on Different Weekdays

Analyzing the results of the violence intensity study of different weekdays we would like to analyze further to discover any insight patterns from the weekly violence data tuples. To discuss some background on the hypothesis we would like to claim there is no significant difference between the different days of the week except in Bangladesh we have two weekly holidays which are Friday Saturday among the 7 days of the week. So we would like to compare the violence occurrence frequency of the working days and the public holiday of the week. We resampled the dataset based on their reporting weekday and calculated the violence frequency of the weekdays on all the months of the demi-decade and compared the distribution of the weekday with the Wilcoxon signed rank test:

$$W_p = \begin{cases} \sum_i^y \sum_j^m \prod_{hd} \prod_{wd} [\text{sgn}(V_f, hd_{ji} - V_f, wd_{ji}) * R_{ji}] \\ \text{where, } hd \rightarrow \{ \text{Friday, Saturday} \} \\ \text{where } sg \in \{ \text{Sunday, Monday, Tuesday, Wednesday, Thursday,} \} \\ \text{where, } j \rightarrow \{ \text{January : December} \} \\ \text{where, } i \rightarrow \{ 2014 : 2018 \} \end{cases} \quad (4.5)$$

H_0 : There is no difference between the violence frequency occurrences of the weekdays and the holidays.

H_1 : The violent frequency occurrences of the weekdays and the holidays are not the same.

Table 4.6: Events Wilcoxon Signed-Rank Test Pvalues Comparison Table of Weekdays.

Day	Friday		Saturday	
	statistic	pvalue	statistic	pvalue
Saturday	7.0	0.5625	—	—
Sunday	0.0	0.03125	3.0	0.15625
Monday	0.0	0.03125	1.0	0.0625
Tuesday	0.0	0.03125	0.0	0.03125
Wednesday	0.0	0.03125	1.0	0.0625
Thursday	0.0	0.03125	3.0	0.15625
Friday	—	—	7.0	0.5625

From the Wilcoxon test, we get a P_{val} that is less than the threshold ($P_{val} < \alpha = 0.05$) which denotes Both Saturday and Friday have the same distributions with other days like Sunday, Monday, Tuesday, Wednesday, Thursday. But when compared to each other the P_{val} points that Friday and Saturday are in different distribution with each other. Hence we reject the alternative hypothesis (H_1) and the null hypothesis (H_0) remains.

4.4.11 Geographical Area Analysis Based on Violence Intensity

In the following experiment, we have shown the rate of violent events in all of the geographical level of areas in Bangladesh. There is a total of 8 divisions in Bangladesh and within every division, there are

multiple districts moreover within districts there are several towns. So from the violent news article data samples, we extracted the places of the incident occurred by a simple rule-based approach as we can see from the data sample that the news starts with mentioning the place name of occurred, and it can also appear at the end of the article after describing the whole ominous incident.

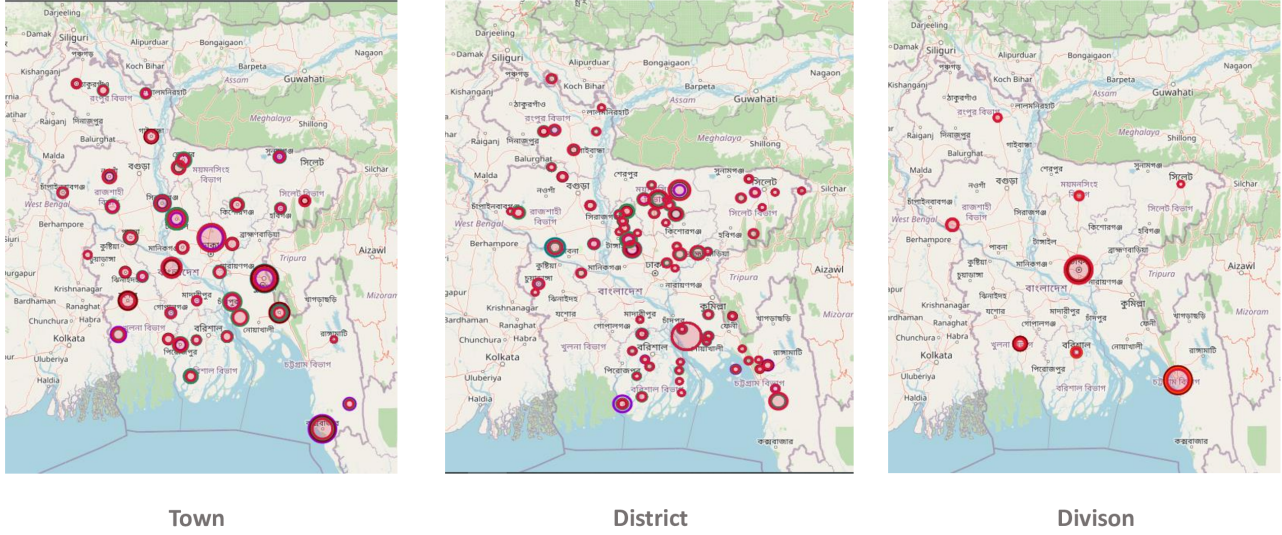


Figure 4.15: Geographical Area Analysis Based on Violence Intensity.

So in the above figure, we can see three composite figures that were plotted based on the geographical level of the incident place so we could identify the most impacted places in all levels. The bubble we can see above the map is based on the occurrence frequency ratio of the violent events in that area so the highest the ratio is the bigger the diameter of the circle gets. We used 6 distinctive color bands for the unique violent events and also the bubble was filled with that color. We can identify the most impacted places by these violent events from the bubble area map graph on all geographical levels of towns, districts, and Divisions.

4.4.11.1 Hypothesis Test - On Divisions Violence Event Ratio

From the above fig-16, we can see that the division Dhaka and Chittagong has circles with bigger diameters hence pointing out to be a more violent occurring state compared to the other six divisions. Eventually, Dhaka and Chittagong is the most developed division among the 8 divisions To test the claim we collected the frequency of violent events of all these divisions on 12 different months of the existing 5 years of the demi-decade dataset. Hence compared the frequency distribution of all these divisions with each other using the Wilcoxon signed rank test:

$$W_p = \begin{cases} \sum_i^y \sum_j^m \prod_{td} \prod_{sd} [sgn(V_f, hd_{ji} - V_f, wd_{ji}) * R_{ji}] \\ \text{where, } td \rightarrow \{ Dhaka, Chittagong \} \\ \text{where } sg \in \{ Rangpur, Sylhet, Barisal, Khulna, Mymensingh, Rajshahi \} \\ \text{where, } j \rightarrow \{ January : Decmber \} \\ \text{where, } i \rightarrow \{ 2014 : 2018 \} \end{cases} \quad (4.6)$$

H_0 : Dhaka and Chittagong aren't the most violence occurring divisions.

H_1 : Dhaka and Chittagong is the most violence occurring divisions.

Table 4.7: Wilcoxon Signed-Rank Two-Tailed test P-value Comparison table of Divisions.

Division	Dhaka		Chittagong	
	statistic	pvalue	statistic	pvalue
Dhaka	–	–	–	–
Rajshahi	1.775592090	0.07580017458	1.775592090	0.07580017458
Sylhet	2.6111648393	0.009023438818	2.6111648393	0.009023438818
Chittagong	0.0	1.0	–	–
Mymensingh	2.40227165	0.016293603621	2.40227165	0.016293603621
Khulna	1.9844852778	0.04720176769	1.775592090	0.07580017458
Barishal	2.40227165	0.016293603621	2.40227165	0.016293603621
Rangpur	2.6111648393	0.009023438818	2.40227165	0.016293603621

The P_{val} derived from the test is lesser than the threshold ($P_{val} < \alpha = 0.05$) when compared to all the Division with Dhaka and Chittagong except each other. So we have sufficient evidence to reject the null hypothesis (H_0) and accept the alternative (H_1). The population medians are significantly different from Dhaka Chittagong to other divisions.

4.4.11.2 Co-relation test between Divisions population count and Violence Ratio

In this hypothesis, we would like to analyze the relationship between the population count of a division and its relation with the occurring violence frequency. For that, we collected population count data of divisions from Wikipedia of those particular years mentioned in the demi-decade and compared the violence occurring frequently with the Pearson correlation coefficient test. In the equation below (r_{dp}) is the Pearson correlation coefficient, (V_{fd}) values of frequency of divisions, ($\bar{V}_f$) is the mean values of frequency of divisions, (P_c) count of population of divisions, ($\bar{P}_c$) is the mean values of Population count of divisions:

$$r_{dp} = \frac{\sum(V_{fd} - \bar{V}_f)(P_c - \bar{P}_c)}{\sqrt{\sum(V_{fd} - \bar{V}_f)^2 \sum(P_c - \bar{P}_c)^2}}$$

After conducting the test we got (r_{dp}) value of 0.926055 and P_{val} of 0.0009555431 and since the correlation coefficient is close to 1, this tells us that there is a strong positive association between the two variables population and Violence count. Also, the p-value is less than the threshold ($P_{val} < \alpha = 0.05$), we conclude that there is a statistically significant association between the two variables population and Violence count.

4.4.11.3 Co-relation test between Divisions population count per Km and Violence Ratio

In this hypothesis, we have further analyzed the relationship between the violence event frequency and population count. Therefore we have broken down the units of the population per kilometer which was also collected from Wikipedia. Therefore we have the division population count per kilometer (P_{ckm})

and the violent event frequency of divisions that were present already. Finally, we have conducted the test with the Pearson correlation coefficient, the equation is given below:

$$r_{dp} = \frac{\sum(V_f, d - \bar{V}_f)(P_{c(km)}, d - \bar{P}_{c(km)})}{\sqrt{\sum(V_f, d - \bar{V}_f)^2 \sum(P_{c(km)}, d - \bar{P}_{c(km)})^2}}$$

The test results indicated that there isn't a strong positive association between the two variables population per Kilometer and Violence count as the value of (r_{dp}) is far from 1 ($r_{dp} = 0.545811$). Since the corresponding p-value is greater than 0.05, ($P_{val} > \alpha = 0.05$) where ($P_{val} = 0.1616920$) so we conclude that there is no statistically significant association between the two variables population per Kilometer and Violence count.

4.4.11.4 Co-relation test between Divisions Area Km square and Violence Ratio

For our new experiment, we collected the division's Area Km square data (A_{km}) from Wikipedia and calculated the correlation with the division's violence event frequency ratio. This study will suggest if there is any relation between the size of the divisions with the occurrence frequency of violence. Hence, we propose the Pearson correlation coefficient, to evaluate the claim and the equation is given below:

$$r_{dp} = \frac{\sum(V_f, d - \bar{V}_f)(A_{(km)}, d - \bar{A}_{(km)})}{\sqrt{\sum(V_f, d - \bar{V}_f)^2 \sum(A_{(km)}, d - \bar{A}_{(km)})^2}}$$

From, this correlation experiment we got the relation (r_{dp}) value of 0.807361 which is very close to 1 and it is indicating a strong positive association between the divisions are km square and the violence event ratio. the p-value is less than the threshold ($P_{val} < \alpha = 0.05$) where ($P_{val} = 0.015389$), we conclude that there is a statistically significant association between the two variables population and Violence count.

4.4.12 Co-Relation Between Target Violence Events

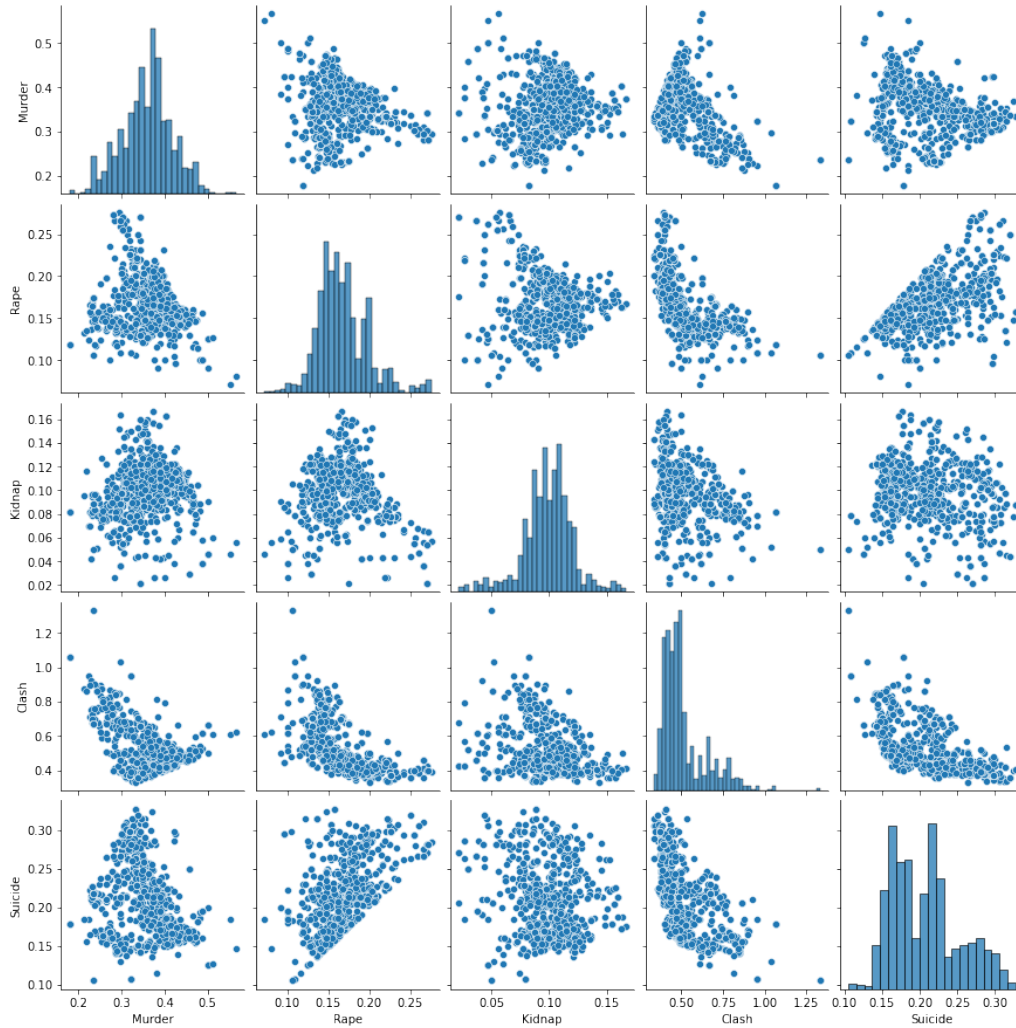
In the following experiment, we would like to analyze the correlation between violent events as our problem is a multi-class classification so there is more than one label associated with one new article data sample. Therefore we would like to deem that if any of the violent events are co-related with each other. For this, we picked a random date from the demi-decade dataset and also a random number of day intervals and calculated the frequency of each of the violent events. So after taking n number of samples from the demi-decade dataset we performed the Pearson correlation coefficient equation on the reformed data samples:

$$r_{dp} = \frac{\sum(V_f, i - \bar{V}_f, i)(V_f, j - \bar{V}_f, j)}{\sqrt{\sum(V_f, i - \bar{V}_f, i)^2 \sum(V_f, j - \bar{V}_f, j)^2}}$$

From the Pearson correlation (r) value table, we can see that there is a strong positive correlation between Suicide and Rape violent events also we can see in the plot that the grid indexes Rape and Suicide we can see the linear relationship between these two different events. Also between Murder

Table 4.8: Pearson Relation Value Table Between Violence Events.

Events	Murder	Rape	Kidnap	Clash	Suicide
Murder	1.000000	-0.156026	0.159635	-0.471064	-0.290150
Rape	-0.156026	1.000000	-0.176986	-0.599732	0.597381
Kidnap	0.159635	-0.176986	1.000000	-0.278234	-0.213892
Clash	-0.471064	-0.599732	-0.278234	1.000000	-0.560638
Suicide	-0.290150	0.597381	-0.213892	-0.560638	1.000000

**Figure 4.16:** Pearson Relation Between Violence Events.

and Kidnap events, there is a weak but positive correlation and in the figure above we can also see that the data samples slightly show a linear relationship with each other located in the Murder, Kidnap grid indexes.

Chapter 5

Conclusions

In this chapter, we would like to summarise our entire work that has been presented and discussed throughout the article. In a nutshell, we have collected Bangla news article data from various online sources using the crawling technique and also constructed a separate time series data of five years. We have conducted several experiments on the collected dataset and tested the data with various statistical tests to analyze trends rest of the summary is briefly described in the next section 5.1,. In the final section 5.2, of this chapter, we have discussed some of the potential future works related to this work. The works that couldn't be conducted in this and some further modifications that will assist to overcome some of the limitations of the proposed method:

5.1 Summary

To summarize our work we started by collecting Bengali news article data from various online portals and publicly available APIs. After that, we labeled the large volume dataset using the pre-trained model on the small volume dataset so that we could focus on potential violent data samples hence constructing the large volume news article training dataset by human intervention evaluation. Constructing the large volume dataset trained several models using different textual feature extractors and elected the best-performing BERT classifier model amongst them. We applied this model to the time series demi-decade dataset to extract some insight patterns related to these violent events and proposed several hypotheses based on the outlook of the results which we eventually evaluated with different statistical tests to establish some of the valid facts related to these violent events.

5.2 Future Work

In our future work, we intend to introduce a text similarity module that can ignore the same news article which will help to eliminate redundant data samples and make the frequency distribution more authentic. Alongside we want to extract more information from the articles by making our existing information-extracting procedures more reliable and authentic.

Bibliography

- [1] KR1442 Chowdhary. Natural language processing. *Fundamentals of artificial intelligence*, pages 603–649, 2020. 1
- [2] Caitlin Dreisbach, Theresa A Koleck, Philip E Bourne, and Suzanne Bakken. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International journal of medical informatics*, 125:37–46, 2019. 1
- [3] Tetsuya Nasukawa and Tohru Nagano. Text analysis and knowledge mining system. *IBM systems journal*, 40(4):967–984, 2001. 2
- [4] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*, 2021. 2
- [5] Iftakhar Ali Khandokar, Imtiaz Mamun, Tasmia Ishrat Alam Chadni, Zubair Ahmed Anas, and Swakkhar Shatabda. Event detection and knowledge mining from unlabelled bengali news articles. In *2020 Emerging Technology in Computing, Communication and Electronics (ETCCE)*, pages 1–6. IEEE, 2020. 2, 10, 11, 13
- [6] M Thangaraj and M Sivakami. Text classification techniques: a literature review. *Interdisciplinary Journal of Information, Knowledge, and Management*, 13:117, 2018. 4
- [7] Marcin Michał Mironczuk and Jarosław Protasiewicz. A recent overview of the state-of-the-art elements of text classification. *Expert Systems with Applications*, 106:36–54, 2018. 4
- [8] Kamran Kowsari, Kiana Jafari Meimandi, Mojtaba Heidarysafa, Sanjana Mendu, Laura Barnes, and Donald Brown. Text classification algorithms: A survey. *Information*, 10(4):150, 2019. 4
- [9] Ammar Ismael Kadhim. Survey on supervised machine learning techniques for automatic text classification. *Artificial Intelligence Review*, 52(1):273–292, 2019. 4
- [10] Zhichao Li, Helen Gurgel, Nadine Dessay, LuoJia Hu, Lei Xu, and Peng Gong. Semi-supervised text classification framework: An overview of dengue landscape factors and satellite earth observation. *International Journal of Environmental Research and Public Health*, 17(12):4509, 2020.

-
- [11] Hu Linmei, Tianchi Yang, Chuan Shi, Houye Ji, and Xiaoli Li. Heterogeneous graph attention networks for semi-supervised short text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4821–4830, 2019. 4
 - [12] Danilo Croce, Giuseppe Castellucci, and Roberto Basili. Gan-bert: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 2114–2119, 2020. 5
 - [13] Devendra Singh Sachan, Manzil Zaheer, and Ruslan Salakhutdinov. Revisiting lstm networks for semi-supervised text classification via mixed objective function. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6940–6948, 2019. 5
 - [14] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised neural text classification. In *proceedings of the 27th ACM International Conference on information and knowledge management*, pages 983–992, 2018. 5
 - [15] Yu Meng, Jiaming Shen, Chao Zhang, and Jiawei Han. Weakly-supervised hierarchical text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6826–6833, 2019. 5
 - [16] Cristian Cardellino, Serena Villata, Laura Alonso Alemany, and Elena Cabrio. Information extraction with active learning: A case study in legal text. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 483–494. Springer, 2015. 5
 - [17] Michael Altschuler and Michael Bloodgood. Stopping active learning based on predicted change of f measure for text classification. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 47–54. IEEE, 2019. 5
 - [18] Mohamed Goudjil, Mouloud Koudil, Mouldi Bedda, and Nouredine Ghoggali. A novel active learning method using svm for text classification. *International Journal of Automation and Computing*, 15(3):290–298, 2018. 5
 - [19] Petar Ristoski, Anna Lisa Gentile, Alfredo Alba, Daniel Gruhl, and Steven Welch. Large-scale relation extraction from web documents and knowledge graphs with human-in-the-loop. *Journal of Web Semantics*, 60:100546, 2020. 5
 - [20] Mahnoosh Kholghi, Lance De Vine, Laurianne Sitbon, Guido Zucco, and Anthony Nguyen. Clinical information extraction using small data: An active learning approach based on sequence representations and word embeddings. *Journal of the Association for Information Science and Technology*, 68(11):2543–2556, 2017. 5
 - [21] Shervin Minaee, Nal Kalchbrenner, Erik Cambria, Narjes Nikzad, Meysam Chenaghlu, and Jianfeng Gao. Deep learning-based text classification: a comprehensive review. *ACM Computing Surveys (CSUR)*, 54(3):1–40, 2021. 5

-
- [22] Weili Fang, Hanbin Luo, Shuangjie Xu, Peter ED Love, Zhenchuan Lu, and Cheng Ye. Automated text classification of near-misses from safety reports: An improved deep learning approach. *Advanced Engineering Informatics*, 44:101060, 2020. 5
- [23] Usman Naseem, Matloob Khushi, Shah Khalid Khan, Kamran Shaukat, and Mohammad Ali Moni. A comparative analysis of active learning for biomedical text mining. *Applied System Innovation*, 4(1):23, 2021. 6
- [24] Jiazhu Dai, Chuanshuai Chen, and Yufeng Li. A backdoor attack against lstm-based text classification systems. *IEEE Access*, 7:138872–138878, 2019. 6
- [25] Liang Yao, Chengsheng Mao, and Yuan Luo. Graph convolutional networks for text classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7370–7377, 2019. 6
- [26] Zeynep H Kilimci and Selim Akyokus. Deep learning-and word embedding-based heterogeneous classifier ensembles for text classification. *Complexity*, 2018, 2018. 6
- [27] Víctor Suárez-Paniagua, Renzo M Rivera Zavala, Isabel Segura-Bedmar, and Paloma Martínez. A two-stage deep learning approach for extracting entities and relationships from medical texts. *Journal of biomedical informatics*, 99:103285, 2019. 6
- [28] Sasa Arsovski, Hasmik Osipyan, Muniru Idris Oladele, and Adrian David Cheok. Automatic knowledge extraction of any chatbot from conversation. *Expert Systems with Applications*, 137:343–348, 2019. 6
- [29] Iqra Safder, Saeed-Ul Hassan, Anna Visvizi, Thanapon Noraset, Raheel Nawaz, and Suppawong Tuarob. Deep learning-based extraction of algorithmic metadata in full-text scholarly documents. *Information processing & management*, 57(6):102269, 2020. 6
- [30] Ning Pang, Zhen Tan, Xiang Zhao, Weixin Zeng, and Weidong Xiao. Domain relation extraction from noisy chinese texts. *Neurocomputing*, 418:21–35, 2020. 6
- [31] Sunil Kumar Sahu and Ashish Anand. Drug-drug interaction extraction from biomedical texts using long short-term memory network. *Journal of biomedical informatics*, 86:15–24, 2018. 6
- [32] Berke Oral, Erdem Emekligil, Seçil Arslan, and Gülşen Eryiğit. Information extraction from text intensive and visually rich banking documents. *Information Processing & Management*, 57(6):102361, 2020. 6
- [33] Katerina Goseva-Popstojanova and Jacob Tyo. Identification of security related bug reports via text mining using supervised and unsupervised classification. In *2018 IEEE International conference on software quality, reliability and security (QRS)*, pages 344–355. IEEE, 2018. 7
- [34] Francesca Greco and Alessandro Polli. Emotional text mining: Customer profiling in brand management. *International Journal of Information Management*, 51:101934, 2020. 7

-
- [35] Abeed Sarker and Graciela Gonzalez-Hernandez. An unsupervised and customizable misspelling generator for mining noisy health-related text sources. *Journal of biomedical informatics*, 88:98–107, 2018. 7
 - [36] Mangi Kang, Jaelim Ahn, and Kichun Lee. Opinion mining using ensemble text hidden markov models for text classification. *Expert Systems with Applications*, 94:218–227, 2018. 7
 - [37] Tian Xia and Xuemin Chen. A discrete hidden markov model for sms spam detection. *Applied Sciences*, 10(14):5011, 2020. 7
 - [38] Stéphan Tulkens, Simon Šuster, and Walter Daelemans. Unsupervised concept extraction from clinical text through semantic composition. *Journal of biomedical informatics*, 91:103120, 2019. 7
 - [39] Fethi Fkih and Mohamed Nazih Omri. Hidden data states-based complex terminology extraction from textual web data model. *Applied Intelligence*, pages 1–19, 2020. 7
 - [40] Qiang Wei, Zongcheng Ji, Zhiheng Li, Jingcheng Du, Jingqi Wang, Jun Xu, Yang Xiang, Firat Tiryaki, Stephen Wu, Yaoyun Zhang, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 27(1):13–21, 2020. 7
 - [41] Fatima N Al-Aswadi, Huah Yong Chan, and Keng Hoon Gan. Automatic ontology construction from text: a review from shallow to deep learning trend. *Artificial Intelligence Review*, 53(6):3901–3928, 2020. 7
 - [42] Hyosun An and Minjung Park. Approaching fashion design trend applications using text mining and semantic network analysis. *Fashion and Textiles*, 7(1):1–15, 2020. 7
 - [43] Bruno Justino Garcia Praciano, João Paulo Carvalho Lustosa da Costa, João Paulo Abreu Maranhão, Fábio Lúcio Lopes de Mendonça, Rafael Timoteo de Sousa Júnior, and Juliano Barbosa Prettz. Spatio-temporal trend analysis of the brazilian elections based on twitter data. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1355–1360. IEEE, 2018. 7
 - [44] Babak Sohrabi, Iman Raeesi Vanani, and Ehsan Abedin. Human resources management and information systems trend analysis using text clustering. *International Journal of Human Capital and Information Technology Professionals (IJHCITP)*, 9(3):1–24, 2018. 7
 - [45] Frederik Hogenboom, Flavius Frasincar, Uzay Kaymak, Franciska De Jong, and Emiel Caron. A survey of event extraction methods from text for decision support systems. *Decision Support Systems*, 85:12–22, 2016. 7
 - [46] David A Juckett, Eric P Kasten, Fred N Davis, and Mark Gostine. Concept detection using text exemplars aligned with a specialized ontology. *Data & Knowledge Engineering*, 119:22–35, 2019. 8

- [47] Jose L Martinez-Rodriguez, Ivan Lopez-Arevalo, and Ana B Rios-Alvarado. Openie-based approach for knowledge graph construction from text. *Expert Systems with Applications*, 113:339–355, 2018. 8
- [48] Amina Boufrida and Zizette Boufaïda. Rule extraction from scientific texts: Evaluation in the specialty of gynecology. *Journal of King Saud University-Computer and Information Sciences*, 2020. 8
- [49] Chien-wen Shen, Min Chen, and Chiao-chen Wang. Analyzing the trend of o2o commerce by bilingual text mining on social media. *Computers in Human Behavior*, 101:474–483, 2019. 8
- [50] Dhaka Tribune. <https://www.dhakatribune.com/>. Accessed: 2023-01-21. 9
- [51] beautifulsoup4 4.11.1. <https://pypi.org/project/beautifulsoup4/>. Accessed: 2023-01-23. 10
- [52] Kaler Kontho Newspaper. <https://www.kalerkantho.com/>. Accessed: 2023-01-21. 10
- [53] Jai Jai Din Newspaper. <https://www.jaijaidinbd.com/>. Accessed: 2023-01-21. 10
- [54] Amader Shomoy Newspaper. <https://www.dainikamadershomoy.com/>. Accessed: 2023-01-21. 10
- [55] Ajker Patrika Newspaper. <https://www.ajkerpatrika.com/>. Accessed: 2023-01-21. 10
- [56] Bhorer Kagoj Newspaper. <https://www.bhorerkagoj.com/>. Accessed: 2023-01-21. 10
- [57] bnlp-toolkit 3.2.0. <https://pypi.org/project/bnlp-toolkit/>. Accessed: 2023-01-23. 10
- [58] Natural Language Toolkit. <https://www.nltk.org/>. Accessed: 2023-01-23. 10
- [59] bangla-stemmer 1.0. <https://pypi.org/project/bangla-stemmer/>. Accessed: 2023-01-23. 10
- [60] scikit-learn: machine learning in Python — scikit-learn 1.2.0. <https://scikit-learn.org/stable/>. Accessed: 2023-01-23. 11
- [61] Tensorflow. https://www.tensorflow.org/api_docs/python/tf. Accessed: 2023-01-23. 11
- [62] Fasttext. <https://fasttext.cc/docs/en/python-module.html>. Accessed: 2023-01-23. 11
- [63] Prothom Alo Newspaper. <https://www.prothomalo.com/>. Accessed: 2023-01-21. 12, 16