Combinatorial Approach of Feature Generation for Traffic Event Detection using Social Media Data: CFGA

Narayan Chaturvedi (narayanchaturvedi@gmail.com)
Indian Institute of Technology Roorkee

Durga Toshniwal
Indian Institute of Technology Roorkee

Manoranjan Parida
Indian Institute of Technology Roorkee

Research Article

Keywords: Machine Learning, Feature Generation, Traffic Data, Congestion Detection, Twitter

Posted Date: September 27th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-876469/v1

Combinatorial Approach of Feature Generation for Traffic Event Detection using Social Media Data: CFGA

Narayan Chaturvedi[1*], Durga Toshniwal[2†] and Manoranjan Parida[3†]

[1*]CTRANS, Indian Institute of Technology Roorkee, Roorkee, 247667, Uttarakhand, India.
[2]CSE, Indian Institute of Technology Roorkee, Roorkee, 247667, Uttarakhand, India.
[3]CE, Indian Institute of Technology Roorkee, Roorkee, 247667, Uttarakhand, India.

*Corresponding author(s). E-mail(s): narayanchaturvedi@gmail.com;
Contributing authors: durgatoshniwal@cs.iitr.ac.in; mparida@ce.iitr.ac.in;
†These authors contributed equally to this work.

Abstract

Confined use of road sensors limits the effectiveness of traffic disturbing event detection. In this context, Twitter is becoming popular among the people to share the events that affect the daily life. In this paper, a novel dictionary formation and a new feature generation approach is proposed to build an integrated machine learning framework to detect the traffic events. The proposed novel combinatorial feature generation approach (CFGA) uncovers appropriate associations among the keywords of tweet and extracts the correlated keyword sets to the data collected. Such keyword sets are denoted as *set phrases*. The *set phrases* may comprise of single or multiple words of a tweet. These *set phrases* may be used as keywords for event-related data collection or further analysis. The frequently occurring *set phrases* are identified using the notion of support, which signifies the percentage of tweets containing relevant keywords. Since the nature of different events may also vary; therefore, a hardcoded value for support threshold will not be beneficial. Hyper-parameter designated as support ($\boldsymbol{\omega}$) is tuned for finding threshold value that is used to obtain the *set phrases*. This process sets up a database of frequently occurring *set phrases* that can signalize the traffic-related events using ML classifier. The results of the proposed approach suggest that if suitable support is chosen, proposed CFGA increases the accuracy of supervised classification models for extracting traffic information from Twitter data. The classification results obtained by using the proposed approach outperform their existing counterparts in terms of precision, recall, and F-measure.

Keywords: Machine Learning, Feature Generation, Traffic Data, Congestion Detection, Twitter

1 Introduction

Traffic event identification in real-time is always a significant challenge in urban transport maintenance. Many countries are using road sensors and cameras for maintaining the smooth flow of city traffic. But due to the high cost of installation and maintenance of road sensors and other similar electronic devices, traffic data production and its processing become expensive. However, social media platforms and microblogging websites like Facebook, Twitter, etc. are becoming popular among the people to share the events and things which affect their daily life. In India, a tremendous increase in the number of social media users has been observed in recent years. They are using these platforms for expressing concerns that affect their daily lives. Microblogging service like Twitter has 328 million active users every month who tweet at 230000 messages per minute [1] thus; it has been proven a economical data source for collecting the large amount of real-time data.

Detection of transport service disruption, travel complaint resolution, seasonal information to travellers, and traffic event detection are the important applications of social media for transportation [2]. Researchers evaluated the potential of Twitter data for transit customer satisfaction [3] and studied traffic incident detection using the tweets containing road events [4]. To characterize the traffic events, relevant information extraction from social media data is the essential step. In text mining, words of short messages are the phrases or tokens, which are the basic unit of features [5] and these phrases are the key to knowledge discovery from social media messages. Therefore, determining relevant features and pruning unwanted ones is the first step in the direction of knowledge discovery from such vast text datasets. Due to tweets' noisy and unstructured nature, pruning of irrelevant features plays an essential role in feature selection. Most tweets are not related to events; thus, the information we seek is a challenge for text classification.

The majority of the existing approaches use machine learning (ML) techniques for event detection using social media data and represent text as a

Bag-of-Words (BOW); namely, the features in document vectors represent the weighted event frequencies of individual words [6]. In addition, the feature generation approach directly affects the performance of text classification using ML based models. BOW is the popular textual feature selection technique for machine learning text classification models [7]. This is very effective when categorized by very easily identifiable keywords. However, it does not show good performance for demanding tasks such as traffic event detection. Implementation of the classical Term Frequency-Inverse Document Frequency (TF-IDF) [8] strategy is also based on the BOW model. These models take the multiple sentences as the input and produce the sparse vectors (sparse matrix) as the output. In the next step, this output will be the input for machine learning models. The large sparse matrix makes the process of model training computationally inefficient. BOW models only care about the presence of words or word frequency in the text document. But semantics, context or ordering of words are not taken into considerations while developing BOW models. Further, the limitations of the BOW model can be written in two points: 1. It completely ignores semantics and doesn't preserve the context. 2. The bow approach needs enormously high computation time and large memory for big textual data.

In knowledge discovery, combinations of frequent keywords are highly informative than single frequent keywords. For example, frequent keyword "*congestion*" does not clearly indicate road traffic jams, but a combination of consecutive keywords such as "*road congestion*" explicitly states for road traffic jam conditions. A tweet like "*jiocare is taking follows up from last2 years. But no improvement in connectivity and congestion at m...*" contains the keyword "*congestion*". But the tweet doesn't belong to the study and irrelevant to our research. TF-IDF is useful for lexical level keyword extraction. Still, it cannot capture semantics and fails to determine the collection of frequent keywords with high associativity as a single token set.

Human-written tweets are relevant to a topic. Still, while applying BOW and TF-IDF models, the tweets have been disambiguating into individual terms, and the terms are collected in a bin for further steps of sparse matrix creation. This approach of disambiguation loses the context in terms of human-written tweets. Our proposed method, however, does not lose the context of the terms in the tweets. Instead of disambiguating the tweets into individual terms, it directly extracts the frequent terms as immutable frozen-sets from the human written paraphrases using an Apriori-based approach. The frequent terms extracted from the human-written tweets include a sense of the subject on which it is written. In this paper, the immutable frozen-sets extracted from the human written paraphrases are denoted as set phrases. The set phrases may be a single term or combination of correlated terms of a tweet.

One of the significant challenges in characterizing traffic events from Twitter data is candidate phrase selection. An efficient method for phrase selection will improve the overall performance of the traffic event detection task. This paper presents a strategy for creating the database of frequent set-phrases, which can be further utilized in the training machine learning model. In this paper, our objective is to empower machine learning techniques for text classification with broader knowledge for traffic event characterization work. We wish to identify the kind of words a tweet contains, which intuitively reveals the topic of tweets whether it contains any traffic event. Just like the BOW and TF-IDF model, which uses frequency and weight values as a numeric value of phrases, the proposed approach CFGA utilizes support value to represent the phrases. Therefore, the models like BOW and TF-IDF have been used to compare the results of the CFGA for traffic data characterization from Twitter data. However, few unsupervised models such as word2vec and GloVe [9] word embeddings generate word vectors, that require neural network for training and label predictions. Word embeddings represent the same representation for similar keywords that find n-dimensional vectors for each word. Word embedding models are very effective in many text analysis tasks. However, such a corpus-based word embedding method fails to record the exact meanings of several words accurately. Such as synonym and antonym words are semantically related, but they can be different words in the same contexts. Therefore, these word embedding models may lead to some synonym word vectors much closer in learned embedding vector than many antonym words.

Moreover, if two semantically similar keywords have different frequencies in the corpus, then these word embeddings do not catch the precise relationships in the learned embedding space [10]. In this paper, a domain-specific dataset (In this paper, the context of the dataset is Traffic-related) is used, which means that we may not need to generate corresponding embedding vector space from trained words (like in word2vec, etc.). An example of the frequently used same word in different context: ”Crash” can be used for a road accident, in the stock market crash, for the sound of cymbals happening simultaneously or to attend a dinner party without an invitation, etc. Word embeddings come at the cost of having memory-intensive and more complex implementation.

The novelty and main contributions of our work are summarized below.

- The proposed CFGA approach disambiguates the tweets into set phrases without losing the sense of subject for which it is written. In addition, the concept of choosing frequently occurring set phrases reduces the feature count that results in a reduced memory requirement.
- Based on the degree of associativity among the tokens, feature set phrases are generated from the preprocessed tweets. Further, the set phrases vector is classified to describe the written text. In order to tune the parameter *min_supp* for selecting candidate set phrases, we use three different Twitter datasets collected using different traffic related keywords. The results present that the proposed method improves the accuracy of machine learning-based text classification models.
- We evaluate the gain of empowering inductive learning techniques, using the task of text classification for the transport domain. Empirical evaluation proves a significant improvement of event detection results in the host of transportation-related Twitter datasets.

Table 1: Collected Tweets' Sample

Sl. no	Keyword	Tweet
1	"bus"	police 'kill 40 militants' in raids after tourist bus blast
2	"road"	Hello there. We are picking up delays in the area of the ring road.
3	"fog"	The fog today crazy jewertoll
4	"traffic"	Chandni chowk was full of vehicles, traffic nightmare, illegal parking and vendors capturing stretch meant for pedestrian. B...
5	"road, congestion "	Illegal parking of double lane on the main road between netaji subash place towards pitampura metro station causing congestion.

- A dictionary containing main traffic-based keywords popular in social media and in daily communication of citizens has been created.

The rest of the paper is organized as follows. Section 2 covers the literature survey part. Section 3 describes the preliminaries concepts used in the study. Raw data collection and preprocessing describes in section 4 and section 5 respectively. In section 6, details about basic notations, problem formulation, hyper-parameters, and proposed approach have been described. Section 7 contains results and discussion. The last section finally concludes the whole study.

2 Literature Survey

In this section, literature related to the application of social media data in transportation has been reviewed. Micro-blogging platform Twitter provides broad coverage at an economical way. Therefore, Twitter data has been utilized using various data mining and machine learning tools for various research areas like urban planning, traffic incident identification, etc. Further, different feature extraction and selection approaches have also been reviewed because features defined from the tokens contained in the collected text data impact the traffic events identification task's performance. The related work has been studied in two phases as social media based transportation work and feature generation of text documents for the machine learning model.

2.1 Social interactions in Transportation

This section discusses the brief review of Twitter data harnessing for transportation. Twitter is an extremely popular microblogging platform that allows people to write and share short messages among digital users. In addition to desktop web browsing, Twitter is also on the mobile web, which makes Twitter users portable while posting messages. This conveniences of this platform make it most popular in sharing urgent and time-sensitive messages. Twitter messages often contain daily happening events with location data, that suits well in identifying traffic events. Recently, Twitter data harnessing for various transport-related applications like incident detection, accessing transport information [11], travel behavior [12], [13] and mobility pattern analysis [14], [15], [16] [17], [18], [19], [20], etc. are widely popular. In [21], authors detected the traffic accidents using geotagged Twitter data. In [22], authors presented a broad literature survey on mobility pattern mining from social media data. ML models have shown great promises in text classification [23]. [24] proposed a system that monitors transportation-related incidents and events using Twitter data. In their paper, the method utilizes machine learning algorithms for clustering and classification of tweets. Further, the paper also concluded that Twitter is a more robust data source in identifying traffic incidents and events quickly than other data sources.

In [25], Support vector machine was implemented to classify traffic-related and non-traffic related tweets and in [26], Naive Bayes classifier was applied to classify traffic and non-traffic tweets, and further Latent Dirichlet Allocation classifier was also incorporated to categorize the traffic incidents in traffic-related tweets. Authors applied bag-of-words for the feature selection from the tweets written in Italian language and used support vector machines classifier to perform binary as traffic-related tweets or non-traffic related tweets and then 3-class classification over traffic-related tweets as congestion, crash or traffic due to an external event [27]. Paper [28] addressed the challenges that occur in detecting events from Twitter data and presented a survey of event detection techniques from Twitter data streams. Authors developed a comprehensive approach to identify traffic-related tweets using historical tweets and further applied tf-idf to generate the influential word set [29]. In [13], clustering was applied to group the tweet locations which infer the travel behavior. The results were compared with traditional household survey methods and verify the potential of social media data for travel behavior studies.

2.2 Approaches for the selection of relevant features

Previous studies related to relevant textual feature selection have provided assistance and motivation for this work. BOW and TF-IDF are prevailing modeling techniques in the text representation.

In [30], the authors proposed an approach to detect events by analyzing their temporal and textual elements. The time span had been divided in to intervals and tweets were categorized accordingly. Frequent features were extracted from the tweets at particular time intervals, and if the frequency of some features increases at contiguous time intervals, then those words were considered as event representative keywords. The entities and relations to domain specific Twitter data are also identified to facilitate information retrieval [31]. In [21], the authors applied two different feature selection strategies for individual token features and paired token features related to traffic accident. To select individual token features phi coefficient is used that measures the correlation between label and token [32]. The paired token features had been chosen based on association rule established between labels and stemmed token. Latent Semantic Indexing (LSI) was proposed for retrieving textual information from large text corpus to improve the detection of relevant text documents [33]. In LSI, singular-value decomposition technique was incorporated that decomposes the large term by document matrix. However, in [34], the authors presented that the LSI model implemented with support vector machines shows performance degradation. Paper [35] reviewed the state of the art techniques for traffic event detection using geotagged social media data. Machine learning methods like support vector machines, logistic regression, and Naive Bayes were applied to study Twitter data for event detection around London [36]. A combination of Support vector machine classifier and Labeled Latent Dirichlet Allocation (L-LDA) was also used to analyze 700,010 tweets and extracted transportation-related information for New York City [37]. In [38], the authors proposed a unified statistical framework that combines the topic modeling based language model and hinge loss Markov random fields (HLMRFs) based model for the road traffic congestion monitoring using Twitter data. In [38], traffic congestion related Twitter data has been collected for the Washington D.C. and Philadelphia cities. Tweets have been obtained using terms like crowded, pedestrian, traffic, accident, driver, stuck, block, and crash.

3 Preliminaries

This section contains the techniques and concepts used directly or intermediary in the study. The section discusses the BOW, TF-IDF feature representation techniques in brief.

3.1 Bag-of-Words Model

Bag-of-Words (BoW) is a technique for feature extraction from a text document, and extracted features are used to train machine learning models. It is the most natural approach to represent written text document as an unordered collection of phrases. BoW keeps the frequency of terms present in a text document, and this term frequency is used to represent text in numeric form. This feature extraction approach is the best choice in building baseline models for text classification, Natural language processing and information retrieval from text documents. BoW models can be constructed in two ways: 1. by counting the word occurrence in which word occurrence represents the word's importance in the text. 2. Normalize frequency count, which is taken to reduce the dominance of extremely high occurrence keywords. We have used CountVectorizer from Sklearn library to count the occurrence of words and further used the occurrence count to convert keywords in the form of a numeric representation for machine learning classifier implementation.

Large feature dimensions and large-sized sparse representation of text are the typical limitations of the Bow approach observed in the literature. BoW approach signifies the existence of words and suits well in domain-specific small datasets because it does not include any feature selection step to reduce the dimensionality of feature size.

3.2 Term Frequency-Inverse Document Frequency Model

In text mining, Term Frequency-Inverse Document Frequency known as TF-IDF [39] is a widely used technique to convert written text into a meaningful representation of numeric values. TF-IDF is a popular feature extraction technique across many Natural language processing applications, which is based on the assumption that infrequent/rare words' contribution is more than high occurrence words in the document. In contrast to the BOW term frequency-based approach, which only holds the knowledge about the presence and absence of words in the document, TF-IDF is based on a weighing scheme, which allocate more weight to uncommon but important phrases of the written text document to highlight rare words. Further, BoW is the simplest model to convert text into numeric form. TF-IDF is another popular approach to extract features from text and to transform the written text representation of raw document into a numeric sparse matrix of TF-IDF features or a vector space model (VSM).These numeric sparse matrix elements are the significance value of a word (TF-IDF) in the collected text. The term frequency ($TF(w, T_n)$) and inverse term frequency ($IDF(w, T_n)$) and final significance value ($TF - IDF(w, T_n)$) of a word w are calculated in a Tweeter dataset T_n as per equations (1), (2) and (3). The score for a tweet-corpus is the sum over word w in both tweet t and corpus c as given in eq (4).

$$TF(w, T_n) = count(w)/T_n.total\_word\_count \tag{1}$$

$$IDF(w, T_n) = log(T_n/w') \tag{2}$$

$$TF - IDF(w, T_n) = TF(w, T_n) * IDF(w, T_n) \tag{3}$$

$$Score(t, c) = \sum_{w \in t \cup c} TF.IDF_{w,c} \tag{4}$$

Where T_n is the total number of tweets and w' is the number of tweets containing w.

In many NLP-based applications, TF-IDF shows better results than BOW that is only based on word frequency count (TF). This is due to the IDF capability that minimizes the weight of frequently occurring words and increases the significance of uncommon words. However, in text mining applications like Tweet classification for traffic incidence identification, the IDF work of weight reduction does not have the significance. In text mining, a task like traffic event detection, supervised machine learning classification technique should itself decide the relevance of phrases and, more importantly, a combination of phrases. The CFGA provides *set phrases* with a support value that have been used in sparse matrix phrase set generation. The *set phrases* clear the context of the tweet more transparently. Better features provide a key effect in traffic tweet characterization tasks using machine learning algorithms.

Bow and TF-IDF are the extensively used models to transform written text to the sparse matrix of numeric features, next to machine learning classifiers. However, TF-IDF's contribution to classification using machine learning models is not that vital because machine learning can learn the significant words of a corpus. Thus no additional down weighting/up weighting is required in machine learning classifiers. Further, our proposed approach does not use the concept of weight assignment; it merely calculates the value of support for each word and combinations of words. The proposed method CFGA chooses all the terms and combinations of terms for threshold support. Thus, it reduces the feature size and performs feature selection for the classification.

3.3 Apriori Algorithm

Apriori algorithm [40] originally developed and applied for market basket data for mining frequent itemsets and related association rules. The algorithm uses the iterative approach to find the n itemset. The exploration of n itemsets needs knowledge of n-1 itemsets. In general terminology, 1 itemset exploration is known as the L_1 phase of the iteration process, 2 and 3 itemset exploration are called as phase L_2 and L_3 respectively, i.e., L_n is the phase which explores the n-item sets from the dataset, where $n \in \mathbb{N}$. This dependency of phases can be best described by following property of Apriori algorithm.

1. If $P(x)<min\_supp$, and an item y is put together with itemset x than resultant itemset $(x \cup y)$ cannot appear more frequent than x i.e. $P(x \cup y)<min\_sup$ Where $P(x)$ and $P(x \cup y)$ represents the support of x and $(x \cup y)$ respectively.
2. If $(L_1(1) = L_2(1))$, $(L_1(2) = L_2(2))$, ...$(L_1(n-3) = L_2(n-3))$, $(L_1(n-2) = L_2(n-2)) \Rightarrow (L_1(n-1)<L_2(n-1))$. It means if (n-2) items of x and y itemsets are the same in phase L_{n-1}, then x and y can be united in phase L_n.
3. If (n-1) items of n-itemsets do not belong to phase L_{n-1}, then itemset cannot be frequent and therefore, can be pruned.

The support *supp* and *min_supp* of keywords in collected datasets can be defined as:

- Support of a keyword w in collected tweets is the ratio of tweets with the keyword. If T_w tweets contain keyword w from the dataset of T_n number of tweets, then support of keyword w can be defined as in equation 5.

$$Supp(w) = \frac{T_w}{T_n} \tag{5}$$

- In order to shrink the size of the extracted features/keywords, while keeping the important keywords preserved, we need to set a value of minimum support. *min_supp* is the parameter that needs to be set to exclude irrelevant keywords.

Table 2: Traffic related Keywords used to collect Tweets from literature

Author & Year	Keywords for collecting Twitter data
Mai & Harnac, 2013 [4]	"crash", "highway", "accident", "traffic", "road", "freeway"
D'Andrea et al., 2015 [27]	"crash", "queue", "traffic"
Nguyen et al., 2016 [24]	"crash", "delay","traffic", "accident"
Yazici et al., 2017 [41]	"crash", "highway", "accident", "traffic", "road", "delay", "freeway", "lane", "car", "cars", "incident", "collision", "NB", "northbound", "southbound", "westbound", "eastbound", "EB", "SB", "WB", "blocked", "street", "st", "road", etc.

4 Raw Data and Study Area

With the increased character limit of tweets from 140 to 280 characters per tweet, the potential of Twitter data for Information Retrieval has also grown. Each tweet may contain double keywords that represent an increased scope for the traffic event detection using Twitter data. Twitter allows keywords based tweet collection, and also, it provides the option of a bounding box for location-based tweet collection using the streaming API. To tune the hyper-parameters, we have collected three Twitter datasets named as $[\vec{\mathcal{D}}]_1$, $[\vec{\mathcal{D}}]_2$ and $[\vec{\mathcal{D}}]_3$. Three different types of Keyword databases have been used with the Twitter streaming API to collect these three datasets. Twitter data $[\vec{\mathcal{D}}]_3$ is obtained using transport-related keywords like congestion, trafficjam, potholes, road, street, etc. To get collect Twitter data $[\vec{\mathcal{D}}]_2$, weather-related keywords like cloud, cloudy, rain, storm, lightning, rainy, etc. are combined with transport-related keywords. On October 19, 2018, two passenger trains crashed into a crowd of people on the eastern outskirts of Amritsar, Punjab. Twitter data-set $[\vec{\mathcal{D}}]_1$ contains the tweets that are collected from October 19, 2018 to October 23, 2018 for this sad event. 1350224, 100000, and 100000 tweets of collected datasets $[\vec{\mathcal{D}}]_1$, $[\vec{\mathcal{D}}]_2$ and $[\vec{\mathcal{D}}]_3$ have been used to tune the threshold value *min_supp*.

4.1 Dictionary Formation

The data collection strategy is one of the interesting factors of this study. A dictionary of traffic-based keywords is created and used to collect social media data. The proposed approach to create an augmented dictionary of traffic-related keywords is as follows:

- In the beginning, traffic-related popular keywords are collected from news reports, research papers, and research articles and validated by the transport expert.

Table 3: Sample of using the WordNet database and Wu and Palmer method for Dictionary formation

Keyword	Synonym from WordNet database	Similarity Score
`"expressway"`	“freeway”, “motorway”, “pike”, “state_highway”, “superhighway”, “thruway”, “throughway”	1.0
`"car"`	“auto”	1.0
`"road"`	“highway”	.933
`"bus"`	“motorcoach”	1.0

- Synonyms of collected keywords have been fetched using a popular word-net database using python library.
- The similarity score of word and its synonym word is calculated using Wu and Palmer method [42].
- If similarity score $\geq .90$, then the synonym word is in the dictionary. Otherwise the synonym word is not taken in the dictionary.

Dictionary assembled for tweet collection consisted of the phrases collected using the CFGA method. Table 4 presents some of the phrases collected using CFGA. The proposed data collection approach represents that the collected Twitter data-set contains both types of tweets: traffic-related and non-traffic-related, but most of the non-traffic tweets have different reference/domain, containing some of the similar keywords of other class. Therefore, such tweets are hard to distinguish using different traditional approaches.

4.2 Tweet Labeling

In order to train an ML-based model, we need to label each tweet with a class name. In this study, we have collected geo-tagged tweets based on traffic-related keywords. But people use the same phrase in many different contexts; therefore, collected data also contains Tweets that do not provide traffic-related information. For example, some keyword like *accident* is a popular keyword for a road accident, but people may often use such keyword in other references. Because of this use of the same keyword in several different references, tweets containing those words but not related to our reference get collected. Table 1 shows some of the sample tweets collected. In order to train and test ML-based models to classify such tweets, 3000 tweets are manually labeled into 2-class: 1. “t” and 2. “n”, Where “t” represents traffic related tweet and, “n” denotes the non-traffic related tweet.

Table 4: Sample of phrases collected using proposed CFGA approach

Keyword1	Keyword2	Keyword3
"Heavy"	“delay”	–
"public"	“transport”	“system”
"traffic"	“flow”	
"congestion"	“peak”	“time”
"road"	“traffic”	“accident”
"road"	“traffic”	“collision”
"congestion"	“peak”	“time”
"Heavy"	“traffic”	“jam”
"bus"		
"bus"	“accident”	–
"accident"	“road”	..
"road"	“accident”	“killed”
"family"	“car”	“accident”
"family"	“killed”	“injured”

5 Data Preprocessing and Feature Generation

5.1 Tweets Preprocessing

The main objective of the preprocessing is to make the tweets eligible for the feature selection and classification task. Special characters like punctuation and stop words (“a”, “an”, “the”, etc.) are frequently used in tweets, and these stop words do not contain any meaningful information. Therefore, these special characters do not contribute any knowledge in traffic event identification task, and so the removal of such special characters and hyper-links are the primary steps of preprocessing. We have done the following preprocessing steps:

- Upper case text could not be interpreted the same so the whole text is changed to lowercase.
- Words with # symbol could be interpreted differently so phrases like “#roadaccident” are changed to “roadaccident” by removing hashes.
- In Twitter, tweet starts with “ @” symbol with the username. The symbol “ @” is removed from the username.
- **Stopwords** are very frequently occurring common words in a language that do not poses any meaning (like a, an, the in the English language). These words only take memory and valuable program execution time. Stopwords do not help in identifying or indexing of a text, so such words can be removed by preparing a list of words which consider as stopwords. However, in this research, these words are removed [39] with the use of the Natural language Toolkit library.

- **Stemming** and removal of stopwords are the necessary dimensionality reduction steps of the data preprocessing process, which improves the performance of text classifiers. Stemming is performed to group similar words into their root form or stem. In this paper, we have applied the Porter Stemmer [43] to extract inflexional endings from English words.
- **Tokenization** is the process of splitting up a sentence into pieces such as words, keywords, or other elements like phrases, groups of words, which are known as set-phrases. Phrases/set-phrases are the keywords, words, or combinations of keywords (as n-grams) of the sentences. The process of tokenization produces a bag of tokens for further processing.

5.2 Feature Generation

Feature generation is the process of defining features from the unstructured traffic tweets. Dimensionality reduction and performance improvement are the two main objectives of the feature generation process [44], [45]. Feature generation is a prior step of the tweet classification task and independent to later. Therefore, features generated once can be used again and again in different machine learning classification techniques. The process includes the following:

Feature Extraction

To capture patterns about the training dataset, we need to extract out features representing the training instances.While the noisy tweet data may contain words, symbols, marks, or special characters with no utility for the learner classifier, usable and informative tokens need to be taken out from the tweet text. The steps are taken to extract out usable features from the text, and leaving the inappropriate behind is known as feature extraction [46]. Thus, extracted features represent the tweet text, and hence the performance of predictive algorithms depends a lot on the ability of the feature extraction approach.

Feature Selection

The process of feature selection is the collection of steps taken to shrink the extracted features' size while keeping the most important preserved. The accuracy of the traffic event identification task from Twitter data depends on the relevant, frequently occurring keyword selection strategy. In this paper, a novel feature generation approach is proposed to select the combinations of frequently occurring keywords. It is necessary to extract features from the text and convert the extracted text features into numeric or vector forms. BOW and TF-IDF are popular techniques in various text mining applications for such a task.

6 Problem Description and Methodology

This section covers the necessary notations and acronyms used in the paper, problem description, hyper-parameters in the proposed approach, and detailed method description.

6.1 Basic Notations

In this section, notations, and acronyms used in the paper are denoted. The vector of *set phrases* is represented as $\vec{v}$, where v and $\rightarrow$ depicts the name of the vector and its behavior, respectively. The collected dataset of tweets is represented as $[\vec{\mathcal{T}}]$ while preprocessed and labeled dataset is $[\vec{\mathcal{D}}]$, which contains N number of tweets each in the form of $\{tweet_i, label_i\}$, where $i \in \{1, N\}$. Hyper-parameters used in the study are support and minimum support values, which are denoted as *supp* as ω and *min_supp* as $\mathcal{S}$ and $\{\omega, \mathcal{S}\} \in \mathscr{R}$, respectively. Matrix of *set phrases* as $[\vec{\mathcal{M}}]$, $\mathcal{K}$: Dictionary of identified Traffic keywords for tweet collection, $\mathcal{BB}$: location information for bounding box, $\vec{\mathcal{C}}$: Corpus of tweets, CR: Classification Report, Acc: Accuracy, Prec: Precision, Rcl: Recall, Afim: Apriori based frequent item-set Method, KNN: K-nearest neighbor, SVM: Support vector machines.

6.2 Problem Description

The objective of this study is to detect tweets containing traffic-related information. The problem can be defined as: for a corpus of tweets and labels, where label "t" represents the class of tweets containing traffic event and label "n" represents the class of non-traffic tweets, we wish to propose a methodology for faster and accurate prediction of labels on the test dataset. Frequent features with respective ω value is used for feature selection and numeric sparse matrix generation, which is used to train the classification model with corresponding labels and further test the trained model using the test dataset. The frequent feature extraction problem at a given $\mathcal{S}$ can be formally defined as:

Let Tweets = $\{t_1, t_2, ...\}$ be a set of distinct tokens/features, usually called items. An item-set or feature set F is a subset of Tweets. If $\mid F \mid$ = n, then F is called n-feature set. A tweet is a short message $< tID, T >$, where tID is a unique tweet identifier, and T is the text content of the tweet. A feature F is the part of tweet $< tID, X >$ if $F \subseteq X$. Given a preprocessed Twitter dataset $[\vec{\mathcal{D}}]$, the subset of tweets that contain a feature set F is represented as $[\vec{\mathcal{D}}]_F$. At a given *min_supp* $\mathcal{S}$, a feature is called frequent in $[\vec{\mathcal{D}}]$ if $\omega_{[\vec{\mathcal{D}}]}(\mathrm{F}) \geq \mathcal{S}$ which is further represented as constraint C_{ss}. The frequent feature extraction problem can be defined as the mining of all frequent feature $FR((C_{ss})_{T_{freq}})$.

6.3 Proposed Methodology

The workflow to extract the traffic-related information is explained further in this section. Steps are depicted graphically in Figure 1 with all the steps taken, starting from tweets crawling to a trained ML-based classifier. The proposed

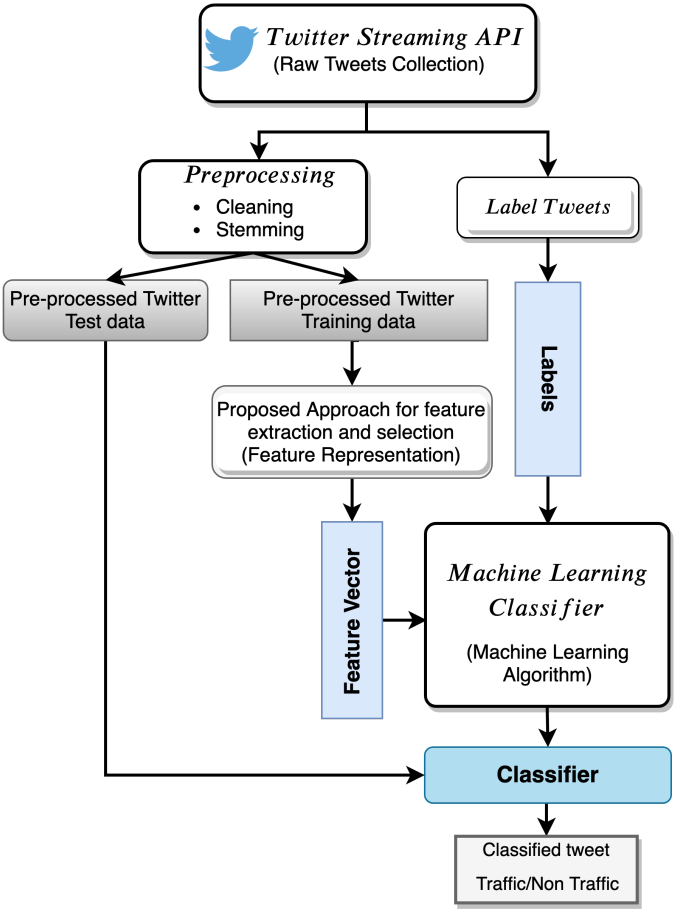

Twitter Streaming API
(Raw Tweets Collection)
Preprocessing
• Cleaning
• Stemming
Label Tweets
Pre-processed Twitter Test data
Pre-processed Twitter Training data
Labels
Proposed Approach for feature extraction and selection (Feature Representation)
Feature Vector
Machine Learning Classifier
(Machine Learning Algorithm)
Classifier
Classified tweet
Traffic/Non Traffic

Fig. 1: Detailed Steps of Proposed Methodology

method to extract traffic information from Twitter based social media data is as follows.

The Algorithm 1 presents the proposed CFGA methodology. The initial steps of Algorithm 1, represents the dataset collection and further preprocessing of the collected dataset. The raw data collection, labeling and preprocessing have been explained in previous sections. Further, the points [9-11] in Algorithm 1 lists the frequent *set phrases* that have been obtained using the notion of ω. The parameter $\mathcal{S}$ is used for *set phrases* selection and pruning of unnecessary items. Since the nature of different events may also be different, therefore, a hardcoded value for support threshold will not be beneficial. A minimum threshold value for support has been decided to prune the unnecessary *set phrases* from the dataset. Frequent *set phrases* extraction and further sparse matrix conversion of selected frequent *setphrases* have been shown in points 5 to 8 and 9 to 11 of Algorithm 1. Point 11 depicts the keywords arrangements in

Algorithm 1 CFGA($[\vec{\mathcal{D}}], \omega, \mathcal{S}, \rho$)

```
$[\vec{\mathcal{C}}] = TwitterStreaming([\mathcal{K}], [\mathcal{BB}])$
$DataCollection$:
$[\vec{\mathcal{D}}] = PreProcessing([\vec{\mathcal{T}}])$
$Preprocessing$:
for $i = 0$ to $len(\vec{\mathcal{D}})$ do
    $l_i = l_i.appent(word\_tokenize(\vec{\mathcal{D}}_i))$
end for
$Tokenization$:
$\vec{\mathfrak{d}}_t = Afim([O\vec{hary}], \mathcal{S})$
$FrequentFeatureSet(frozenset)$:
$[\vec{\mathcal{D}}], \vec{\omega}, \vec{\mathcal{L}} =$ArrangeData$([\vec{\mathcal{D}}], \vec{\mathfrak{d}}_t)$
$dim = length([\vec{\mathcal{D}}][0])$
$[w\vec{\_m}] =$FDF$([T\vec{rain}], dim, \omega, \mathcal{S})$
$[tr\vec{ai}n], [te\vec{s}t] = SplitRandomly([\vec{\mathcal{D}}], SplitIndex)$
$Training$ :
$[X_{T\vec{rain}}] =$Transformm$([w\vec{\_m}], \vec{n}, [X_{T\vec{rain}})]$
$[X_{\vec{test}}] =$Transformm$([w\vec{\_m}], \vec{n}, [X_{\vec{Test})}]$
$t\ n, Acc, CR = CM([X_{T\vec{rain}}], [Y_{T\vec{rain}}], X_{T\vec{est}}])$
return (t n, Acc, CR)
```

frozenlist. the study proposes a novel approach for *set phrases* generation and further numeric sparse matrix creation for text data representation, which is applied on top of the classification models to detect the tweets containing traffic events accurately. The proposed *set phrases* generation approach has been denoted as combinatorial feature generation approach (CFGA) in the paper.

6.3.1 Tunable Parameters

The proposed approach contains a hyper-parameter called as $\mathcal{S}$, which represents the minimum value of ω used for frequent item-set selection while generating the numeric sparse matrix and ω of a feature is the fraction of tweets containing particular features. The variation of $\mathcal{S}$ with the number of frequent item-sets on three different datasets $[\vec{\mathcal{D}}]_1$, $[\vec{\mathcal{D}}]_2$, $[\vec{\mathcal{D}}]_3$ has been shown in Figure 2. The graphs between $\mathcal{S}$ and the number of frequent item-sets for the dataset $[\vec{\mathcal{D}}]_2$, $[\vec{\mathcal{D}}]_3$ represents almost similar patterns for $\mathcal{S}$ interval [1% - 2%]. Dataset $[\vec{\mathcal{D}}]_1$ has been collected at the time of the Amritsar accident disaster. Therefore, dataset $[\vec{\mathcal{D}}]_1$ more tends to include the disasters more prominently and this may be one of the reasons behind showing different behavior in Figure 2a compared to Figure 2b and 2c. The traffic event detection approach has also been employed K- Nearest Neighbor (KNN), which contains hyper-parameter K, which needs to be tuned in model training to obtain K's optimal value for better accuracy. K-value represents the number of participating neighbors in the KNN algorithm. In this study, K has been tuned based on misclassification error. Figure 3 shows the tuning of parameter K for its optimal value using a brute-force approach while using a BOW for feature extraction. When TF-IDF is applied for features in KNN model training, Figure 4 represents how to tune the optimal value of K using the same brute-force approach. Misclassification error minimizes at K=3 and K=5 in Figure 3 and Figure 4, respectively.

6.3.2 KNN Classifier

K-nearest neighbors is a non-parametric machine learning classifier, which scores its nearest neighbors in training data, and k-top-scored neighbors' class is used to classify the new input data [47], [48]. The decision rule can be written as eq 6.

$$Score(t,c) = \sum_{t_n \epsilon KNN(t)} similarity(t,t_n) D(t_n,c) \tag{6}$$

$D(t_n,c)$ is the categorization for tweet t_n with respect to category c that is defined as eq 7.

$$D(t_n,c) = \begin{cases} 1, & \text{if } t_n \in c \\ 0, & \text{if } t_n \notin c \end{cases} \tag{7}$$

6.3.3 Naive Bayes Classifier

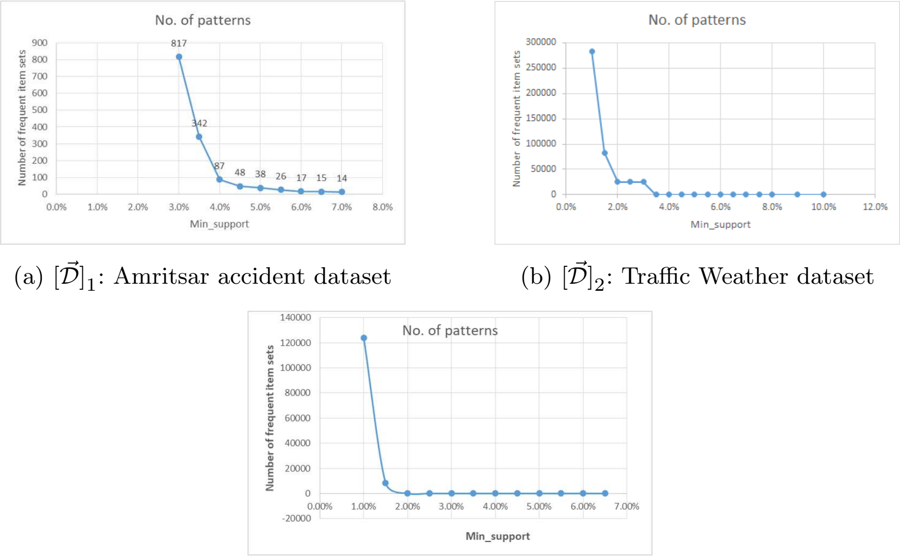
Naive Bayes classifier is based on the assumption that the probability of one feature being in a class is independent of any other feature in the class. The naive Bayes model is based on Bayes' theorem and performs well on big datasets. It is used in text classification that uses a method of estimating the possibility of different classes based on different features of the written text document.

6.3.4 Support Vector Machines

Support vector machines (SVM) are the supervised classification techniques that categorize the data. The basic SVM model training is performed as follows: 1. Plot features in a dimensional space where each plotted point coordinates are known as support vectors. 2. A separating line needs to be found in such a way so that each point should be farthest from the separating line. 3. In the testing phase, depending on the test, where the data is land on either side of the line, we can categorize new data.

7 Results and Analysis

This section contains the performance evaluation of machine learning classifiers with a combination of CFGA, BOW, and TF-IDF models.Further, we have also examined the false positive and false negative tweets. False-positive and false-negative tweets are tweets that are falsely identified.



(a) $[\vec{\mathcal{D}}]_1$: Amritsar accident dataset

(b) $[\vec{\mathcal{D}}]_2$: Traffic Weather dataset

(c) $[\vec{\mathcal{D}}]_3$: Traffic dataset1

Fig. 2: Variation of $\mathcal{S}$ with frequent *set phrases*

7.1 Performance Metrics

In this study, prevalent classification metrics like accuracy, precision and recall, F-measure, have been used to measure the performance of the event detection model.

- Accuracy - the fraction of rightly classified Twitter messages is known as the accuracy of trained model classifier.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \tag{8}$$

Where TN is true negative, TP is true positive, FN is false positive and FP is false positive whereas t is the positive class and negative class is n.

- Precision and Recall - Precision is the percentage/fraction of relevant tweets while recall is the percentage of actual positives rightly identified.

$$Precision = \frac{TP}{TP + FP} \tag{9}$$

$$Recall = \frac{TP}{TP + FN} \tag{10}$$

- F-measure - harmonic average of precision and recall. F-measure is the function of precision P and recall R and can be defined as in eq11.

$$F - measure = \left(\frac{P^{-1} + R^{-1}}{2} \right)^{-1} \tag{11}$$

7.2 Feature Vector

Here we evaluated the use of frequent *set phrases* derived after pruning at the hyper-parameter $\mathcal{S}$ using the proposed framework as a feature vector to represent the tweets group of phrases. After the preprocessing, the items and combination of items have been derived at $\mathcal{S}$ window of 0.010 to 0.08 at an interval of 0.05. In order to accurately detect the class, the optimal value of $\mathcal{S}$ =0.015 has been chosen based on frequent *set phrases* and accuracy achieved. The selected frequent *set phrases* have been converted to numeric sparse matrix representation as a method of prerequisites of the learning algorithm. Higher support means more individual token features and less high order paired token features. Table 5 presents that CFGA generates smaller size matrix than TF-IDF. Smaller matrix size reduces the computation speed of data intensive task.

7.3 Classifier Performance

Table 6 represents the performance of three ML-models trained on top of proposed CFGA, BOW, and TF-IDF feature extraction models to characterize traffic tweets. In the proposed approach, each tweet is converted to a row of the ω matrix $[\vec{\mathcal{M}}]$, and frequent phrases-*set phrases* have been written in a column vector. The ω matrix can be visualized as the collection of sparse vectors in which the count of non-zero values in the vector represents the count of words and its *set phrases* in the tweet. Thus, the matrix $[\vec{\mathcal{M}}]$ generated using the proposed approach has the shape of $R * C$, where $R \in$ number of tweets and C is the count of selected *set phrases.* The weight of *set phrases* is the support of the *set phrases* determined using apriori algorithm. During training, the proposed method discovers the frequent *set phrases* in the corpus. In order to implement KNN classifier, the optimal value of k=3 in the case of

Table 5: Generated sparse matrix size with TF-IDF and proposed CFGA method. CFGA generates smaller feature vector that reduces its large primary memory requirement in computation and decreases computation time in data-intensive application.

Feature generation method	ω	Input Training items	generated sparse matrix size
TF-IDF	-	140	[140, 3888]
CFGA	.015	140	[140, 1121]

Table 6: Performance of Machine Learning models when built on top of different feature generation approaches in terms of Accuracy, and Precision, Recall, F-measure on traffic (t) and non-traffic (n) classes of Twitter data.

ML Model	Feature generation method	Accuracy	Precision		Recall		F-Measure	
Traffic (t) & Non-traffic (n) Class		→	t	n	t	n	t	n
KNN	BOW	66.2	.70	.64	.68	.66	.67	.66
	TF-IDF	68.6	.68	.70	.71	.67	.69	.68
	CFGA (Proposed approach)	**77.7**	.77	.78	.72	.82	.74	**.80**
NB	BOW	63.8	.66	.62	.61	.66	.62	.65
	TF-IDF	65.4	.66	.65	.63	.68	.64	.66
	CFGA (Proposed approach)	**66.2**	.85	.63	.29	.96	.44	**.76**
SVM	BOW	64.7	.69	.62	.66	.63	.58	.67
	TF-IDF	65.9	.68	.64	.59	.73	.63	.68
	CFGA (Proposed approach)	**73.6**	.83	.70	.52	.91	.64	**.79**

bow and k=5 in the case of TF-IDF has been used (see Figure 3 and 4). The CFGA implementation uses the optimal value of K=1 for KNN classification (see Figure 5). The accuracy results in Table 6 depict that all three machine learning models' performance varies between 63%-78% in terms of Acc. KNN and SVM, on top of the proposed CFGA outperform other ML-based classifier Naive Bayes by almost 10%. However, SVM classifier with CFGA presents higher precision for t class of tweets, and it successfully brings about the Acc and F-measure of 73.6% and [64% (for class t), 79% (for class n)], respectively, for the traffic-related Twitter dataset labeled for binary classification.

Since the KNN model with the CFGA has achieved the best performance, this classifier is treated as the representative of machine learning models for the traffic event classification task. Comparing the Prec, Rcl, f1-score, and Acc achieved through different feature extraction models reveals the effectiveness of feature extraction techniques in modeling the tweet tokens. The study results presented in Figure 6 depicts that the proposed CFGA technique significantly increases the performance of all three machine learning classifiers (KNN, NB, and, SVM).

Table 6 represents the Prec, Rcl, and F-Measure for the test dataset. Further, the effectiveness of the ML-based classification model is directly proportional to Prec, Rcl, and F-Measure for both classes. Some studies reported the lower Prec and Rcl for traffic class tweets. The reason behind this lower value may be the smaller number of traffic class tweets. However, in our case, we have dedicated data collection strategy as described in section 4; due to that, we have collected a high number of traffic class tweets compared to non-traffic class tweets number. Therefore, as indicated in the Prec results of CFGA

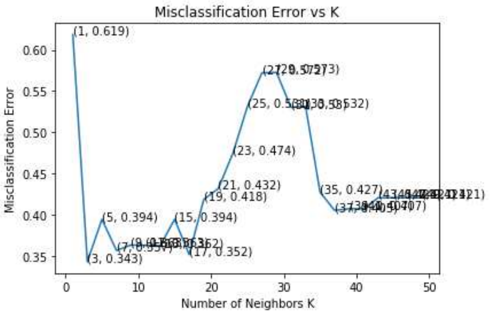


Fig. 3: *BoW*: Tuning hyper-parameter K for KNN classifier

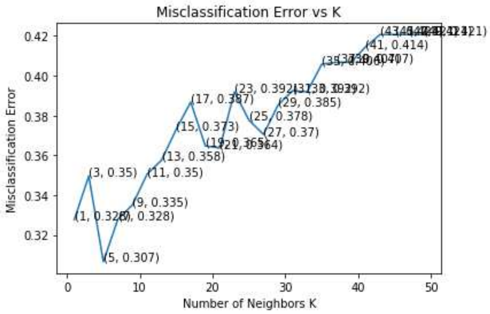


Fig. 4: *$TF - IDF$*: Tuning hyper-parameter K for KNN classifier

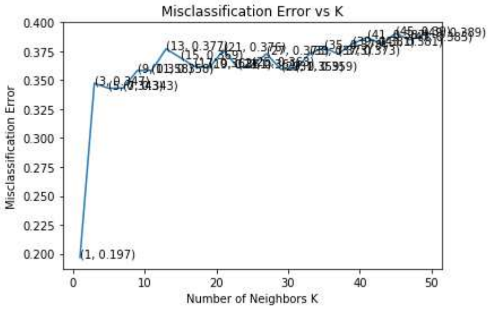


Fig. 5: *CFGA* Tuning hyper-parameter K for KNN classifier

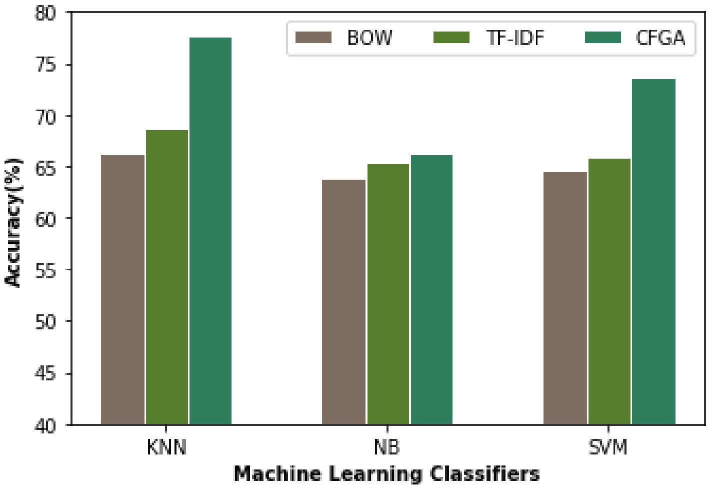


Fig. 6: Accuracy improvement of machine learning classifiers using CFGA.

Table 7: Sample of Tweets extracted using proposed approach

Sl no.	Tweet	Location	Keywords	Date-Time
1.	A vehicle overturned at Maa flyover, Park circus bound.	Maa flyover, Park circus	vehicle	May 8, 2019-7:17 PM
2.	there is a huge traffic jam on Maa Flyover.. any update on it ? Not able to understand the reason for this..	Maa Flyover	traffic jam	May 8, 2019-7:42 PM
3.	there is always a severe traffic snarl on Rajdanga Road due to DPS Junior School. It's a huge risk to children, all of whom are below 6 years. During pick up and drop hours can the traffic be restricted to West to East one way? #suggestion	Rajdanga Road	traffic	May 13, 2019-9:26 AM
4.	Simple solution to avert the traffic jam in the Dhalai bridge area near Shahid Khudiram Metro station, Kolkata.Reduce the green light duration in the Garia Station road which has less traffic and increase it for the road to Kamalgazi/Ruby which has more traffic	Dhalai bridge, Shahid Khudiram Metro station	traffic jam, road,	May 13, 2019-11:13 AM
5.	A wire is lying on Maa flyover which seems to be dangerous for riders near sonar bangla hotel if u r going to park circus via the flyover.	Maa flyover. sonar bangla hotel	riders	May 25, 2019-5:45 PM
6.	Sir,road between Esplande to Sealdah via Lenin sarani remains highly congested due to illegal parking, slow bus movement, illegal road blockage.Hence it is requested to deploy police to avoid congestion & suffering of daily commuters.pls look into it soon	Esplande to Sealdah via Lenin sarani	congestion, road, commuters	Jun 20, 2019-6:51PM
7.	Choas as no vehicle is allowed to take maa flyover towards park circus. A vehicle break down 20 minutes back and created this choas	maa flyover	vehicle	Jul 3, 2019-9:43 AM
8.	Convent Road was having waist height of water at 11 am. I got stuck in the middle of the road with my self driven car	Convent Road	road, car	Aug 17, 2019-3:14 PM

(proposed approach) in Table 6, the value of Prec readings for non-traffic class tweets has been obtained lower.

7.4 Validation

The study presents a ML-based methodology to identify tweets containing traffic events from the social media messages. The classification model need to train on labeled tweets to detect traffic events from Twitter data, and then tweets can be analyzed for identifying traffic events. In this study, 70% of labeled tweets have been selected to train the model while the remaining labeled tweets have worked as the test dataset. The classifier's performance is examined on a test dataset with varying proportions of traffic and non-traffic tweets. As part of the validation, we have also tested our proposed approach using social media data of an Indian city, Kolkata. We started Kolkata traffic-related tweet collection from 16 April 2019, 08:18:30, using a keyword-based approach. The collected dataset has been used to validate the results of the study. Park Circus Seven Point Crossing, Behla Chowrasta, Shyambazar Five Point Crossing, MG Road, Gariahat, Jadavpur Police Station Crossing, Altadanga, Science City Connector, Esplanade, BBD Bag are important cross-roads which are considered to be the predominantly congested traffic points of Kolkata city [49]. Almost all locations are in the text of Tweet that has extracted using the proposed approach. Table 7 presents the sample of tweets obtained through the proposed approach. The tweet in Table 7 includes traffic-related events in tweets 1, 4, 5, 6, 7, and 8, while tweets 2 and 3 do not contain any specific events that have traffic-related terms. The date and time of tweet creation also play an essential role in such work. For example, Tweet 2 is about asking the reason behind the crowd. At the same time, tweet 1 has the response to tweet 1. The text for Tweet1 and 2 includes the location keyword Maa flyover, and the date-time of Tweet creation is the same.

8 Conclusion

The study's final results show that the micro-blogging sites are the potential source of information to identify traffic events. The basic idea of the proposed CFGA is derived from the Apriori algorithm, which is mainly used for market-basket analysis. ML-based classifiers show better performance in Twitter data analysis when built on top of the CFGA. As discussed in the results section, the proposed approach presents superior results over other comparable methods and identifies the tweets containing traffic-related events more accurately. The feature pruning with the help of hyper-parameter $\mathcal{S}$ incorporates the dimensionality reduction while converting tweet sentences to machine-readable form for the learning algorithm. This phase of dimensionality reduction improved the performance of the overall traffic data characterization system not only in terms of accuracy but also in memory and computational power requirements.

Social media data enables us to resolve many data mining task-related problems, but still, knowledge mining from social media data brings some limitations:

- Very few social media users open their location information while posting messages; therefore, the percentage of geo-tagged messages is very less. Some studies claim that less than 1% of Twitter users share their location while posting tweets. This is the major challenge in the analysis of location-based traffic-related tweets.
- Being a popular micro-blogging platform, Twitter is the most significant source of the big-data generation, bringing new challenges in storing and preprocessing this data. The proposed strategy shows a substantial reduction in the size of the numeric feature sparse matrix. For example, while classifying preprocessed social media data, TF-IDF feature representation creates the [140, 3888] size sparse matrix with the input of 140 training items. In contrast, our proposed CFGA creates only [140, 1121] size sparse matrix with ω= 0.015 and same input training data without loss of generality. Further, This reduction of sparse matrix size reduces the overall processing power of the computer for the classification task.

To get the location of the traffic event, the geo-code of the tweet has to be plotted correctly. This study with an efficient geocode-plotter can optimize traveler information system. The travelers can decide the better route and get the traffic event information quickly. This obtained traffic information will be from the tweets of actual traffic users. Further, the transport authority can easily detect the traffic disturbances, which can help to resolve the problem.

Acknowledgement

This research is supported by All India Council for Technical Education, under Department of Higher Education, Ministry of Human Resource Development, Govt. of India.

Acknowledgment:

We sincerely acknowledge the Indian Institute of Technology Roorkee and All India Council for Technical Education, under Department of Higher Education, Ministry of Human Resource Development, Govt. of India for supporting this study.

Declaration:

Ethics Approval and Consent to Participate: No participation of humans takes place in this implementation process

Human and Animal Rights: No violation of Human and Animal Rights is involved.

Funding: No funding is involved in this work.

Conflict of Interest: Conflict of Interest is not applicable in this work.

Authorship contributions: There is no authorship contribution

Acknowledgment: We sincerely acknowledge the Indian Institute of Technology Roorkee and All India Council for Technical Education, under Department of Higher Education, Ministry of Human Resource Development, Govt. of India for supporting this study.

Conflict of Interest

Manuscript title: Combinatorial Approach of Feature Generation for Traffic Event Detection using Social Media Data: CFGA

Corresponding Author Name: Narayan Chaturvedi, CTRANS, Indian Institute of Technology Roorkee, 247667, Uttarakhand, India.

We have no conflict of interest to declare. This statement is to certify that all authors have seen and approved that the manuscript is submitted. On behalf of all Co-Authors, the corresponding author shall bear full responsibility for the submission.

All the authors sign this statement to indicate agreement that the above information is accurate and correct.

The corresponding author shall have to complete and submit this form to the Editorial Office. I certify that there is no actual or potential conflict of interest with this article.

NChaturvedi

Narayan Chaturvedi

Corresponding Author

References

[1] Ashtari, O.: The super tweets of# sb47. Twitter. com Blog (2013)

[2] Gal-Tzur, A., Grant-Muller, S.M., Minkov, E., Nocera, S.: The impact of social media usage on transport policy: issues, challenges and recommendations. Procedia-Social and Behavioral Sciences **111**, 937–946 (2014)

[3] Wu, B., Idris, A.O.: Measuring and visualizing transit customers satisfaction using twitter data. Technical report (2018)

[4] Mai, E., Hranac, R.: Twitter interactions as a data source for transportation incidents. Technical report (2013)

[5] Wang, Z., Cui, X., Gao, L., Yin, Q., Ke, L., Zhang, S.: A hybrid model of sentimental entity recognition on mobile social media. EURASIP Journal on Wireless Communications and Networking **2016**(1), 253 (2016)

[6] Sebastiani, F.: Machine learning in automated text categorization. ACM computing surveys (CSUR) **34**(1), 1–47 (2002)

[7] Bian, J., Topaloglu, U., Yu, F.: Towards large-scale twitter mining for drug-related adverse events. In: Proceedings of the 2012 International Workshop on Smart Health and Wellbeing, pp. 25–32 (2012). ACM

[8] Wu, H.C., Luk, R.W.P., Wong, K.F., Kwok, K.L.: Interpreting tf-idf term weights as making relevance decisions. ACM Transactions on Information Systems (TOIS) **26**(3), 13 (2008)

[9] Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[10] Liu, Q., Jiang, H., Wei, S., Ling, Z.-H., Hu, Y.: Learning semantic word embeddings based on ordinal knowledge constraints. In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 1501–1511 (2015)

[11] Huang, A., Gallegos, L., Lerman, K.: Travel analytics: Understanding how destination choice and business clusters are connected based on social media data. Transportation Research Part C: Emerging Technologies **77**, 245–256 (2017)

[12] Rashidi, T.H., Abbasi, A., Maghrebi, M., Hasan, S., Waller, T.S.: Exploring the capacity of social media data for modelling travel behaviour:

Opportunities and challenges. Transportation Research Part C: Emerging Technologies **75**, 197–211 (2017)

[13] Zhang, Z., He, Q., Zhu, S.: Potentials of using social media to infer the longitudinal travel behavior: A sequential model-based clustering method. Transportation Research Part C: Emerging Technologies **85**, 396–414 (2017)

[14] Cebelak, M.K.: Location-based social networking data: doubly-constrained gravity model origin-destination estimation of the urban travel demand for austin, tx (2013)

[15] Jiang, B., Yin, J., Zhao, S.: Characterizing the human mobility pattern in a large street network. Physical Review E **80**(2), 021136 (2009)

[16] Jin, P.J., Cebelak, M., Yang, F., Ran, B., Walton, C.M., Zhang, J.: Location-based social networking data: exploration of use of doubly constrained gravity model for origin-destination estimation. In: The 93rd Annual Meeting of Transportation Research Board, Washington DC, pp. 14–5314 (2014)

[17] Hasan, S., Ukkusuri, S.V.: Urban activity pattern classification using topic models from online geo-location data. Transportation Research Part C: Emerging Technologies **44**, 363–381 (2014)

[18] Alesiani, F., Gkiotsalitis, K., Baldessari, R.: A probabilistic activity model for predicting the mobility patterns of homogeneous social groups based on social network data. In: Transportation Research Board: 93rd Annual Meeting, pp. 14–1013 (2014)

[19] Ni, M., He, Q., Gao, J.: Using social media to predict traffic flow under special event conditions. In: The 93rd Annual Meeting of Transportation Research Board (2014)

[20] Hasan, S., Zhan, X., Ukkusuri, S.V.: Understanding urban human activity and mobility patterns using large-scale location-based data from online social media. In: Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing, p. 6 (2013). ACM

[21] Zhang, Z., He, Q., Gao, J., Ni, M.: A deep learning approach for detecting traffic accidents from social media data. Transportation research part C: emerging technologies **86**, 580–596 (2018)

[22] Manca, M., Boratto, L., Roman, V.M., i Gallissà, O.M., Kaltenbrunner, A.: Using social media to characterize urban mobility patterns: State-of-the-art survey and case-study. Online Social Networks and Media **1**, 56–69 (2017)

[23] Yang, Y., Liu, X., *et al.*: A re-examination of text categorization methods. In: Sigir, vol. 99, p. 99 (1999)

[24] Nguyen, H., Liu, W., Rivera, P., Chen, F.: Trafficwatch: Real-time traffic incident detection and monitoring using social media. In: Bailey, J., Khan, L., Washio, T., Dobbie, G., Huang, J.Z., Wang, R. (eds.) Advances in Knowledge Discovery and Data Mining, pp. 540–551. Springer, Cham (2016)

[25] Gutierrez, C., Figuerias, P., Oliveira, P., Costa, R., Jardim-Goncalves, R.: Twitter mining for traffic events detection. In: 2015 Science and Information Conference (SAI), pp. 371–378 (2015). IEEE

[26] Gu, Y., Qian, Z.S., Chen, F.: From twitter to detector: Real-time traffic incident detection using social media data. Transportation research part C: emerging technologies **67**, 321–342 (2016)

[27] D'Andrea, E., Ducange, P., Lazzerini, B., Marcelloni, F.: Real-time detection of traffic from twitter stream analysis. IEEE transactions on intelligent transportation systems **16**(4), 2269–2283 (2015)

[28] Atefeh, F., Khreich, W.: A survey of techniques for event detection in twitter. Computational Intelligence **31**(1), 132–164 (2015)

[29] Fu, K., Nune, R., Tao, J.X.: Social media data analysis for traffic incident detection and management. Technical report (2015)

[30] Parikh, R., Karlapalem, K.: Et: events from tweets. In: Proceedings of the 22nd International Conference on World Wide Web, pp. 613–620 (2013). ACM

[31] Alajlan, S., Coenen, F., Konev, B., Mandya, A.: Ontology learning from twitter data. In: Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (2019). SCITEPRESS-Science and Technology Publications

[32] Cramér, H.: Mathematical Methods of Statistics vol. 43. Princeton university press, ??? (1999)

[33] Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K., Harshman, R.: Indexing by latent semantic analysis. Journal of the American society for information science **41**(6), 391–407 (1990)

[34] Wu, H., Gunopulos, D.: Evaluating the utility of statistical phrases and latent semantic indexing for text classification. In: 2002 IEEE International Conference on Data Mining, 2002. Proceedings., pp. 713–716 (2002). IEEE

[35] Xu, S., Li, S., Wen, R.: Sensing and detecting traffic events using geosocial media data: A review. Computers, Environment and Urban Systems **72**, 146–160 (2018)

[36] Suma, S., Mehmood, R., Albeshri, A.: Automatic detection and validation of smart city events using hpc and apache spark platforms. In: Smart Infrastructure and Applications, pp. 55–78. Springer, ??? (2020)

[37] Khan, S.M., Chowdhury, M., Ngo, L.B., Apon, A.: Multi-class twitter data categorization and geocoding with a novel computing framework. Cities **96**, 102410 (2020)

[38] Chen, P.-T., Chen, F., Qian, Z.: Road traffic congestion monitoring in social media with hinge-loss markov random fields. In: 2014 IEEE International Conference on Data Mining, pp. 80–89 (2014). IEEE

[39] Manning, C., Raghavan, P., Schütze, H.: Introduction to information retrieval. Natural Language Engineering **16**(1), 100–103 (2010)

[40] Agrawal, R., Srikant, R., *et al.*: Fast algorithms for mining association rules. In: Proc. 20th Int. Conf. Very Large Data Bases, VLDB, vol. 1215, pp. 487–499 (1994)

[41] Yazici, M.A., Mudigonda, S., Kamga, C.: Incident detection through twitter: Organization versus personal accounts. Transportation Research Record **2643**(1), 121–128 (2017)

[42] Wu, Z., Palmer, M.: Verbs semantics and lexical selection. In: Proceedings of the 32nd Annual Meeting on Association for Computational Linguistics, pp. 133–138 (1994). Association for Computational Linguistics

[43] Porter, M.F., *et al.*: An algorithm for suffix stripping. Program **14**(3), 130–137 (1980)

[44] Raymer, M.L., Punch, W.F., Goodman, E.D., Kuhn, L.A., Jain, A.K.: Dimensionality reduction using genetic algorithms. IEEE transactions on evolutionary computation **4**(2), 164–171 (2000)

[45] Islamaj, R., Getoor, L., Wilbur, W.J.: A feature generation algorithm for sequences with application to splice-site prediction. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 553–560 (2006). Springer

[46] Trier, Ø.D., Jain, A.K., Taxt, T.: Feature extraction methods for character recognition-a survey. Pattern recognition **29**(4), 641–662 (1996)

[47] Cover, T.M., Hart, P., *et al.*: Nearest neighbor pattern classification. IEEE

transactions on information theory **13**(1), 21–27 (1967)

[48] Han, X., Liu, J., Shen, Z., Miao, C.: An optimized k-nearest neighbor algorithm for large scale hierarchical text classification. In: Joint ECML/PKDD PASCAL Workshop on Large-Scale Hierarchical Classification, pp. 2–12 (2011)

[49] Chowdhury, I.R.: Traffic congestion and environmental quality: a case study of kolkata city. Int. J. Humanit. Soc. Sci. Invent **4**(7), 20–28 (2015)