



Article Using Machine Learning to Detect Events on the Basis of Bengali and Banglish Facebook Posts

Noyon Dey ^{1,†}, Md. Sazzadur Rahman ^{1,*,†}, Motahara Sabah Mredula ¹, A. S. M. Sanwar Hosen ² and In-Ho Ra ^{3,*}

- ¹ Institute of Information Technology, Jahangirnagar University, Dhaka 1342, Bangladesh; noyondey8@gmail.com (N.D.); mmredula12@gmail.com (M.S.M.)
- ² Division of Computer Science and Engineering, Jeonbuk National University, Jeonju 54896, Korea; sanwar@jbnu.ac.kr
- ³ School of Computer, Information and Communication Engineering, Kunsan National University, Gunsan 54150, Korea
- * Correspondence: sazzad@juniv.edu (M.S.R.); ihra@kunsan.ac.kr (I.-H.R.)
- + These authors contributed equally to this work.

Abstract: In modern times, ensuring social security has become the prime concern for security administrators. The widespread and recurrent use of social media sites is creating a huge risk for the lives of the general people, as these sites are frequently becoming potential sources of the organization of various types of immoral events. For protecting society from these dangers, a prior detection system which can effectively detect events by analyzing these social media data is essential. However, automating the process of event detection has been difficult, as existing processes must account for diverse writing styles, languages, dialects, post lengths, and et cetera. To overcome these difficulties, we developed an effective model for detecting events, which, for our purposes, were classified as either protesting, celebrating, religious, or neutral, using Bengali and Banglish Facebook posts. At first, the collected posts' text were processed for language detection, and then, detected posts were pre-processed using stopwords removal and tokenization. Features were then extracted from these pre-processed texts using three sub-processes: filtering, phrase matching of specific events, and sentiment analysis. The collected features were ultimately used to train our Bernoulli Naive Bayes classification model, which was capable of detecting events with 90.41% accuracy (for Bengalilanguage posts) and 70% (for the Banglish-form posts). For evaluating the effectiveness of our proposed model more precisely, we compared it with two other classifiers: Support Vector Machine and Decision Tree.

Keywords: Banglish; Bengali; Bernoulli Naïve Bayes; decision tree; event detection; social media; support vector machine

1. Introduction

The scale and interactivity of social media sites result in the generation of a massive volume of data in the form of audio, video, text, and images relevant to users' personal, social, political, and economic lives. These features have allowed participants to organize events around the world, which were reflected in the George Floyd protests in the United States on 26 May 2020 [1] and the Arab Spring of 2010–2012 [2]. Similar events have been organized in Bangladesh, including student protests in Dhaka in 2018 [3] and road safety protests in the same city in 2019 [4]. The harshness of these frightening events has appalled us, which led us to analyze user-generated social media data for building an event detection framework so that security authorities could acquire appropriate information at the right time and maintain social security.

A great deal of research over the last decade has focused on event detection and prediction using information pulled from social media. Numerous researchers have examined



Citation: Dey, N.; Rahman, M.S.; Mredula, M.S.; Hosen, A.S.M.S.; Ra, I.-H. Using Machine Learning to Detect Events on the Basis of Bengali and Banglish Facebook Posts. *Electronics* **2021**, *10*, 2367. https:// doi.org/10.3390/electronics10192367

Academic Editor: Juan M. Corchado

Received: 13 July 2021 Accepted: 14 September 2021 Published: 28 September 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https:// creativecommons.org/licenses/by/ 4.0/). traffic [5,6], disaster [7,8], disease [7,9], sporting [10], earthquake [8], and crime events [10], to name but a few. These studies have examined events described in English [11,12], Hindi [12], Mandarin [13], Urdu [14], Japanese [8], Korean [15], Arabic [16], and other languages, and have examined data gathered from different platforms to conduct their research including Twitter [17–19] or Sina Weibo [20,21]. Although these studies focused on different languages and platforms from different countries, none of them have analyzed Bengali and its related complexities in detecting events.

To date, no research team has focused on detecting or predicting religious events. Religious, cultural, and political events occur throughout the year in Bangladesh and it is common to organize events on Facebook since it is used by 86.73% of all social media users in Bangladesh [22]. In fact, approximately 41 million people use Facebook on a daily basis across the country [23]. As Facebook is a popular social media site, quite a number of researchers have conducted their experiments using Facebook data [24,25]. Moreover, existing works have only considered social media posts of short pre-determined lengths, like those possible on Twitter, i.e., 280 characters per post [26], while Facebook posts are often longer and more complex. This study is also unique for its consideration of posts in Bengali and Banglish, i.e., Bengali words written in the English alphabet, as they are widely used in Bangladesh for posting on Facebook. Especially, Banglish offers significant challenges in text processing due to its multiple possible representations of a word, i.e., a word with various spellings but the same pronunciation and meaning.

Though some researchers have used specific keywords for conducting their research [8,20], no researchers have acknowledged Bengali and Banglish event-specific word lists for their experiments. Event-specific phrase lists for Bengali and Banglish are also unacknowledged until now. In this paper, we discovered separate lists for Bengali and Banglish event-specific words as well as for event-specific phrases. So far, none have recognized multiple representations of Banglish words and phrases. In our work, we recognized every possible representation of Banglish words and phrases.

To the best of our knowledge, this is the first study that analyzes Facebook posts written in Bengali and Banglish to detect events. Though there exists some works in the Bengali language [24,25,27–29], all these works are of opinion or sentiment analysis. In this study, we extended our previous work where we only worked with Bengali Facebook posts with an accuracy of 87.5% in detecting events [30]. Our previous work utilized the same procedures, i.e., data collection, pre-processing, features extraction, model training, and detection, which we followed in this work. Building on this prior work, we increased our previous work's event-detection accuracy and also worked with the Banglish-form posts. We modified our previous model by recognizing the multiple representations of the Banglish words. Besides, multiple representation of both Bengali and Banglish phrases' were also added in this model to enhance the detection accuracy. We also improved the language detection procedure for the Banglish-form post detection since it is not directly detectable by available libraries. Furthermore, sentiment score of the Banglish posts were also calculated in a different way compared with Bengali posts since it is not directly calculable using current software libraries. Eventually, in this study, we relied on Facebook posts written in Bengali-language and Banglish-form. We performed a detailed study on event-related posts in Bengali and Banglish, and determined commonly used event words and specific event phrases. We also recognized multiple representations of Banglish words and Bengali and Banglish phrases. This multiplicity of word and phrase presentations, if unacknowledged, would have created incorrect feature values and incorrect detections. Besides, we analyzed the sentiments of both Bengali and Banglish posts and finally detected four types of events: celebrating, protesting, religious, and neutral.

The rest of the paper is structured as follows. In Section 2, we review the relevant literature study that predates our work. In Section 3, we describe our proposed model, and we evaluate its performance in Section 4. In Section 5, we offer our concluding thoughts and propose some directions for future research.

2. Literature Review

The huge data repository, i.e., social media, has fascinated researchers, and it comes as no surprise that event prediction and detection has been a recurring theme in the literature. Different approaches, including machine learning, neural network, rule-based approach, and many other techniques have been adopted by researchers to implement in their work. Table 1 presents a tabular representation of some of these background studies.

For predictive analysis, a neural network provides the best analysis, as it makes use of hidden layers. Many researchers have thus turned to neural networks to effectively detect events from social media posts [31–33]. Chen et al. [31] used a neural network to introduce an online event identification module. At first, a classification framework was trained to track events, on the basis of a posting's content. Then, a clustering-based effective method was adopted to identify and trace the events, and a memory component was utilized to store and refresh event portrayal. Bekoulis et al. [32] focused on the sequential nature of data streams to detect events from social media data with a neural sequence model. This team not only detected the existence of sub-events from Twitter data, but also managed to identify the type of the detected sub-event. They presented a baseline model for binary classification and showed that their proposed model outperformed the state-of-the-art in identifying the presence or absence of the sub-event. To achieve these results, they used AVG (Average) pooling and MLP (Multi-Layer Perceptron). Aldhaheri et al. [33] used neural networks to suggest a unique outline for event detection. A temporal approach that converted social media data streams into consecutive images was proposed, and, ultimately, they proved that the transformation of a data stream to images covers the overall complication of the social media chain and enhances event-detection accuracy.

To accurately detect events, some researchers have relied on unsupervised learning (deemed 'the clustering approach'). In this method, a given data set is split into similar objects. P.N. et al. [9] used this clustering technique to propose an event transformation framework based on user interest (the 'hot event evolution model'). A clustering algorithm (the 'hot event automatic algorithm') was used to associate the mini-texts cluster by cluster with their relevant topics. An online clustering approach has also been attempted. Alsaedi et al. [18] proposed an assimilated model that detected events from Arabic Twitter data. This research team was primarily interested in filtering disruptive events from social media data streams, and differentiating between these and other events. Disruptive events included different types of protests, terrorist attacks, transport loss, and certain crimes. They mentioned several stages, including data collection, pre-processing, classification, clustering, and summarization, and used the Naive Bayes (NB) classification model along with an online clustering algorithm for event detection. Automatically Named Entity Recognition (NER), dictionaries, and various tweet features (hashtags and retweet ratio) were also used to improve event detection.

Some event detection methods focus on identifying how events evolve over time. One example of this technique was presented by Fedoryszak et al. [11], who applied bundling on a large stream to produce a dynamically rationalized set of events. This method was arrayed to work in real-time at the scale of Twitter-sized postings. Both offline and online estimation of their model were presented. Li et al. [34] proposed a procedure that consisted of three steps: tweet classifications into relevant semantic classes, calculation of the class relationship between the tweets, and integration. They used different semantic classes, including proper noun, location, mention, verb, common noun, and hashtag. They also introduced a chronological information identification module to identify temporal information in a cluster. This module indicates whether an event is new or old.

Event detection research has reached into the domain of sports as well. Kannan et al. [35] presented a novel procedure for identifying critical real-time events from live tweets associated with cricket. They applied the Locality Sensitive Hashing (LSH) technique to fulfill the online incremental huddling of tweets of the cricket domain and key events were acknowledged by leveraging the event lexicon. Some detection approaches also account for location. Feng et al. [36] used LSH to achieve improved similarity comparisons. Their proposed method to extract location information was based on Part-of-Speech (POS) tagging and a Support Vector Machine (SVM) classifier. They presented a unique similarity dimension that considered message content, as well as its time and location, to improve the speed and accuracy of event detection.

Other researchers have concentrated on the domain of natural disasters [7,8,12,37,38] using data from case studies with the goal of improving mankind's resilience in the face of adverse natural calamities. Imran et al. [7] proposed an effective framework capable of extracting disaster-related data from tweets. By analyzing the behavior of social-media content propagated during two different natural disasters, they trained a Conditional Random Fields (CRF)-based model to identify valuable information. Their system detected between 40% and 80% of the tweets that contained information relevant to the disaster. Sakaki et al. [8] analyzed Twitter data to detect real-time earthquake events. A spatiotemporal model based on probability was built to detect this circumstance. For location estimation, Kalman Filtering and Particle Filtering were used. Other researchers have used the available data for wellness event detection. Akbari et al. [39] presented a framework capable of identifying wellness events from users' published social media postings. Their skeleton utilized the content of microblogged texts, along with the relation between event categories, to extract Personal Wellness Events (PWE). Their process relied on a trained Multi-Task Learning (MTL) model with a Lasso regularizer to evaluate the model for every task, and they finally calculated the correlations between individual tasks.

While most researchers have concentrated on Twitter and Facebook data, some researchers have worked with other social media data. Using this data, Panagiotou et al. [40] rearranged the complication of fuzzy concepts and summarized a number of different approaches to event detection. Moreover, they developed an easy-to-use language for describing the state-of-the-art method under the same scheme. Some researchers have emphasized the test corpus used for detecting events [41], while others have focused only on the chronological relationship between available data [42,43].

Koyla et al. [41] explored events from various newspapers where vectors (person, place, time, date, etc.) were mainly considered. They basically intended to find out if two news documents of the same time period indicate the same event or not. For this, a predefined threshold value was considered in order to check if the total number of vectors of two separate documents matched by at least this threshold. Using the SVM-based NER, they generated their event vectors. Zhao et al. [42] generated an intermediate semantic level Microblog Clique (MC) capable of analyzing the vast number of interrelationships among microblogs. They unitedly engaged the textual, visual, and social content in microblog evaluation for truly exploring the inherent interrelation among the diverse data. Shi et al. [43] proposed a Hypertext-Induced Topic Search (HITS) based on the Topic-Decision method (TD-HITS) in conjunction with a Latent Dirichlet Allocation (LDA)-based Three-Step model (TS-LDA). TDHITS was able to precisely determine the total number of topics from a vast collection of posts while identifying connected key posts. Some researchers have undertaken a literature review of available papers and ably summarized the studied methods [44–49].

Nurwidyantoro et al. [44] focused on methods of detecting disasters, traffic congestion, pandemics, and news topics, and presented some definitions relevant to event detection. Zarrinkalam et al. [45] provided a comprehensive overview of available methods and some impediments to their effective functioning, including short length, noisiness, and the informality of the social contents. They classified available techniques as either specified or unspecified event-detection types. Moreover, they provided some event-detection techniques based on potential applications. Dou et al. [46] presented four tasks to detect events: new event detection, event tracking, event summarization, and event association.

Papers with Publishing Year	Method	Objective Used Language		Used Data	Dissimilarity from Our Work
Ref. [9] "Event detection and text summary by disaster warning", 2019.	LDA	-Proposed a user interest-based model. -Their model outputs a brief summary of the microblogging comments.	-	-	-Examined only Twitter data. -Only summarized comments of microblog data. -Discarded some tweet attributes (i.e., embedded URL) while computing.
Ref. [11] "Real-time event detection on social data streams", 2019.	Clustering algorithm	-Handled event progress over time. -Estimated both online and offline performance.	-English	-	-Analyzed only Twitter data. -Their dataset only consisted of English tweets.
Ref. [18] "Arabic event detection in social media", 2015.	NB and online clustering scheme	-Recognized disruptive events from Arabic tweets.	-Arabic	-1.7 million tweets	-Adopted only the Arabic language while ignoring local languages used in tweeting. -Detected only disruptive events while ignoring other types of events.
Ref. [24] "Depressed People Detection from Bangla Social Media Status using LSTM and CNN Approach", 2021.	Hybrid CNN-LSTM	-Utilized a hybrid algorithm. -Detected depressed people.	-Bangla	-7163.	-Only the Bangla language was considered. -The dataset was still in the update stage.
Ref. [25] "Bangla hate speech detection on social media using attention-based recurrent neural network", 2021.	LSTM and GRU (hybrid) model El	-Disclosed Bangla hate speech. -Classified news comments into seven categories.	-Bangla	-	-Experimented only with Bangla language. -Analyzed only news comments rather than whole posts.
Ref. [27] "Attention-based convolutional neural network for Bangla sentiment analysis", 2021.	CNN	-Effectively incorporated attention mechanism. -Analyzed Bangla sentiment from comments and reviews.	-Bangla	-2979 reviews and comments	-Did not occupy semantic meanings of individual words. -Could not bypass word-sense ambiguity.
Ref. [28] "Identifying and Categorizing Opinions Expressed in Bangla Sentences using Deep Learning Technique", 2020.	Deep learning networks (CNN and LSTM)	-Categorized sports news comments based on their sentiment. -Explored four types of sentiments: happiness, sadness, advice, annoyance.	-Bangla	-2492 sentences	-Only employed the Bangla language. -Utilized news comments only.
Ref. [29] "Bangla Text Classification using Transformers", 2020.	Multilingual BERT and XLM-RoBERTa	-Explored different transformer models for classifying text. -Conducted work in the domain of sentiment analysis, news categorization, emotion detection, and authorship distribution.	-Bangla	-Youtube comment dataset (15,686) -News comment sentiment dataset (13,802) -Authorship attriution dataset (14,047) -News classification dataset (11,284)	-Dataset consisted of Bangla articles only.

Table 1. A summary of some significant papers of the literature review including their main objective, used method, data, and language along with some dissimilarities from our approach.

Papers with Publishing Year	Method	Objective	Used Language	Used Data	Dissimilarity from Our Work
Ref. [31] "Online event detection and tracking in social media based on neural similarity metric learning", 2017.	NN	-Distinguished and traced events. -Utilized memory module.	-	-9,563,979 tweets	-Conducted their experiment only with Twitter data.
Ref. [32] "Sub-event detection from twitter streams as a sequence labeling problem", 2019.	Long Short Term Memory (LSTM), MLP	-Detected the existence and type of sub event. -Basically, focused on the chronological relation of the tweets.		-2 M	-Used Twitter data.
Ref. [33] "Event detection on large social media using temporal Analysis", 2017.	Neural Network (NN)	-Proposed a temporal approach of event detection. -Also detected the complexity of social media chains.	-	-17 GB	-Worked with Twitter data only. -Did not consider Bengali or Banglish posts.
Ref. [34] "Real-time novel event detection from social media", 2017.	Clustering algorithm	-Focused on improving event detection performanceEnglis -Also identified temporal information.		-120 million tweets collected (finally 100 k from them were used)	-Used only Twitter data.
Ref. [35] "Sportsbuzzer: detecting events at real time in twitter using incremental clustering", 2018.	er: -Identified events from cricket domain. sing 2018. -Used event lexicon for identification of event.		-	-Tweets of 44 games with a file size of over 6 GB	-Detected sports event. -Exploited Twitter data.
Ref. [36] "Efficient location-based event detection in social text streams", 2015.	LSH and SVM classifier	-Proposed a location-based event detection method. -Considered message content along with its time.	-	-257,872 messages	-Collected microblogs only from Sina Weibo.
Ref. [42] "Real-time multimedia social event detection in microblog", 2018.	Hypergraph cut method [50] and transfer cut method [51]	-Considered the correlation among data. -Generated an intermediate semantic level.	-	-3 million microblogs	-Dataset consisted of microblogs from Sina Weibo. -Considered neither the Bengali nor Banglish languages.

Table 1. Cont.

Deep learning, machine learning, and other techniques have all been tried as a means of detecting or predicting events. In this work, the NB classification model, which is capable of quick calculations and high accuracy, was used in addition to language detection, pre-processing, filtering, phrase matching of specific events, and sentiment analysis for detecting events. In the following section, we provide a detailed explanation of our module as well as the tools and technologies used for its implementation.

3. Proposed Model

Our proposed model includes data collection, language detection, data pre-processing, feature extraction, model training, and event detection steps. These steps are depicted in Figure 1. This model allowed us to detect celebrating, protesting, religious, and neutral events after analysis of Bengali and Banglish Facebook posts' text. 'Celebrating events' include marriages, cultural events, and etc. Student or human-chain protests (a demonstration of individuals who form a chain structure by holding hands for showing solidar-ity [3,4]) or rallies were classified as 'protesting events'. Eid celebrations, funerals, venerations, and etc. were classified as 'religious events'. Any events outside these frameworks were deemed 'neutral events'.



Figure 1. Model for event detection on the basis of the text of Bengali and Banglish-form Facebook posts.

To detect events, we used only real Bengali and Banglish posts collected from Facebook. Algorithm 1 depicts the pseudocode of the overall process of our event-detection procedure. The language in which the collected posts were written was detected using the "langdetect" python library [52], and this procedure was performed using Algorithm 2. After language detection, pre-processing of the detected posts was performed using the *pre-Process*(*P*) function and features were extracted afterwards. Feature extraction consisted of three sub-processes: filtering, phrase matching of specific events, and sentiment analysis. These three sub-processes extracted common and specific event words and common and specific event phrases using Algorithm 3, and sentiment scores using the V(P) procedure, respectively. V(P) calculated sentiment using the Valence Aware Dictionary and Sentiment Reasoner (VADER) [53]. Ultimately, all of the features extracted from the post, were run through the Bernoulli Naïve Bayes (BNB) classification model and the model then detected events based on the feature values. These steps are described in greater detail in the following subsections.

Algorithm 1. Event Detection

1: Input: $P_{i} = ('hn' / an')$ M list = ('number of a list = ('nhraces')						
1. Input. $F, L = \{bn, en,, \}, w_{list} = \{words, F_{list} = \{phrases\}$						
2: Begin						
3: Algorithm 2 (P)						
4: <i>if</i> Algorithm 2 == True <i>then</i>						
5: P_processed = preProcess (P)						
6: W_count = Algorithm 3 (P_processed)						
7: $P_count = Algorithm 3 (P)$						
8: $S_score = V(P)$						
9: Train Model						
10: Detect Event						
11: else						
12: Take new post						
13: end						
14: Output: Event Type						

3.1. Data Collection

As no dataset of Bengali and Banglish Facebook posts existed to detect events, we made our own by manually collecting posts from various public pages maintained by a variety of groups and individuals (e.g., Supplementary Material Document S1). For collecting data regarding different student-relevant events, we considered various public and private university Facebook pages and groups, as most of the students and other university officials frequently post about protesting, celebrating, and religious event-related information in these groups and pages. Moreover, for covering countrywide protesting, celebrating, and religious event-related information, we selected quite a good number of popular public Facebook groups and pages which have a significant number of followers. The collected posts contained information related to recent and prior events. The dataset skewed towards previous events due to restrictions on public gatherings put in place in response to the COVID-19 outbreak. Our datasets were kept small due to this situation and used no automatic data collection method. We did not build any automatic data collection tool, and there are also no benchmark datasets available in Bengali and Banglish for event detection. For these reasons, we had to collect posts manually from different Facebook pages, groups, or individual accounts. This manual collection limited our data collection efforts. The collected posts were then checked to determine whether they were in Bengali or Banglish.

3.2. Language Detection

Posts not written in Bengali or Banglish were excluded from this study. Hence, Bengali and Banglish were detected using "langdetect" — a python library that supports 55 different languages. Using the "detect" method allows a user to input a text and receive the detected language's short form (in the case of Bengali, that is 'bn'). Algorithm 2 depicts the process of language detection in which post (*P*) and list of supported languages (*L*) were the inputs. *length* (*L*) is a procedure that provides the number of elements in a list. *detect* (*P*) is a procedure that determines the language of the post, and *bnb* (*P*) is another procedure that converts the encountered Banglish post into Bengali using 'bnbphoneticparser' [54]. *processFurther* (*P*) is a procedure that further processes the post for event detection.

```
Algorithm 2. Language Detection
```

```
1: Input: P, L = { 'bn', 'en', .... }
2: Begin
3:
     S = length(L)
     if (Dl = detect(P)) == 'bn' then
4:
5:
         processFurther(P)
6:
     else
7:
        c = 0
       for each K_i \in L do
8:
9:
           if\{(Dl = detect(P)) = L_i \& Dl! = en'\} then
10:
                c = c + 1
11:
           else
12:
              c = 0
13:
              break
14:
        end
15:
       if c = 0 then
16:
           cnv \ post = bnb(P)
17:
           if (Dl = detect(P)) == 'bn' then
              processFurther (P)
18:
19:
           else
20:
              Discard P and goto step 1
21:
           enđ
22:
        else
23:
           Discard P and goto step 1
24:
        end
     enđ
25:
26: Output: P
```

For Banglish post-detection, the algorithm follows a different approach since Banglish is not directly supported by any of the available libraries. Besides, there is a possibility that the library may detect Banglish as an English post since the post contains only English letters. That is why the post is checked for English posts as well in lines 8 to 14 of Algorithm 2. If the condition of line 15 is false, the post is discarded as a non-Bengali and non-Banglish post, and otherwise it defines a post with possibilities both in English or Banglish. Since the post must be either in English or Banglish, we employed a function, *bnb*. The *bnb* is a library function that takes Banglish as input and converts it into Bengali. If any other language's text is given as input, the output will not be a Bengali text. Hence, we used this as a determining factor for Banglish since only Banglish can be converted into Bengali by bnb. After conversion, we again detect the language; if it is 'bn', then it is confirmed as a Banglish post. Otherwise, it is discarded.

3.3. Data Pre-Processing

Tokenization and stopwords-removal processes were applied for data preprocessing. Pre-processing was applied because postings made on social media platforms typically are written informally and reflect a user's idiosyncratic writing style. This style of writing provides significant challenges and hence the use of pre-processing reduces impact of this issue. In this pre-processing step, stopwords were removed, and tokenization was applied subsequently. Stopwords are those words, characters, or special characters that offer no valuable insight into the text, and were accordingly excluded. Tokenization breaks the text stream into its component words for further processing. Overall, data preprocessing renders posts more suitable for feature extraction.

3.4. Feature Extraction

The frequency of common event keywords, the frequency of specific event phrases, i.e., celebrating, protesting, and religious, the frequency of specific event keywords, i.e., celebrating, protesting, and religious, and the sentiment analysis score were used as features in this work. At first, common event words were detected from the post since commonevents words primarily select a post as a probable candidate for an event. Afterwards, specific event-related words were detected from the post and their occurrence further bolstered the possibility of a post being a type of event. Along with the words, event-specific phrases were detected and used as the most important type of feature since this feature greatly influences a post's classification of an event category. Eventually, the sentiment score was calculated to further improve the possibility of a post's type since sentiments are correlated with any kind of events. In short, a gradual selection of common-events words, specific event words, specific event phrases, and eventually the sentiment score detected an event. Filtering, phrase matching of specific events, and sentiment analysis are the three sub-processes that were applied in this feature extraction process.

3.4.1. Filtering

Filtering determines the frequency of common event and specific event keywords from the post's text. Algorithm 3 describes this process.

Here, P (posts), C {'a', 'b', ... } (words lists) and $Ph{'abc', 'def', ...}$ (phrase lists) are inputs, W_{type} and Ph_{type} are the types of words and phrases respectively determined by a function length(), W_{count} and Ph_{count} stores every type of words and phrases' occurrences respectively, W_x is a word from P and K_y is a word from W[i] (specific type of word), and P_{phrase} is a phrase from Ph[j] (specific type of phrase). Filtering was performed in steps. In each step, the event words from W were detected and their frequency was recorded. Common and event-specific words were detected using the steps from line 7 to 18 of Algorithm 3. In these steps, every type of word from W was checked to see if they match with a word from the post using a loop. Those types' occurrences were also tracked using a designated storage W_{count} [i].

Common-event keywords are defined as commonly used words to describe an event. Both Bengali and Banglish employ common event keywords, as well as specific eventrelated words, some of which are presented in Table 2. The first row of Table 2 shows Bengali event words along with their equivalent English words in brackets and the second row shows the corresponding Banglish words.

Since Banglish may include multiple representations of a word, all possible combinations of the relevant words were identified and used in this work. Figure 2 describes this multiple representations of Banglish words.

In the Banglish writing form, a word's pronunciation and meaning remain the same, while that word's spelling can be presented in multiple ways. For example, "My" is an English word, while the equivalent Bengali form is "আমার" and its Banglish representations can be "Amar", "Amr", "Amaar", or "Aamar". Such cases where there are more than four alternative spellings of a word are common. Such complexity among representations affects sentiment scores.

Common Event's Words	Celebrating Event's Words	Protesting Event's Words	Religious Event's Words
ঘটেছ (happening),ঘটবে (will happen), সমাবেশ (assembly), জমায়েত (gathering),সমাগম (gathering), সভা (meeting)	আনন্দমেলা (funfair), অনুষ্ঠান (ceremony), বিয়ে (marriage), পুনঃমিলনী (reunion), মেলা (fair)	মিছিল (procession), মানববন্ধন (human chain), বিক্ষোভ (demonstrato), সংঘর্ষ (conflct), হামলা (attack), আন্দোলন (protest)	গুয়াজ (waaz), মাহফিল (religious concert), নামাজ (prayer), দাফন (burial), জানাযা (funeral), পূজা (worship)
Ghotche, ghotbe, shomabesh, jomayet, shomagom, shova	Anondomela, onusthan, biye, punomiloni, mela	Michil, manobbondhon, bikkhov, shongghorsho, hamla, andolon	Waaz, mahfil, namaz, dafon, janaza, puja

Table 2. Common event words and specific events' keywords used for filtering.

Algorithm 3. Filtering of Words and Phrases

```
Input: P, W{ 'a', 'b', ... }, Ph{ 'abc', 'def', ... }
1.
2:
     Begin
3:
     W_{type} = length(W)
4
     Ph_{type} = length (Ph)
5:
     Wcount[Wtvpe], Phcount [Phtvpe]
     i \equiv 0, j \equiv 0
6:
     while i < W_{type}
7:
8.
         for each W_x \in P do
9:
              d = 0
10:
              for each K_v \in W[i] do
                   if W_x == K_y then
11:
12:
                       d \equiv d + 1
13:
                   end
14:
               end
15:
          end
16:
         W<sub>count</sub>[i] = d
17:
         i += 1
18
     end
     while j < Ph_{type}
19:
         for each P_{phrase} \in Ph[j] do
20:
             d = 0
21:
             if P<sub>phrase</sub> in P then
22:
23:
                 d \equiv d + 1
24:
             end
25:
          end
26:
          Phcount[j] = d
27:
         j += 1
28:
       end
29:
       Output: Wcount, Phcount
```



Figure 2. An example of multiple representations of Banglish words.

3.4.2. Phrase Matching of Specific Event

A phrase is a collection of words that collectively define a meaning. Bengali and Banglish both employ specific phrases to describe events. Phrase matching of specific events follows the same process as Algorithm 3. Phrases and their occurrences were tracked along with their types using steps from line 19 to 28. For every phrase from Ph[j], it was checked to see if it was in the post P. If it was there, its occurrence was recorded in its designated storage $Ph_{count}[j]$, otherwise a new phrase was checked.

Some of the specific event phrases we looked for are presented in Table 3. The first row shows event phrases in Bengali along with their equivalent English phrases in brackets. Corresponding Banglish representations of the phrases are presented in the second row. All possible combinations of phrases were identified and used in this work since multiple representations of phrases is also a common scenario for both Bengali and Banglish phrases. This multiple-phrase representation is similar to multiple-words representation where a phrase's meaning and pronunciation remain intact while its written form can be of several types. Figure 3 shows this scenario, wherein three equivalent Banglish and two equivalent Bengali phrases are present and they all mean the same phrase.

Table 3.	Specific	events'	phrases.
----------	----------	---------	----------

Celebrating Event's Phrases	Protesting Event's Phrases	Religious Event's Phrases
জন্মবার্ষিকীর অনুষ্ঠান (birthday celebration), বিজয়মিছিল চলছে (victory process is going on), মিলনমেলা চলছে (The reunion is going on), আয়োজিত হবে (will be organized),জমকালো র্যালি হবে (will be splendid rally)	উন্তাল অবস্থা তৈরী (created turbulent condition), মিছিলের আহবান (call of procession), আন্দোলনের ডাক (call of movement), রাজপথে নামতে হবে (have to take the highway), প্রতিবাদ সভা হবে (will be a protest meeting)	ওয়াজ মহফিল অনুষ্ঠিত হবে (waaz mahfil will be held), জানাজা হবে (will be zanaja), দাফন করা হবে (will be buried), পূজা হবে (will be worship)
Jonmobarshikir onusthan, bijoymichil cholche, milonmela cholche, aayojit hbe, jomkalo rally hbe	Uttal obostha toiri, michiler ahoban, andoloner dak, rajpothe namte hbe, protibad shobha hbe	Owaaz mahfil onusthito hbe, janaza hbe, dafon kora hbe, puja hbe



Figure 3. An example of multiple representations of Bengali and Banglish phrases.

3.4.3. Sentiment Analysis

As a general rule, positive sentiment is correlated with celebrating events, while negative sentiment is correlated with protesting events, and positive or neutral sentiment is correlated with religious events. Sentiment score, therefore, plays an important role in event detection. To determine sentiment scores, each post's text was subjected to "vadermulti," an upgraded version of VADER, well-suited for social media datasets, that takes account for emoticons as well as varied shorthand phrases and sentence structures. Every word was assigned a score within a range of -4 through +4 according to its valence score in VADER's dictionary with -4 including the most negative words and +4 including the most positive. A post's valence score was then normalized between -1 to +1 using Equation (1), with -1 indicating an extreme negative text and +1 indicating an extreme positive text.

$$X_{compound} = \frac{X_{valance}}{\sqrt{X_{valance}^2 + \alpha}}$$
(1)

Here, $X_{compound}$ is the compound score in the range [-1, 1], $X_{valance}$ is the summed valence score in the post, and α is the constant of normalization with a default value of 15 [55].

We used compound scores to determine whether a post is positive, negative, or neutral. When the compound score was greater than or equal to +0.05, the sentiment was treated as positive. Similarly, when the compound score was less than or equal to -0.05, the sentiment was treated as negative. Sentiment was treated as neutral if the compound score was greater than -0.05 and less than +0.05.

A slightly different approach was followed to determine the sentiment scores of Bengali and Banglish posts. For Bengali posts, the post was directly applied to the "vadermulti". Banglish posts, in contrast, were converted into Bengali and then were applied to "vader-multi". The Banglish-to-Bengali conversion was performed using another python library, "bnbphoneticparser". In both cases, appropriate values were collected from the post and labeled according to the features.

Features collected from the three sub-processes were then used for model training purposes.

3.5. Model Training and Detection

The BNB model was used as the main algorithm for this task while the SVM and the DT algorithms were used for performance comparisons. For the SVM model, polynomial kernels with degree 3 were selected. For the DT, we used the Classification and Regression Trees (CART). Since we used the scikit learn package of Python and also used numerical variables, i.e., binary features values, that is why CART was used. The BNB was used since our datasets are small and this algorithm trains faster and accurately with small datasets and also provides good accuracy. The same reason also applies for selecting DT and SVM algorithms. Other methods such as Deep Learning require large training data that is not currently available to us since there is no benchmark data. Each post's extracted features were used to train the model. Actual feature values were converted into binary before they were applied to the training model. If the value of any feature was greater than or equal to 1, the value was set to 1; in all other circumstances, the value was set to 0. Equations (2) and (3) reflect the underlying process of the NB classification model.

$$P(C|F) = \frac{P(F|C) * P(C)}{P(F)}$$
⁽²⁾

Or,

$$P(C|F) = P(F_1|C) \times P(F_2|C) \times \dots P(F_n|C) \times P(C)$$
(3)

In which, P(C|F) refers to the probability of C (*Class*) being true given that F(Features) has already occurred. P(F|C) refers to the probability of F being true given that C is true. P(C) refers to the probability of C being true. P(F) refers to the probability of F being true.

Although there are three types of NB classification models (i.e., Bernoulli, Multinomial, and Gaussian), we used the BNB model as it provides superior results when working with binary features [56]. Equation (4) reflects the BNB classification model's equation:

$$p(f|C_k) = \prod_{j=1}^n p_{kj}^{f_j} \left(1 - p_{kj}\right)^{(1-f_j)}$$
(4)

In this case, $p(f|C_k)$: the likelihood of an event given a class C_K . f_j : is a Boolean feature vector that expresses the occurrence or absence of *j*th term. p_{kj} : the probability of class C_k generating the feature vector f_j .

After the model was trained with the collected features and assigned labels, the model then detected events on the basis of the associated features in the post.

4. Performance Evaluation

4.1. Performance Metrics

Precision, recall, F1-score, accuracy, Receiver Operating Characteristics (ROC), Area Under Curve (AUC), True Positive Rate (TPR), True Negative Rate (TNR), False Positive Rate (FPR), False Negative Rate (FNR), and confusion matrix were referenced as performance metrics.

TPR and TNR: True positive is an actually positive class or label that the model detected to be positive. Likewise, true negative is a negative class that the model detected to be negative.

FPR and FNR: False positives are those where the model incorrectly detected a positive class, while false negatives are incorrectly detected negative classes.

Precision: Precision is defined as the ratio of detected true positive cases to all detected positive instances. All positively identified instances also contain false positive instances.

Recall: Recall is the ratio of the detected true positive instances to all the actual positive instances. All actual instances may or may not have been detected by the model.

F1-score: The weighted average of the Precision and Recall scores is referred to as the F1-score. False positives and false negatives are accounted for in this measurement.

Accuracy: Accuracy is defined as the ratio between the total number of correct detections to all detections. Accuracy is used to measure general detection ability.

ROC: The ROC curve shows the TPR versus the FPR at different threshold values. The ROC curve identifies the probability of a true or false positive at a certain threshold value. An optimal result would be 0 FPR and 1 TPR.

AUC: AUC is defined as the ability to distinguish between classes. AUC can fall within a range of 0 to 1 for a specific class, with 1 indicating that a class is categorically identified and 0 indicating that it is never identified. An AUC closer to 1 indicates a model that classifies well.

Confusion Matrix: A confusion matrix contains all the output labels used for classification. Every diagonal value in the confusion matrix defines the true detected classes for each label, while all the other cells define the false positive labels. Precision, recall, and accuracy can be calculated by referencing this matrix.

4.2. Used Datasets

We created two separate datasets for Bengali and Banglish Facebook posts. The datasets were manually labeled by two experts. They went through each post in the datasets and labeled each post according to the types of events. Then, each of the dataset labeling was cross checked by another expert. The Bengali dataset contained 364 real Facebook posts, including four types of posts (celebrating = 47, protesting = 64, religious = 42, and neutral = 211). On the other hand, 200 real Facebook posts were collected for the Banglish dataset. The Banglish dataset also contained four types of event posts (celebrating = 56, protesting = 69, religious = 27, and neutral = 48). In both the datasets, 80% of data were used for training purposes and the remaining 20% were used for testing purposes. There were 4 celebrating, 15 protesting, 6 religious, and 48 neutral posts in the Bengali testing data. In the case of Banglish testing data, there were 10 celebrating, 16 protesting, 3 religious, and 11 neutral posts. Experimental results gained from testing data are provided in the experimental results subsection.

4.3. Experimental Results

4.3.1. Results of Our Model's Assessment of Bengali-Language Posts

Figure 4 presents the AUC-ROC curve of the BNB model as it detected events on the basis of Facebook posts written in Bengali. Celebrating (AUC = 0.93), protesting (AUC = 0.99), religious (AUC = 1.00), and neutral (AUC = 0.88) events are presented as red, blue, orange, and green lines respectively. Based on these scores, the BNB model performed well at detecting events since AUC values are closer to 1. The nearer the value of AUC to 1, the better the model's performance.

For showing the effectiveness of our performance evaluation, our event-detection model was also tested with the SVM and Decision Tree (DT) classifiers on the same dataset. Table 4 compares the performance of all three models, and includes the metrics of precision, recall, F1-score, and accuracy. In sum, the data in Table 4 show that the BNB model outperformed the other models. Tables 4 and 5 and Figures 4 and 5 depict only the evaluation of our testing data. We did not incorporate our training data while forming these tables and figures.



Figure 4. BNB classification model's AUC-ROC curve for detecting events from Bengali-language Facebook posts.

Table 4. Performances of BNB, SVM, and DT classification models in detecting events from Bengali Facebook posts.

Method	Event Type	Precision	Recall	F1-Score	Accuracy
	Celebrating	0.50	0.50	0.50	
DND	Protesting	0.83	1.00	0.91	0.0041
DIND	Religious	1.00	1.00	1.00	0.9041
	Neutral	0.96	0.90	0.92	-
	Celebrating	0.50	0.50	0.50	
CVD /	Protesting	0.83	1.00	0.91	0.97(7
SVM	Religious	1.00	0.67	0.80	0.8767
	Neutral	0.91	0.90	0.91	-
	Celebrating	0.60	0.75	0.67	
DT	Protesting	079	1.00	0.88	0.97(1
	Religious	1.00	0.67	0.80	0.8/01
	Neutral	0.93	0.88	0.90	-

Table 5 presents the true and false event-identification rates the BNB, SVM, and DT models produced. The BNB performed very well with respect to protesting, religious, and neutral events, but showed a comparatively low true-event rate for celebrating events. SVM's accuracy at detecting celebrating and neutral events was low, though it performed better for the other classes. DT showed a low true rate for neutral events but a comparatively good true rate in other event classes. Overall, the BNB performed well, achieving nearly the same accuracy at all the classes.

Figure 5 presents the Bernoulli, Multinomial, and Gaussian NB models' confusion matrices. Protesting, celebrating, neutral, and religious events are referred to by the numbers in the 1st, 2nd, 3rd, and 4th column respectively. The diagonal values in the matrices are the true events detected by the NB Classification model while other cells express the number of false-positive events. Overall, Figure 5 suggests that the BNB model outperformed the comparators.











(c) Gaussian

Figure 5. NB classification model's confusion matrices on the basis of Bengali-language posts.

	Prote	esting	Celeb	rating	Relig	gious	Neu	ıtral
Method	True Protesting	False Protesting	True Celebratig	False Celebratig	True Religious	False Religious	True Neutral	False Neutral
BNB	1.00	0.00	0.50	0.50	0.89	0.11	1.00	0.00
SVM	1.00	0.00	0.50	0.50	0.89	0.11	0.67	0.33
DT	1.00	0.00	0.75	0.25	0.87	0.13	0.67	0.33

Table 5. BNB, SVM, and DT models' true and false detection performances on every output class in detecting events from Bengali Facebook posts.

The BNB classification was 90.41% accurate at detecting events on the basis of Bengalilanguage Facebook posts, while the SVM and DT models were only 87.67% and 87.61% accurate, respectively.

We performed 10-fold cross validation of our event-detection model. Table 6 shows the results of 10-fold cross validation along with their standard deviation for BNB, SVM, and DT classifiers for the Bengali language. The values of the table are approximately same as our average result, which means our model works uniformly for all the data.

Table 6. The 10-fold cross validation's result of Bengali Facebook posts along with their standard deviation.

	Method		S	tandard Deviatio	on
BNB	SVM	DT	BNB	SVM	DT
0.9041	0.8767	0.8904			
0.8767	0.8904	0.8904	_		
0.8904	0.8493	0.8904	_		
0.8630	0.8630	0.8767	_		
0.8082	0.8219	0.8356	0.027707	0.023926	0.00.4000
0.9041	0.8767	0.8767			0.024082
0.8493	0.8630	0.8630	_		
0.8767	0.8630	0.8767	_		
0.8904	0.8767	0.8356	_		
0.89041	0.9178	0.9178	_		
0.8753	0.8698	0.8753		Average	

Table 7 shows the comparison among our current work, previous work [30], and some similar works of the literature review. Our work showed better results than the works mentioned above in terms of F-score by a significant margin. Besides, this comparison also shows the improvement from our previous event detection task.

4.3.2. Results of Our Model's Assessment of Banglish-Form Posts

Figure 6 presents the AUC-ROC curve of BNB model as it detected events on the basis of Banglish-form Facebook posts. Celebrating events are denoted by a red line, protesting events are denoted by a blue line, religious events are denoted by an orange line, and neutral events are denoted by a green line in Figure 6. The AUC values for these events were celebrating (0.92), protesting (0.92), religious (0.98), and neutral (0.73). The closeness of these values to 1 confirms that the model detected events well.

Author Name with Reference	F-Score	
Sakaki et al. [8]	73.69	
Alomari et al. [16]	83	
Alsaedi et al. [18]	80.24	
Dey et al. [30]	82.5	
Our approach	92	

Table 7. Comparison of our approach and other approaches in terms of F-scores.



Figure 6. BNB classification model's AUC-ROC curve in detecting events from Banglish-form Facebook posts.

The performances of the SVM and DT models were also assessed using the same dataset for demonstrating the efficiency of our model, with Table 8 comparing the precision, recall, F1-score, and accuracy of all three models. Table 8 reflects the superior performance of the BNB model.

Method	Event Type	Precision	Recall	F1-Score	Accuracy
	Celebrating	0.64	0.70	0.67	
	Protesting	0.78	0.88	0.82	0.70
BINB	Religious	0.75	1.00	0.86	0.70
	Neutral	0.57	0.36	0.44	
	Celebrating	0.60	0.60	0.60	
C1714	Protesting	0.85	0.69	0.75	0.65
SVM	Religious	0.75	1.00	0.86	0.65
	Neutral	0.46	0.55	0.50	
DT	Celebrating	0.54	0.70	0.61	
	Protesting	0.79	0.94	0.86	0.47
	Religious	0.75	1.00	0.86	0.67
	Neutral	0.50	0.18	0.27	

Table 8. Performances of BNB, SVM, and DT classification models in detecting events from Banglish Facebook posts.

Table 9 presents the true and false event-identification rates of all three models. The

BNB model performed very well with respect to protesting, religious, and celebrating event classes, but showed a comparatively low true event rate in neutral events. SVM was most accurate at detecting religious events but demonstrated a low true rate in other classes. DT showed a high true rate for celebrating, religious, and protesting events, but a comparatively low rate for neutral events. Overall, the BNB model performed well at detecting every class of event. Tables 8 and 9 and Figures 6 and 7 present the outcome of our testing data, excluding our training data.

Table 9. BNB, SVM, and DT models' true and false detection performances on every output class in detecting events from Banglish Facebook posts.

	Prote	esting	Celeb	rating	Relig	gious	Neı	ıtral
Method	True Protesting	False Protesting	True Celebrating	False Celebrating	True Religious	False Religious	True Neutral	False Neutral
BNB	0.87	0.13	0.70	0.30	1.00	0.00	0.38	0.62
SVM	0.69	0.31	0.60	0.40	1.00	0.00	0.55	0.45
DT	0.93	0.07	0.70	0.30	1.00	0.00	0.19	0.81

Figure 7 presents the Bernoulli, Multinomial, and Gaussian NB models' confusion matrices. The 1st, 2nd, 3rd, and 4th columns refer to the protesting, celebrating, neutral, and religious events, respectively. The diagonal cells reflect accurately detected events and the remaining cells reflect false positive detections. On the whole, the BNB model outperformed the two other NB models according to Figure 7.

The BNB classification model was 70.0% accurate at detecting events using Banglishform Facebook posts. SVM and DT, in contrast, were 65.0% and 67.0% accurate, respectively.

The 10-fold cross validation for the Banglish language was also performed and is presented in Table 10. This table not only shows the result of 10-fold cross validation, but also shows the result of standard deviation and the average value of the results. From this table, we can state that our detection model uniformly works for all the data values, as the results of this table are approximately same as our average result.



Figure 7. Cont.



Figure 7. NB classification model's confusion matrices in detecting events on the basis of Banglishform posts.

Table 10. The 10-fold cross validation's result of Banglish Facebook posts along with their standard deviation.

	Method		S	Standard Deviation	on
BNB	SVM	DT	BNB	SVM	DT
0.675	0.675	0.65			
0.7	0.675	0.675			
0.725	0.675	0.65			
0.75	0.725	0.75			
0.65	0.625	0.625			
0.8	0.725	0.725	0.06	0.061033	0.070755
0.775	0.75	0.675			
0.85	0.825	0.85			
0.7	0.675	0.625			
0.675	0.6	0.6			
0.73	0.69	0.68		Average	

Table 11 shows the results of the statistical tests that were performed on the datasets using STAC [57]. In terms of the Bengali dataset, the overall *p*-value was 0.97639 and the H0 was accepted. In all other comparisons with SVM and DT, the H0 was accepted. This suggests that our result was always close to the average result. Similarly, in the Banglish

dataset, the *p*-value was 1.0 and H0 was accepted in all the cases. This result, again, suggests that the result was close to the average result.

Datasets –	ANOVA between Cases Test			Bonferroni-Dunn Test			
	Statistic	<i>p</i> -Value	Result	Compare	Statistic	<i>p</i> -Value	Result
Bengali		0.97639	H0 is accepted	NB vs. SVM	0.18943	1.00000	H0 is accepted
	0.02392			SVM vs. DT	0.18940	1.00000	H0 is accepted
				NB vs. DT	0.00003	1.00000	H0 is accepted
Banglish –0.236		23622 1.00000	H0 is accepted	NB vs. SVM	0.48853	0.94368	H0 is accepted
	-0.23622			SVM vs. DT	0.17447	1.00000	H0 is accepted
				NB vs. DT	0.66300	0.76943	H0 is accepted

Table 11. Statistical tests performed on the datasets.

Our proposed approach worked well in detecting events from Bengali and Banglish Facebook posts. It worked better since we selected BNB as our main algorithm, and the features of our method were binary in nature. Moreover, we detected suitable common-event words and phrases as well as event-specific words and phrases, which better captured the events. Besides, sentiment analysis of the Bengali and Banglish posts expedited our eventdetection task.

Once the event is detected by this model, the authorities can keep an eye on that particular event and keep up with the pre-preparation. If there is no event, there is no need for precautionary action. Thus, this detection model can help the authorities to ensure social security.

5. Conclusions, Limitations and Future Scope

5.1. Conclusions

In this paper, we proposed an event-detection scheme that analyzes Bengali and Banglish-form Facebook posts and detects events as celebrating, protesting, religious or neutral types. For this purpose, we extracted Facebook posts from different popular public Facebook pages and groups along with various public and private university groups and pages. Collected posts were first checked for language detection that considers only Bengali and Banglish posts and discards other language's posts. Detected posts were then pre-processed and then features were collected. The collected features were used for the training process and then proceeded for event detection. In this event detection work, common and event-specific words and phrases were detected for both Bengali and Banglish. Besides, words' and phrases' multiple representations were also recognized. We acquired satisfactory results employing this model. We achieved an accuracy of 90.41% for Bengali and 70.0% for the Banglish posts. We employed the BNB model for this detection process and used binary feature values for which BNB worked better in detecting events. Besides, thoroughly recognizing common event words and specific event words and phrases bolstered this model's performance. Another important factor in this good accuracy is the recognition of multiple representations of words and phrases of Bengali and Banglish. Without these multiple representations, we previously achieved the wrong feature values and ultimately detected the wrong types of events. This recognition helps in identifying the same words or phrases in multiple ways and also provides proper feature values.

5.2. Existing Limitations and Future Scope

Because of the unavoidable COVID-19 situation, we collected our data manually rather than using any automated tool such as a web crawler. In the future, we intend to develop our own web crawler for collecting data. We also expect to work with a larger dataset. As of this paper, we mainly considered Bangladeshi peoples' perspectives, and we performed our experiments using Bengali and Banglish-form posts. Our future aim is to build a language-independent model that will provide large-scale service to any other dialects. Moreover, we would like to enlarge our event-specific word and phrase list in the future so that we can achieve better accuracy for the Banglish-form language also. We also wish to attempt to incorporate event-location detection into our model.

Supplementary Materials: The following are available online at https://www.mdpi.com/article/10. 3390/electronics10192367/s1, Document S1: "A dataset for event detection from Bengali and Banglish Facebook posts."

Author Contributions: Conceptualization, N.D., M.S.R. and M.S.M.; methodology, N.D., M.S.R. and M.S.M.; software, N.D. and M.S.M.; validation and formal analysis, N.D., M.S.R. and M.S.M.; writing—N.D., M.S.R. and M.S.M.; writing—review and editing, A.S.M.S.H. and M.S.R.; visualization, N.D., M.S.R. and M.S.M.; supervision, M.S.R.; funding acquisition, A.S.M.S.H. and I.-H.R. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the Korean Institute of Energy Technology Evaluation and Planning (KETEP), Korean Government, Ministry of Trade, Industry, and Energy (MOTIE), under Grant 20194010201800, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea Government [Ministry of Science and ICT (MSIT)], under Grant 2021R1A2C2014333.

Data Availability Statement: Self-made data attached as a Supplementary Document.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations and symbols are used in this manuscript:

AVG	Average
AUC	Area Under Curve
BNB	Bernoulli Naive Bayes
CRF	Conditional Random Field
DT	Decision Tree
FPR	False Positive Rate
FNR	False Negative Rate
HITS	Hypertext Induced Topic Search
LSTM	Long Short-Term Memory
LDA	Latent Dirichlet Allocation
LSH	Locality Sensitive Hashing
MLP	Multi-Layer Perceptron
MTL	Multi-Task Learning
MC	Microblog Clique
NB	Naive Bayes
NER	Name Entity Recognition
PWE	Personal Wellness Events
ROC	Receiver Operating Characteristics
SVM	Support Vector Machine
TD-HITS	Topic Decision HITS
TPR	True Positive Rate
TNR	True Negative Rate
TS-LDA	Three Step LDA
VADER	Valence Aware Dictionary and Sentiment Reasoner

References

- 1. Taylor, D.B. The New York Times. Available online: https://web.archive.org/web/20200602235547/https://www.nytimes.com/ar ticle/george-floyd-protests-timeline.html (accessed on 8 May 2021).
- Robinson, K. Council on Foreign Relations. Available online: https://www.cfr.org/article/arab-spring-ten-years-whats-legacy -uprisings (accessed on 2 April 2021).

- 3. The Economist. Available online: https://www.economist.com/asia/2018/04/21/protests-in-bangladesh-put-an-end-to-a-corru pt-quota-system (accessed on 20 March 2021).
- 4. Firstplot. Available online: https://www.firstpost.com/world/students-end-protests-on-road-safety-in-bangladesh-after-ninedays-education-ministry-to-hold-meet-tomorrow-4913421.html (accessed on 27 March 2021).
- 5. Anantharam, P.; Barnaghi, P.; Thirunarayan, K.; Sheth, A. Extracting city traffic events from social streams. *ACM Trans. Intell. Syst. Technol.* **2015**, *6*, 1–27. [CrossRef]
- 6. Alomari, E.; Mehmood, R.; Katib, I. Sentiment analysis of Arabic tweets for road traffic congestion and event detection. In *Smart Infrastructure and Applications*; Springer: Cham, Switzerland, 2020; pp. 37–54. [CrossRef]
- Imran, M.; Elbassuoni, S.; Castillo, C.; Diaz, F.; Meier, P. Practical extraction of disaster-relevant information from social media. In Proceedings of the 22nd International Conference on World Wide Web, Rio de Janeiro, Brazil, 13–17 May 2013; pp. 1021–1024. [CrossRef]
- 8. Sakaki, T.; Okazaki, M.; Matsuo, Y. Earthquake shakes twitter users: Real-time event detection by social sensors. In Proceedings of the 19th International Conference on World Wide Web, Raleigh, NC, USA, 26–30 April 2010; pp. 851–860. [CrossRef]
- 9. Fathima, P.N.; George, A. Event detection and text summary by disaster warning. Int. Res. J. Eng. Technol. 2019, 6, 2510–2513.
- 10. Ristea, A.; Al Boni, M.; Resch, B.; Gerber, M.S.; Leitner, M. Spatial crime distribution and prediction for sporting events using social media. *Int. J. Geogr. Inf. Sci.* 2020, 34, 1708–1739. [CrossRef]
- Fedoryszak, M.; Frederick, B.; Rajaram, V.; Zhong, C. Real-time event detection on social data streams. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Anchorage, AK, USA, 4–8 August 2019; pp. 2774–2782. [CrossRef]
- 12. Ahmad, Z.; Varshney, D.; Ekbal, A.; Bhattacharyya, P. *Multi-Lingual Event Identification in Disaster Domain*; Indian Institute of Technology Patna: Bihta, India, 2019.
- 13. Shi, K.; Gong, C.; Lu, H.; Zhu, Y.; Niu, Z. Wide-grained capsule network with sentence-level feature to detect meteorological event in social network. *Future Gener. Comput. Syst.* **2020**, *102*, 323–332. [CrossRef]
- 14. Ali, D.; Missen, M.M.S.; Husnain, M. Multiclass Event Classification from Text. Sci. Program. 2021, 2021, 6660651. [CrossRef]
- 15. Choi, D.; Park, S.; Ham, D.; Lim, H.; Bok, K.; Yoo, J. Local Event Detection Scheme by Analyzing Relevant Documents in Social Networks. *Appl. Sci.* **2021**, *11*, 577. [CrossRef]
- Alomari, E.; Katib, I.; Mehmood, R. Iktishaf: A big data road-traffic event detection tool using Twitter and spark machine learning. *Mob. Netw. Appl.* 2020, 1–16. [CrossRef]
- Jain, A.; Kasiviswanathan, G.; Huang, R. Towards accurate event detection in social media: A weakly supervised approach for learning implicit event indicators. In Proceedings of the 2nd Workshop on Noisy User-Generated Text (WNUT), Osaka, Japan, 11 December 2016; pp. 70–77.
- Alsaedi, N.; Burnap, P. Arabic event detection in social media. In Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics, Cairo, Egypt, 14–20 April 2015; Springer: Cham, Switzerland, 2015; pp. 384–401. [CrossRef]
- Suma, S.; Mehmood, R.; Albeshri, A. Automatic event detection in smart cities using big data analytics. In Proceedings of the International Conference on Smart Cities, Infrastructure, Technologies and Applications, Jeddah, Saudi Arabia, 27–29 November 2017; Springer: Cham, Switzerland, 2017; pp. 111–122. [CrossRef]
- 20. Cui, W.; Wang, P.; Du, Y.; Chen, X.; Guo, D.; Li, J.; Zhou, Y. An algorithm for event detection based on social media data. *Neurocomputing* **2017**, *254*, 53–58. [CrossRef]
- Gao, Y.; Zhao, S.; Yang, Y.; Chua, T.S. Multimedia social event detection in microblog. In Proceedings of the International Conference on Multimedia Modeling, Sydney, NSW, Australia, 5–7 January 2015; Springer: Cham, Switzerland, 2015; pp. 269–281. [CrossRef]
- 22. StatCounter GlobalStats. Available online: https://gs.statcounter.com/social-media-stats/all/bangladesh (accessed on 1 March 2021).
- Statista. Available online: https://www.statista.com/statistics/268136/top-15-countries-based-on-number-of-facebook-users/ (accessed on 25 January 2021).
- 24. Mumu, T.F.; Munni, I.J.; Das, A.K. Depressed people detection from bangla social media status using lstm and cnn approach. *J. Eng. Adv.* **2021**, *2*, 41–47. [CrossRef]
- 25. Das, A.K.; Al Asif, A.; Paul, A.; Hossain, M.N. Bangla hate speech detection on social media using attention-based recurrent neural network. *J. Intell. Syst.* 2021, *30*, 578–591. [CrossRef]
- Rozen, A. Twitter Blog. Available online: https://blog.twitter.com/official/en_us/topics/product/2017/tweetingmadeeasier.html (accessed on 25 March 2021).
- 27. Sharmin, S.; Chakma, D. Attention-based convolutional neural network for Bangla sentiment analysis. *AI Soc.* 2021, *36*, 381–396. [CrossRef]
- 28. Rahman, M.; Haque, S.; Saurav, Z.R. Identifying and categorizing opinions expressed in bangla sentences using deep learning technique. *Int. J. Comput. Appl.* **2020**, *975*, 8887. [CrossRef]
- 29. Alam, T.; Khan, A.; Alam, F. Bangla Text Classification using Transformers. arXiv 2020, arXiv:2011.04446.

- Dey, N.; Mredula, M.S.; Sakib, M.N.; Islam, M.N.; Rahman, M.S. A Machine Learning Approach to Predict Events by Analyzing Bengali Facebook Posts. In Proceedings of the International Conference on Trends in Computational and Cognitive Engineering, Dhaka, Bangladesh, 17–18 December 2021; Springer: Berlin/Heidelberg, Germany, 2021; pp. 133–143.
- Chen, G.; Kong, Q.; Mao, W. Online event detection and tracking in social media based on neural similarity metric learning. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 182–184. [CrossRef]
- 32. Bekoulis, G.; Deleu, J.; Demeester, T.; Develder, C. Sub-event detection from twitter streams as a sequence labeling problem. *arXiv* **2019**, arXiv:1903.05396.
- 33. Aldhaheri, A.; Lee, J. Event detection on large social media using temporal analysis. In Proceedings of the 2017 IEEE 7th Annual Computing and Communication Workshop and Conference (CCWC), Las Vegas, NV, USA, 9–11 January 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1–6. [CrossRef]
- Li, Q.; Nourbakhsh, A.; Shah, S.; Liu, X. Real-time novel event detection from social media. In Proceedings of the 2017 IEEE 33rd International Conference on Data Engineering (ICDE), San Diego, CA, USA, 19–22 April 2017; IEEE: Piscataway, NJ, USA, 2017; pp. 1129–1139. [CrossRef]
- 35. Kannan, J.; Shanavas, A.M.; Swaminathan, S. Sportsbuzzer: Detecting events at real time in twitter using incremental clustering. *Trans. Mach. Learn. Artif. Intell.* **2018**, *6*, 1.
- Feng, X.; Zhang, S.; Liang, W.; Liu, J. Efficient location-based event detection in social text streams. In Proceedings of the International Conference on Intelligent Science and Big Data Engineering, Suzhou, China, 14–16 June 2015; Springer: Cham, Switzerland, 2015; pp. 213–222. [CrossRef]
- Arachie, C.; Gaur, M.; Anzaroot, S.; Groves, W.; Zhang, K.; Jaimes, A. Unsupervised detection of sub-events in large scale disasters. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; Volume 34, pp. 354–361. [CrossRef]
- Pekar, V.; Binner, J.; Najafi, H.; Hale, C.; Schmidt, V. Early detection of heterogeneous disaster events using social media. J. Assoc. Inf. Sci. Technol. 2020, 71, 43–54. [CrossRef]
- Akbari, M.; Hu, X.; Liqiang, N.; Chua, T.S. From tweets to wellness: Wellness event detection from twitter streams. In Proceedings of the AAAI Conference on Artificial Intelligence, Phoenix, AZ, USA, 12–17 February 2016; Volume 30.
- 40. Panagiotou, N.; Katakis, I.; Gunopulos, D. Detecting events in online social networks: Definitions, trends and challenges. In *Solving Large Scale Learning Tasks. Challenges and Algorithms*; Springer: Cham, Switzerland, 2016; pp. 42–84. [CrossRef]
- Kolya, A.K.; Ekbal, A.; Bandyopadhyay, S. A simple approach for Monolingual Event Tracking system in Bengali. In Proceedings of the 2009 Eighth International Symposium on Natural Language Processing, Bangkok, Thailand, 20–22 October 2009; IEEE: Piscataway, NJ, USA, 2009; pp. 48–53. [CrossRef]
- 42. Zhao, S.; Gao, Y.; Ding, G.; Chua, T.S. Real-time multimedia social event detection in microblog. *IEEE Trans. Cybern.* 2017, 48, 3218–3231. [CrossRef] [PubMed]
- 43. Shi, L.; Wu, Y.; Liu, L.; Sun, X.; Jiang, L. Event detection and identification of influential spreaders in social media data streams. *Big Data Min. Anal.* **2018**, *1*, 34–46. [CrossRef]
- 44. Nurwidyantoro, A.; Winarko, E. Event detection in social media: A survey. In Proceedings of the International Conference on ICT for Smart Society, Jakarta, Indonesia, 13–14 June 2013; IEEE: Piscaaway, NJ, USA, 2013; pp. 1–5. [CrossRef]
- 45. Zarrinkalam, F.; Bagheri, E. Event identification in social networks. *Encycl. Semant. Comput. Robot. Intell.* 2017, 1, 1630002. [CrossRef]
- Dou, W.; Wang, X.; Ribarsky, W.; Zhou, M. Event detection in social media data. In Proceedings of the IEEE VisWeek Workshop on Interactive Visual Text Analytics-Task Driven Analytics of Social Media Content, Seattle, WA, USA, 14–19 October 2012; IEEE: Piscataway, NJ, USA, 2012; pp. 971–980.
- 47. Said, N.; Ahmad, K.; Riegler, M.; Pogorelov, K.; Hassan, L.; Ahmad, N.; Conci, N. Natural disasters detection in social media and satellite imagery: A survey. *Multimed. Tools Appl.* **2019**, *78*, 31267–31302. [CrossRef]
- 48. Saeed, Z.; Abbasi, R.A.; Maqbool, O.; Sadaf, A.; Razzak, I.; Daud, A.; Aljohani, N.R.; Xu, G. What's happening around the world? A survey and framework on event detection techniques on twitter. *J. Grid Comput.* **2019**, *17*, 279–312. [CrossRef]
- 49. Yu, M.; Bambacus, M.; Cervone, G.; Clarke, K.; Duffy, D.; Huang, Q.; Li, J.; Li, W.; Li, Z.; Liu, Q.; et al. Spatiotemporal event detection: A review. *Int. J. Digit. Earth* 2020, 13, 1339–1365. [CrossRef]
- 50. Zhou, D.; Huang, J.; Schölkopf, B. Learning with hypergraphs: Clustering, classification, and embedding. *Adv. Neural Inf. Process. Syst.* **2006**, *19*, 1601–1608.
- 51. Akaike, H. Information theory and an extension of the maximum likelihood principle. In *Selected Papers of Hirotugu Akaike;* Springer: New York, NY, USA, 1998; pp. 199–213.
- 52. Pypi. Available online: https://pypi.org/project/langdetect/?fbclid=IwAR17pzcUCVFUaWi7PMLHOiD7pqjYhX7rew_DTxSLXX FBKJdGmes6V3qooyU (accessed on 2 January 2021).
- 53. Hutto, C.; Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the International AAAI Conference on Web and Social Media, Ann Arbor, MI, USA, 1–4 June 2014; IEEE: Piscataway, NJ, USA, 2014; Volume 8.

- 54. GitHub. Available online: https://github.com/porimol/bnbphoneticparser?fbclid=IwAR2bXVZioSZyVaijKoIXE8srOEtyhycFmc aTsL88zWnprNhbrRXY4J2NxpY (accessed on 5 January 2021).
- 55. QuantInsti. Available online: https://blog.quantinsti.com/vader-sentiment/#:~{}:text=Compound\$\delimiter"026E30F\$%20VA DER\$\delimiter"026E30F\$%20scores\$\delimiter"026E30F\$%20for\$\delimiter"026E30F\$%20analyzing,1\$\delimiter"026E30 F\$%20(most\$\delimiter"026E30F\$%20extreme\$\delimiter"026E30F\$%20positive) (accessed on 10 March 2021).
- 56. Analytics Vidhya. Available online: https://www.analyticsvidhya.com/blog/2017/09/naive-bayes-explained/ (accessed on 12 August 2021).
- 57. Rodríguez-Fdez, I.; Canosa, A.; Mucientes, M.; Bugarín, A. STAC: A web platform for the comparison of algorithms using statistical tests. In Proceedings of the 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Istanbul, Turkey, 2–5 August 2015; IEEE: Piscataway, NJ, USA, 2015; pp. 1–8.