IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received April 3, 2019, accepted April 14, 2019, date of publication May 15, 2019, date of current version June 4, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2917159

Face off: Travel Habits, Road Conditions and Traffic City Characteristics Bared Using Twitter

AMIT AGARWAL AND DURGA TOSHNIWAL
Department of CSE, IIT Roorkee, Roorkee 247667, India

Corresponding author: Amit Agarwal (aagarwal3@cs.iitr.ac.in)

This work was supported in part by the Information Technology Research Academy (ITRA), MeITY, Government of India, under Grant MIT-1100 CSE.

ABSTRACT The adequacy of traditional transport related issues detection is often limited by physical sparse sensor coverage and reporting incident/issues to the emergency response system is labor intensive. The social media tweet text have been mined so as to identify the complaints regarding various road transportation issues of traffic, accident, and potholes. In order to identify and segregate tweets related to different issues, keyword-based approaches have been used previously, but these methods are solely dependent on seed keywords which are manually given and these set of keywords are not sufficient to cover all tweets posts. So, to overcome this issue, a novel approach has been proposed that captures the semantic context through dense word embedding by employing word2vec model. However, the process of tweet segregation on the basis of semantic similar keywords may suffer from the problem of pragmatic ambiguity. To handle this, Word2Vec model has been applied to match the semantically similar tweets with respect to each category. Furthermore, the hotspots have been identified corresponding to each category. However, due to the scarcity of geo-tagged tweets, we have proposed a hybrid method which amalgamates Named Entity Recognition (NER), Part of speech (POS), and Regular Expression (RE) to extract the location information from the tweet textual content. Due to the lack of availability of the ground truth dataset, model feasibility has been validated from the existing data records (*i.e., published by government official accounts and reported on news media*) and the evaluation results signify that the stated approach identifies few additional hotspots as compared to the existing reports while analyzing the tweets.

INDEX TERMS Incident detection, social media, Named Entity Recognition, Part of Speech, hotspot detection, word embedding, transportation, Word2Vec.

I. INTRODUCTION

In India, four major tier-1 cities (Mumbai, Delhi, Kolkata, and Bengaluru) annually losses 22 billion dollar due to congestion. It mainly induced from non-recurrent events such as *accident, adverse road conditions, construction on roads, potholes, adverse weather condition, and inadequate drainage*. Due to this individual has to spend more than one-a-half hour longer during the peak hour to cover the same distance as on non-peak hour. Furthermore, it's one of the most significant challenge in-front of infrastructural manager and commuters as these events would take most of the time as well as causes a number of deaths. The report published by MORTH (Ministry of Road Transport& Highways) shows that the number of fatalities in India due to potholes from the last five year is 14,296 which is much higher than the casualties due to terrorist or Naxal attacks. Whereas death due to road construction is increased by 50% in 2017 (i.e. 4250). So to overcome, it's essential to identify these events in a timely and efficient manner.

In this study, we identify these non-recurrent events effectively and inexpensively, by leveraging the potential of social networking sites such as *Twitter, Facebook etc*. From last few years, peoples interests are more inclined towards these sites to express their opinion, feeling and suggestion regarding any problem or event in the form of short text. Twitter is one of such platforms which has more than 335 million[1] monthly active users over the globe, where users interact with each other through a textual/visual post that is

The associate editor coordinating the review of this manuscript and approving it for publication was Gang Mei.

[1] https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/

known as "tweet". That results in a vast amount of data records in the form of posts which are very informative and can be used in a number of applications. As a case study, we consider tier-1 cities in India *(Mumbai, Delhi, Hyderabad, Chennai, Kolkata, and Bengaluru)* to show the city characteristics, i.e. (traffic congestion, accidents, commuters travel habits and road condition) by harnessing Twitter data. We broadly categories the non-recurrent events into three categories i.e. *(accident, traffic, and potholes)*.

Previously, some researchers have dedicated their time to identify the traffic incident by developing an algorithm to spot the event in real time by using the physical sensors [17], [18]. However, these algorithms work well over the highways, but not on local arterials because it is costly as well as difficult to cover every locality under the physical sensor. So in this work, our primary motivation is to establish an efficient and cost-effective system to identify non-recurrent incident in both highways as well as on local arterials. Recently, it has been observed that Twitter data have become a rich source of information pertaining to accidents, congestion, poor lighting, potholes [3], [4], [5], [6], [7] [8], [9], [10], [11]. But it is very challenging to identify events from the tweet texts because tweets post is generally informal, brief, unstructured and often contain grammatical mistakes, misspelling and a lot of noise. That makes a challenging task for researchers, to identify linguistic features for building NLP (Natural Language Processing) based application. It might be due to the restriction imposed by Twitter over tweet post length, i.e. 140 character limits. Thus, it makes text classification and information extraction a challenging problem. So we have performed various data prepossessing steps to convert the text into a readable form.

To segregate the tweets, we proposed semantically similar adaptive keyword generative method by leveraging the semantic context through dense word embedding using Word2vec model. The proposed approach overcomes the shortcoming of traditional methods i.e. *keyword based segregation, and classification by using a machine learning algorithm.* As keyword based approached requires human effort to select manually seed keywords, which might be not sufficient to cover all the tweets. Whereas machine learning algorithm requires crowd-sourced workers to annotate the dataset which itself a difficult task to annotate such a large data and also expensive proposes.

This paper presents a methodology to crawl, pre-process and filter freely available tweets. These tweets post then analyzed to extract non-recurrent events information by using deep learning and Natural Language processing (NLP) techniques. Furthermore, we have identified hotspot with respect to identified events. After that we have compared our proposed model feasibility from the news article and various reports which is published by government departments such as (Hyderabad Traffic Police, Delhi Traffic Police, Mumbai Police Traffic, Kolkata Traffic Police etc.) as well as each cities Municipal department reports such as (Brihanmumbai Municipal Corporation (BMC), Municipal Corporation of Delhi (MCD) etc.).

The main contribution of this work can be summarized as follows:

1) **Semantic Similar keywords:**We have proposed and applying an adaptive semi-supervised method for tweets, by leveraging dense word embedding to identify semantic similar keywords for non-recurrent event's.
2) **Handling Pragmatic Ambiguity:** To address the challenge of existing keyword based methods, so that our proposed method results in less false negative.
3) **Data enrichment:** Dataset can be collected by using multiple sources *(government official traffic accounts, Hashtags, and by using bounding box)*. So that large amount of non-recurrent events data collected.
4) **Mention Based Location Extraction:** Proposed a hybrid approach (an amalgamation of NER, POS, and Regular expression) to identify the location information from the textual content.
5) **Hotspot & Critical location Identification:** The frequency w.r.t each location has been utilized to identify the spatial hotspot.
6) **Temporal Analysis over Weekends (WKND) and Weekday(WKD):** Analysis of commuters travels behavior.

The proposed methodology adds a novel perspective, which can be used by the government agencies to take proactive action before any incident took place and also help in locating the most vulnerable place which should be taken care in a priority manner. However, lack of ground truth data we are not able to find out the accuracy in some of the issues like Potholes related complaint. Thus, we are unable to make any definitive conclusion on the timeliness at current. This paper is an organized as follows, we first discuss the related work, in which we review different techniques to identify the transportation-related issues and also to extract the geo-location from tweet content in section II and brief discussions of Word Embedding and LDA in Section III. Section IV describes the proposed methodology of preprocessing, analyze, and providing the geo-coordinates from tweet text, which is then applied over the different Tier-1 cities in India in section V. At last conclusion are drawn and discussed in section VI.

II. RELATED WORK

We have reviewed the recent work into two parts, i.e. recent study to identify the event from social media data and secondly Location Identification from text content.

A. TRANSPORT EVENT IDENTIFICATION

Various research studies have been done to analyze the use of social media data for purpose of event detection [1]–[5]. Abdelhaq et al. [1] developed a system to track the events evolution with time. In [2], proposed a tool named *"Twitcident"* which is a significant tool that filter, analyze, and search for information in real time during emergency broadcasting services. Krstajic et al. [3] identify the real-world events, by keywords occurrence frequency in the text. Schulz et al. [5] identified a small scale incident by leveraging the semantic web and machine learning algorithm. Furthermore, to detect the space and time of incident more precisely they refined irrelevant content by using spatial and temporal filtering.

Twitter data has been emerged out as a significant source of information concerning traffic incident detection, and several methodologies have been developed to use the Twitter data [6] to detect relevant events. Dandrea et al. [6] developed a traffic event monitoring system over twitter data. In which author used support vector machine (SVM) for classifying the tweets and their approach achieved a classification accuracy of 95.75%.

Gu et al. [7] proposed a framework which identifies the tweet related to traffic incident (TI) and not, by using the adaptive keyword-based approach. The author first crawls the tweets by using some of the keywords which are related to traffic incident. Calculate the frequency w.r.t each token, and check if the keyword is present in the initial keywords list, if not then it will added to the list. At last, the author uses the fuzzy matching algorithm and a regular expression to extract the location information from the tweet text.

Zhang et al. [8] use deep learning model, i.e. Deep Belief Network (DBN) and Long Short-Term Memory (LSTM) to detect traffic accident from social media. The author divided their work into three steps, i.e. firstly feature selection, secondly classification, and thirdly validation. After that, they have performed the classification algorithm over individual and paired tokens. Their experimental results show that over paired token DBN achieve higher accuracy (i.e., 85%) then LSTM.

Wang et al. [9] proposed a tweet-LDA to detect traffic related tweets from social media which is the incremental approach of the keyword-based approach and also the author handle the pragmatic ambiguity. Their experimental results show that their model achieves better accuracy than traditional methods like SVM.

Furthermore, a similar study for identification of traffic-related events has also performed over the Chinese social networking platform, i.e. (Sina Weibo). Chen et al. [10] used CNN with its own continuous bag-of-word (CBOW) model to learn word embedding so that they capture the semantics of words. Their approach achieves better accuracy, to classify traffic related tweets while comparing with the traditional methods like SVM and multi-layer perceptron (MLP). Cui et al. [11] developed a system that extracts information from Sina Weibo (Chinese microblogging website) related to traffic statuses and incident detection by using the natural language processing and machine learning method.

Gutierrez et al. [12] directly extracts the tweets from regional traffic agencies, then identify all events related to traffic. After that determine the location information by using NER whereas Tejaswin et al. [13] have proposed an end to end system, i.e. (Continuous Traffic Management Dashboard (CTMD)) to identify the traffic-related events in real time by harnessing the twitter data.

Some of the researchers have combined traditional approaches, i.e. physical sensors with social media data [14] in order to enhance the precision for detecting real-time incidence and also evaluate the total congestion time due to incidents on highways and their clearance time by incorporating ‘‘real-time’’ of incident from social media.

Zhang et al. [15] identify the incident related tweets by using the LDA (Latent Dirichlet Allocation) and document clustering models. Whereas few number of authors uses Twitter data to predicted the traffic flow [16], [17], [18]. He et al. [16] developed a model by using the linear regression and incorporating with traffic data for long-term prediction of traffic flow where the time span beyond one hour. Their experimental results show that their model outperforms the previously existing auto-regression based traffic flow prediction. Furthermore, authors use Twitter data for short time traffic flow prediction during the sports event like FIFA world cup [17]. Their prediction model incorporated with two features, i.e. tweet rate and semantic features. And its experiment results showed that by using the tweet features in their prediction model improve the traffic flow prediction performance. Grosenick [18] predict the traffic speed on a road segment. For achieving that, they have extracted the non-occurring events related to traffic like accident information from the Twitter data and incorporated with their model to predict the speed on a single road segment by using the neural network model.

However, we have also solved the event identification problem from social media text, but by leveraging the semantics between the words by using word2vec embedding. To the best of our knowledge no one previously used the embedding as an incremental approach to generate similar semantic keywords for classification of tweets in the real time. Furthermore, we have also handled the pragmatic ambiguity so that we may decrease the false results and at last proposed a two parser to extract the location information from textual content.

1) IDENTIFIED THE MENTIONED LOCATION

To recognize the Named Entity from formal documents, state of the art machine learning algorithm like conditional random fields (CRF) [19] have been proposed. These algorithm equipped with comprehensive features like POS tagging and capitalization due to which it achieves satisfactory performance [20]. Based on the CRF algorithm number of NER tools like Standford NER, OpenNLP[2] have been developed and released.

Ritter et al. [21] proposes T-NER system, in which they entirely rebuild the NER pipeline for tweets. They have uses Brown clustering [22] to identify the word variations. For example (‘‘Street’’ and ‘‘st’’) and also identified whether all

[2] https://opennlp.apache.org/

the capitalization is informative or not. This system is also used to identify the geo-location from tweets and achieved an F-score 0.77 in extracting the geo-location. Liu et al. [23] has done similar work, [24] in which they have train the normalization model over tweets to correct the noisy words (e.g. "goood" to "good") before performing NER. They have also focused over capitalization word problem like "chandni chowk" which is hard to label within the short tweets, so they train K-nearest neighbor word classifier to inform NER system with global information. Lingad et al. [25] they have compared various NER tools with standard Stanford NER tools and as well as NER model trained on twitter data. They have concluded that existing NER tolls should be re-trained on social media data before being applied to Twitter data.

Malmasi et al. [26] do not use CRF model to identify the geo-location from tweets. They have proposed an approach based on Noun Phrase extraction and conduct fuzzy matching with Geonames.[3] Their proposed approach achieved F-score 0.792 in shared task ALTA 2014. Gelernter and Balaji [27] have used a combination of gazetteer based location parser, a rule-based street/ building parser and a CRF- based recognizer to achieve better recall. Li et al. [28], [29] observe that often user uses the abbreviation in place of the location name. Zhang et al. [30] system rely on location mention recognizer they have proposed in previous study [31].

Summarization: Traditional NER Tools have many limitations over short and noise tweets. For example *"accident @ chandni chowk st."* due to the noise present in a given tweet creates a problem during location identification from previous tools, while these tools work well over the formal documents. It is because the given tweets in formal documents written as "at" in place of ('@'), "street" in place of ('st') and capitalization "Chandni Chowk" all are absent in tweets, but this limitation is taken care by [21], [23], [24] and achieve and better accuracy. There is still some restriction to identify all the mentioned location from the tweets it's due to unstructured language (no proper use of preposition, verbs, etc) tweets. So to overcome this issue Malmasi et al. [26] proposed an approach based on Noun Phrase extraction and conduct fuzzy matching with Geonames and able to achieve better results from the previous study. The author uses the Geonames database which contains most of the POI location names. But not all the location names it covers so to overcome this we make enlarge database for location names and POI names from different resources such as Wikipedia,[4] geonames and Floor address[5] to achieve better accuracy.

III. PRELIMINARIES

A. WORD2VEC EMBEDDING

Each unique word in the vocabulary has to be represented using a vector. Using binary number for each word introduced unrelated dependencies between unrelated words. This issue is addressed with a one-hot vector encoding for words. But this notation resulted in each vector for a word having the length equal to the size of the vocabulary, which was impractical to work with even for moderately sized vocabularies. This was addressed with the development of word embeddings for individual words. Each word in the vocabulary of the dataset is represented as a unique vector. Thomas Mikolav et al. [32], [33] have developed a system that generates a word vector for each word in the vocabulary that represents the semantic and syntactic meaning with the help of a unique neural network model, namely skip-gram and a continuous bag of words model. These models generate the word to capture the context associated with any word in a shallow window.

This model aims to predict between a center word and context words in terms of vectors. For this they have used the two algorithm that is *skip-gram(SG)* which predict context words for a given target, where as another one is *Continous Bag of Words(CBOW)* which predict target word from bag of words context. For creating word embedding we firstly train Word2vec model over the collected tweets. We consider each tweets as a list of tokens $\{t_1, t_2, t_3 \ldots, t_n\}$ and then for every token we have selected a window of size m. It could be simplified as for token t_i the window is $\{t_{i-m}, t_{i-m+1}, \ldots, t_{i+m-1}, t_{i+m}\}$. Then predict the context of every token t_i by considering the embedding of every token.

$$out(token, \theta) = T(\theta)W[token]$$

where the *out* function approximator estimates the probability of a particular token presence in the window of token under consideration. θ are the weights of the function approximator, and W is the size of embedding that is dimensional vector lookup table corresponding to each token denoted as d_{wrd}.

$$L_1(B;\theta) = \frac{1}{B}\sum_{token \in B} out(token,\theta) log P_i^{wrd}$$

where P_i^{wrd} finds the probability of how much that word lies in the context of a token.

We have used the skip-gram for training our model. Working of the skip-gram algorithm is shown in figure 1. In which we input the target word to Skip-gram model, in results it gives the surrounding words for a given target word. For example, in the sentence *"Accident near Vikhroli flyover traffic too much"* input would be *traffic* whereas the output is *"Accident", "near", "Vikhroli", "flyover", "too","much"*, if we assume window size 5. Skip-gram model contains one hidden layer whose dimension is smaller than output/input vector size. At the end of the output, the Hierarchical softmax activation function applied.

IV. METHODOLOGY

The proposed method uses non-recurrent events related keywords as seed words for crawling the tweets using Twitter Streaming API. Our main is to categories the events into three *congestion, accidents, potholes* and after that identify

[3] http://www.geonames.org/

[4] https://en.wikipedia.org/wiki/Category:Roads_in_Delhi

[5] www.flooraddress.in

TABLE 1. Examples of tweets before and after executing the pre-processing steps i.e hashtag & handle removal, url removal, typo correction, abbreviation, and redundant consecutive character removal (RCCR).

Operation	Before	After
Hashtag & Handle removal	#Delhipolice #delhiCM Heavy traffic at Kalindi kunj towards jaitpur road due to unnecessary auto stand @dtptraffic pic.twitter.com/lGAZUsC4eZ	Heavy traffic at Kalindi kunj towards jaitpur road due to unnecessary auto stand.pic.twitter.com/lGAZUsC4eZ
URL removal	Heavy traffic at Kalindi kunj towards jaitpur road due to unnecessary auto stand. pic.twitter.com/lGAZUsC4eZ	Heavy traffic at Kalindi kunj towards jaitpur road due to unnecessary auto stand.
Abbreviation & Slang Replacement	plz do smthing asap there were No. of accidents near Azadpur mandi Delhi @DelhiPolice @dtptraffic	Please do something as soon as possibleas there were No. of acc -idents near Azadpur mandi Delhi @DelhiPolice @dtptraffic
RCCR	pls help stuckkk near #ISBT #Kashm- -irigate #Delhi no traffic police.	pls help stuck near #ISBT #Kashm- -irigate #Delhi no traffic police.

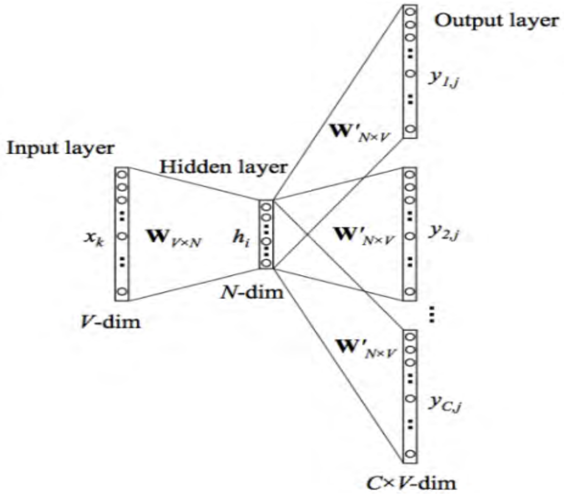


FIGURE 1. Working of skip-gram algorithm.

those locations where relatively higher tweets we are getting for a particular issue. To achieve this, we have divided our methodology into four steps. In the first step, we perform data pre-processing as collected tweets contain a lot of metadata like URL, personal information, etc. which is of no use for the proposed use case. Therefore, we remove irrelevant data. Secondly, we use Word2Vec model to generate similar keywords with respect to seed words. In the third step, we remove the Pragmatic Ambiguity from segregated data based on the above keywords. At last, we perform content-based location identification from a textual tweet.

A. DATA PREPROCESSING

We have collected data by using twitter streaming API.[6] Data crawling can be done in two ways, i.e. by using *hashtags* (H) and another one by using *bounding box* (L). In this study, we collected dataset by using both ways. Twitter post are generally informal, brief, unstructured and often contain grammatical mistakes, misspelling and a lot of noise. This is due to the 140 character limit imposed on tweets. Authors from the previous study claim that users knowingly use the abbreviation, shortened, slang words and also uses an amalg -amation of prefix and suffix of the word [34]. So it becomes very cumbersome to understand some of the tweets like *"Hvy trafic at strt of andheri brdge going 2wrds aiprt & further"*. To overcome this problem previous study [34], [35] uses different text mining techniques like Edit Distance, Longest Common Sub-sequence and Prefix_Suffix match to handle noisy text in SMS. We have used the same approach so that that tweet post could be converted to a readable form like *"Heavy traffic at start of andheri bridge going towards airport & further"*. In table 1 we have shown tweet example of before and after executing the pre-processing steps.

Steps involved in Pre-processing:

- **Plain Text Extraction (PTE)** While collecting the tweets it contains a lot of extraneous information which is not relevant, so we save only the informative information in the given format *(tweetID, Date&Time, Tweet, Location, Hashtags, Mention, Geo-coordinates)*. And all the irrelevant information were removed as shown in table 1.
  - **Hashtag Removal:** # hashtags were used by users before any relevant keyword or phrase to categorize their tweets and show their tweets more easily in Twitter search. So hashtags symbol # were removed as they carry no relevance.
  - **URL Removal:** it is used to share web resources. But generally, they carry no relevance information. So we remove it.

[6] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html

  - **Handle Removal:** "@" is used in Tweet by the user to tag or refer other Twitter users so that they may follow the tweet. So we remove "@" from the tweets.
- **Remove Stop Words:** Stop words are common words in every language. While there use in the Language is crucial, although these words don't mean a lot in the sentence, these words are especially adverb, articles, conjunction etc. were known as stop-words. Stop-word removal is an essential step in text pre-processing. We have used the stop-word list in the NLTK library.
  - **Remove Punctuation** The marks, such as full stop, comma, and brackets, were the common punctuation used to isolate the sentences so that we can understand the elements easily. But often these elements do not present any significant information, so these punctuation marks were removed from the tweets.
  - **Remove RT:** RT means retweet which is used in twitter with syntax 'RT:@username', and followed by tweet text. It is used to share someone post directly, or the user may add some comments, before sharing in his/her timeline.
- **Typo Correction:** Twitter post i.e. *tweets* were generally informal, brief, unstructured and often contain grammatical mistakes, misspelling and a lot of noise. We have listed out types of noise present in a tweet as shown below in table 1.
  - **Abbreviation & Slang Replacement:** Due to character limit over tweet post enforces users to make use of slang words as well abbreviation. So it is very cumbersome for a machine to understated the post which contains a lot of noise. To handle that we have to first convert the post in readable form. For that, we make use of online available slang word list and abbreviation. So to create more robust we manually identified a few more slang words i.e. those words which are common among Indian people and added with available online list.[7]
  - **Redundant Consecutive Character Removal (RCCR)** Some time users uses repeated characters like *("Trafficcccc"for "traffic"), ("stuckkkk" for "stuck")*. So were we find these type of words in the text, which have more than two consecutive character we replace with single character.

TABLE 2. List of similar words generated with respect to each keywords using Word2Vec model.

Categories	Top five Related keywords
Accident [C1]	accidents, crash, wreck, collision, crashes
Potholes [C2]	potholes, roads, road, manholes, holes
Traffic [C3]	traffics, congestion, jams, stuck, intersections

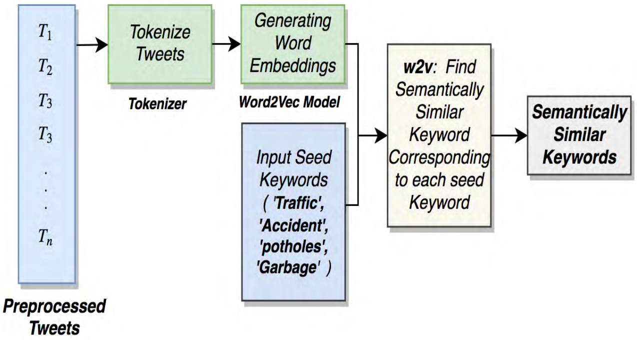

T_1
T_2
T_3
T_3
.
.
.
T_n
Preprocessed Tweets
Tokenize Tweets
Tokenizer
Generating Word Embeddings
Word2Vec Model
Input Seed Keywords ('Traffic', 'Accident', 'potholes', 'Garbage')
w2v: Find Semantically Similar Keyword Corresponding to each seed Keyword
Semantically Similar Keywords

FIGURE 2. A Framework for semantic similar keywords generation.

B. SEMANTIC SIMILAR AUTOMATIC KEYWORD GENERATION

We initially have some of the seed keywords related to different categories corresponding to non-recurrent issues face by commuters in day to day life such as *Accident, Potholes, Traffic* it could be simplified as $\{C_1, C_2 \ldots..C_n\}$. Above categories typically center most of the incident issues. These set of category determined by domain experts and by examining the data. In this study, we have explored the data to classify the tweets corresponding to categories shown in table 2. Now, to segregate the tweets by using seed keywords (SD_KD) have some limitation, such as these set of $\{s_1, s_2 \ldots..s_n\}$ keywords are not sufficient to cover every tweet regarding above categories face by commuters. So to cover all the tweets, we have to consider the semantics of the problem from social media text. This problem closely resembles the same issue of finding a topic from the textual content, i.e. Topic Modeling. Latent Dirichlet Allocation (LDA) [36] is one of the most often used technique for modeling topic. But this technique works over the word co-occurrence pattern. Due to this reason, LDA is not suitable for the sparse, short text of social media [37]. So, to extend the seed keywords list, we use the generalized word embedding representation that is Word2Vec model [32]. For this, we consider each tweet as a list of tokens $\{t_1, t_2, t_3 \ldots, t_n\}$ and for each token we select a window of size m. It could be simplified as for token t_i the window is

$$\{t_{i-m}, t_{i-m+1}, \ldots., t_{i+m-1}, t_{i+m}\}$$

The word2vec model predicts the context of every token by considering the embedding of every token. We have used the most_similar function of Word2Vec model to find a semantically similar word for a given input word. A framework for generating similar semantic keywords shown in figure 2. For the same, a stepwise procedure is explained in Algorithm 1. So the expansion of seed keyword dictionary is made after the inclusion of keywords by using the word2vec embedding (W2V_KD).

$$W2V\_KD' = E(SD_K D)$$

[7] https://www.noslang.com/dictionary/

Algorithm 1 Semantically Extended Keywords Generation

```
Input: Clean tweets $\tau$={$\delta_1, \delta_2, \delta_3 \ldots\ldots \delta_n$}
Output: Expended keyword list $W2V\_KD = E(\mathcal{SD\_KD})$
1: res1=[' ']
2: res3=[' ']
3: res2=[' ']
4: for each tweet in $\tau$ do
5:     $\mathcal{T} = nltk.tokenize(tweet)$
6: end for
7: $model=models.gensim.Word2Vec(\ \mathcal{T})$
8: $\mathcal{SD\_KD} = \{s_1, s_2 \ldots\ldots\ldots s_n\}$
9: res3=$\mathcal{SD\_KD}$
10: for each keyword in $SD\_KD$ do
11:     res1=model.most_similar(Keyword, topn=5)
12:     res2.append (res1)
13:     res2=unique (res2)
14:     if $(res2 - \mathcal{SD\_KD}) = \phi$ then
15:         Stop
16:         Exit
17:     else
18:         res3= res2-$\mathcal{SD\_KD}$
19:         $\mathcal{SD\_KD} = \mathcal{SD\_KD} \cup res3$
20:         goto step 10
21:     end if
22: end for
```

where E is an expansion operator and W2V_KD' is the superset of SD_KD:

$$W2V\_KD' \supseteq E(SD_K D)$$

A list of top five similar keyword examples corresponding to Accident (C1), potholes (C2), Traffic (C3) are shown in Table 2. After that, we segregate the tweets by using the W2V_KD keywords corresponding to each category. Our proposed model is entirely dependent over the initial seed keywords, that is if we choose a seed keyword that is sparse in our dataset than it will not return most semantic keywords. So, to handle this issue, we have taken only those seed keywords as an input to word2vec model which are used during crawling so that the problem of sparsity is resolved.

C. DISAMBIGUATION PROCESS FOR REMOVING PRAGMATIC AMBIGUITY

We have crawled the tweets w.r.t hashtags related to different categories as shown in table 2. However, some hashtags like "traffic" can be used for different purpose, for example, *"80% of Internet traffic will come from video. And one day, people will accept that I can shoot video with this darn phone and it will be nice. #DSPhilly #WECANDOTHIS #Internet #traffic"*, *"#Intrnet #Traffic The majority of Internet traffic is not generated by humans, but bots like Google and Malware."*. This is a cumbersome problem to remove ambiguity. As we know this type of tweet post is not relevant to our work. So to identify the tweets related to different traffic events, it is essential to identify the information about an issue or topic.

TABLE 3. Two dimensional $vectorscore_{m*n}$ between each $t_i \in t$ & $K_p \in K$.

	t_1	t_2	..	t_i	..	t_n
K_1	0.7	0.14	0.4	0.5	0.0	0.2
K_2	0.8	0.8	0.13	0.6	0.12	0.3
.	.	.	..	.	..	.
.	.	.	..	.	..	.
K_p	0.1	0.35	0.9	0.2	0.0	0.1
.	.	.	..	.	..	.
.	.	.	..	.	..	.
K_m	0.3	0.19	0.20	0.6	0.01	0.3

TABLE 4. Word2Vec similarity between terms present in tweets and topics related to traffic events.

	crash	lane	flyover	intersection	animal
accident	0.74	0.14	0.41	0.5	0.0
traffic	0.34	0.19	0.20	0.6	0.01
potholes	0.15	0.35	0.09	0.2	0.0

In our proposed Framework, we are addressing the challenge of keywords based method i.e pragmatic ambiguity. After manual inspection over some random tweets from the collected tweets we get to know that noun and adverb were important to match efficiently. So we use POS tagging and store only Noun and adverb corresponding to each tweets. Subsequently we match tweets $(t) = \{t_1, t_2, t_3 \ldots\ldots, t_n\}$ from the keywords list with respect to each categories $K = \{C_1, C_2 \ldots C_j\}$ where $\{1 \leq j \leq 3\}$. List of categories with keywords were shown in table 1. After that we calculate the similarity between the t_i and keyterms corresponding to each categories K_i. For calculating the similarity we have use Word2Vec (W2V) model. To find whether t_i is related or not related with one of the categories (K). To achieve this we firstly check for each $t_i \in t$ where tweet post $t_i \in \{NN, NNP, NNS, NNPS\}$. Then $\forall k_P \in K | K = \{K_1, K_2, K_3, \ldots . K_m\}$, apply the semantic similarity with each t_i as W2Vt_iK = Word2VecSimilarityt_i, k_P. Now we get two dimensional $vectorscore_{m*n}$ between each $t_i \in t$ & $K_p \in K$ as shown below in table 3.

If $\exists K_p \in K : W2Vt_iK \geq th$ then we can say information/topic is said to be true for tweet t. If $\forall t_i$ and $\forall k_p \in K$, the similarity score $W2Vt_iK \leq th$, then we say topic/ information is said to be false for t. We have taken threshold value *th=0.2*. Table 4 shows the similarity between the various noun terms present in the tweet and category related keywords.

D. MENTIONED BASED LOCATION IDENTIFICATION

A very few (0.01%) number of tweets in our data set have Geo-coordinates. The availability of coordinates is utmost important for the effective resolution of the spatial problems. This information is only available if users share their GPS location while posting a tweet. The location information is available in different forms in a tweet. First one is the user profile location, but this cannot be used to detect the location in real time, because it might be the same location

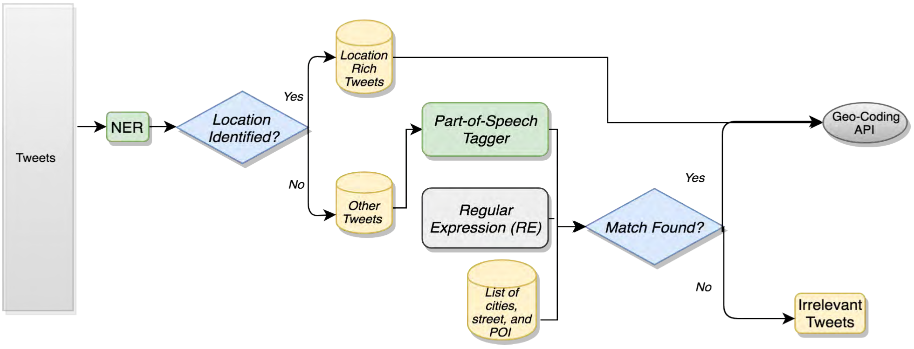


FIGURE 3. A Content based Hybrid module for Location Identification.

TABLE 5. Location extraction from tweets by using hybrid module (NER and POS Tagger).

Example of Tweets	NER / POS
1) A East Delhi Dilshad Garden metro station Circle side roads side traffic seen parking white yellow auto Ghaziabad challan made buses park taking passengers stop Anand vihar side road. 2) roads leading Fortis Hospital Shalimar Bagh Delhi always choked Traffic problem roads leading hospital dangerous 3) Please look area dividing road rohini sector sector near wall Sarvodaya Vidayalaya sector evening hours boys parked cars roads start taking snacks liquor leads unethical traffic.	NER
4) Accident Due To Wrongly Placed Dividers On Road Heavy Traffic Jams near Cantt Road Infront Of Odean Cinema. 5) MCD roadside traffic issue people parking space vehicles causes congestion in Guru Nanak locality . 6) traffic jam at race course area Kemal Ataturk event Suggestion widening road airforce dassapalla. 7) traffic stopping place taxi hawkers bus satand cause traffic jam Same story for mukarba chowk shops buses passengers.	POS

at the time when the account is created. Secondly, directly Geo-coordinates available or city/place name but it is very few in our case. Therefore, it is important to mine the location information from textual content. For that, we propose a hybrid module which is an amalgamation of Named Entity Recognition (NER),[8] Part-of-Speech tagger (POS) and Regular Expression (RE) the detailed architecture of proposed module is shown in the figure 3.

In this, module we first use the Named Entity Recognition(NER) to identify the location. NER identifies named entities from the text & then classifies them into predefined categories like location, organization, person, etc. But for this work, we classify tweets $T = \{t_1, t_2, t_3 \ldots . .t_n\}$ on basis of location only. So, to identify the location informative tweets NER is used, i.e., $NER_T = NERtagger\{T\}$. Some of the examples of location identification by NER were shown in Table 2. But NER does not recognize all the tweets which contain the location information. It might be one of the reason that in India most of the road names and locality name were supposed to be the name of some renowned person. So, it can be the reason for unmatched tweets. As shown in the following examples: *"Too much traffic near hotel royal nest in Ashok Vihar stuck around* 2 *hour in jam"* In the above example NER wrongly identify *"Ashok Vihar"* as a person instead of location.

So we don't neglect remaining tweets, i.e. $S = \{T - NER_T\}$, as some of them in S contains location information in the tweet itself. As shown in table 5, in which some of the examples are recognized by NER and remaining posts which contains the location information is identified by POS tagger.

To achieve this, we firstly identified location indicator terms from our dataset. After that these set of terms classified into Noun indicators, Preposition and Location indicator phrases such as *street, town, living near, residing at, at, from,*

[8] https://nlp.stanford.edu/ner/

to, etc.. Then we have applied a large set of Regular expressions(RE) to extract the street, road names and POI names, etc from it. But due to the noisy nature of tweet text, it is difficult to convert into latitude and longitude. So to overcome this problem we made an automatic web crawler to extract information about the road names, Street names and colony names from different resources like Wikipedia,[9] geonames,[10] commonfloor[11] and store in a dataset (D) and also create list of POI names for different cities individually store in POI list. Then we apply the string matching between the names extracted from the REs and the names store in (D) and POI list. We employ edit distance as a string matching [35] with a maximum difference of length two. If it is matched then passed to the Google place API to convert into coordinates. Some of previous work [21], [23], [24], [26] has also uses same methodology for identification of mentioned location from tweets. But these studies don't have a combination of above all the methods and also list created in the previous study is mainly from Geonames or only from one source. But in our work, we have taken names from different resources so that the most robust name list we can create.

E. SPATIAL HOT-SPOT DETECTION

Spatial hot-spot detection is a problem of finding the location where objects/points were anomalously high. Geo-spatial clustering/ spatial partitioning [38], [39] were different from Spatial hot-spot detection as it is formed where the intensity of objects/ points is higher than the outside. There are a number of applications where hot-spot detection has been used significantly such as public health, epidemiological and recently literature increased in criminology, where finding a hot-spot might help for an official to locate/detect the criminals [40].

We have described in section IV, to extract location from the tweet content and convert into latitude and longitude by using the Google API and this whole process is known as geocoding. To find hot-spot corresponding to different categories as shown in table 2. We cluster those points which are identical and shares the same coordinates. Based on this we plot these points over the map.

V. CASE STUDY FOR TIER-1 CITIES IN INDIA

To show the usefulness of our proposed system we have chosen tier-1 cities in India i.e. *Delhi, Mumbai, Bengaluru, Hyderabad, Kolkatta and Chennai* as a case study. This type of system can help the government to know the places or hotspot from where most of the complaints related to different issues in transportation, faced by commuters in their day to day life. This type of study also helps in identify users tweeting pattern w.r.t each city.

A. EXPERIMENTS AND RESULTS

TABLE 6. Dataset statistic by using Hashtags (H) and Location (L) with respect to Tier-1 cities.

City	No. of Tweets	Unique Tweets
Bengaluru **(H)** **(L)**	254823 433133	107727 373903
Chennai **(H)** **(L)**	201435 172262	59884 137236
Kolkata **(H)** **(L)**	98624 205907	47404 178830
Mumbai **(H)** **(L)**	833574 820313	267079 686943
Delhi **(H)** **(L)**	911539 757885	253786 640502
Hyderabad **(H)** **(L)**	120728 226436	53354 188495

1) DATASET DESCRIPTION

We have collected data by using twitter streaming API.[12] Data crawling can be done in two ways, i.e. by using *hashtags* (H) and the other one by using *bounding box* (L). In this study, we collected dataset by using the above two methods and collected approx 60 million tweets during May-03-2018 to August-23-2018. This case study focus over the tier-1[13] cities in India. So we crawled the dataset by using the different *'#' hashtags such as (congestion, potholes, accident, road, traffic, intersection, crash, highway,construction, etc.), '@' government official Twitter handle i.e. (mumbaitraffic, DelhiTraffcPol, blrcitytraffic, hydcitypolice, KPTrafficDept etc.)* with respect to the cities and also by using whole India bounding box. The dataset statistics about a number of tweets and unique tweets with respect to Hashtags (H) and Location (L) corresponding to each city where shown in table 6.

2) TWEETS SEGREGATION BY USING WORD2VEC

Now we segregate the tweets by using keywords that are semantically similar as explained in section IV. In table 7 shows the comparison between tweets segregate using Word2vec approach (W2V_KD) and by using the Seed keywords (SD_KD). In table 7, we also showed the tweet distribution over differently collected dataset methods, i.e. Hashtags and handle (H) and Bounding Box (L). We found that tweet frequency relative high corresponding to potholes, accident, and traffic in Mumbai as compared to other tier-1 cities.

Our proposed method that is Word2vec is able to classify 40% more tweets than by using the initial SD_KD. It is depicted clearly from figure 4.

[9] https://www.commonfloor.com/delhi-city
[10] http://www.geonames.org/
[11] https://www.commonfloor.com/delhi-city

[12] https://developer.twitter.com/en/docs/tweets/data-dictionary/overview/tweet-object.html
[13] https://en.wikipedia.org/wiki/Classification_of_Indian_cities

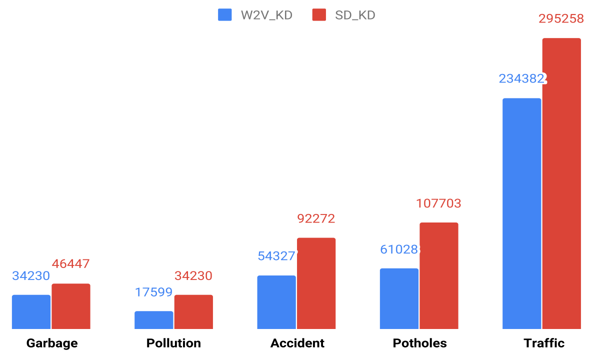


FIGURE 4. Comparison between Seed keywords (SD_KD) and by using Word2Vec approach (W2V_KD).

3) TEMPORAL ANALYSIS

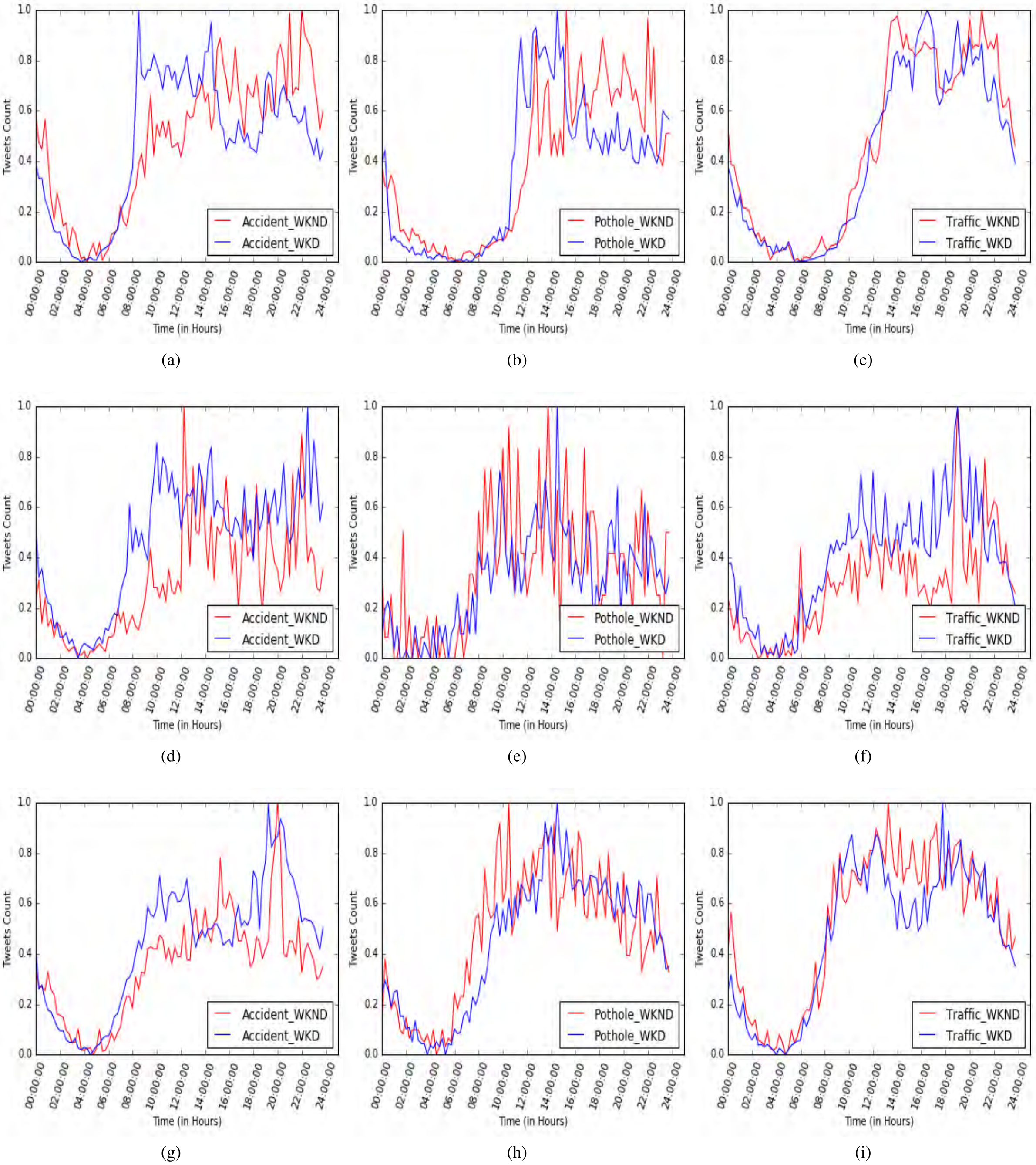
We obtained the classified tweets as an output from the proposed approached. After that we have perform temporal analysis i.e to plot incident frequency with respect to time for different categories such as *{"Accident (C1), Traffic (C2),Potholes(C3)"}* as shown in figure 5 and 6. And also try to find out people tweeting behavior pattern with respect to categories over weekdays (WKD) and weekends (WKND). This type of analyses may help the government agencies to take some proactive action related to the issues. So that commuters or people live their life incident-free over killer roads due to potholes. Below we have Briefly described tweets distribution over WKD and WKND w.r.t each category.

- **Accident Tweets Distribution over Weekdays & Weekends:** We segregated the tweets corresponding to categories, i.e. Accident (C1), Traffic (C2), and Potholes (C3) by using the proposed model. From figure 5 {(a), (d), (g)} and figure 6 {(a), (d), (g)} shown peoples tweeting behaviour corresponding to accident. Users from Mumbai, Chennai, Hyderabad, and Bengaluru follows the same pattern that they tweet more often while traveling during WKD, i.e. *(morning session from* 8 *A.M to 10 A.M and evening session from* 5 *P.M to* 8 *P.M)*. While peoples from Delhi and Kolkata follow a different pattern, individual tweets during the working hour *(morning session from 10 A.M to 11:45 A.M)*. From the graph, we have also concluded peak (highest frequency related to accidents) time, i.e., after the working hour except in Mumbai. While the tweeting behavior pattern on WKND is totally different from WKD, i.e. people mostly tweets after the afternoon session expect Bengaluru and Delhi in which people start tweeting around 10 A.M. At last by using the tf-idf score, we find out the main reasons for accidents are drink and drive, lane-cutting, potholes, intersections, street light not working, due to road condition, wrongly place speed breaker, and one of the most important reasons road construction. The ministry of road transport and highways are shown in the report that deaths due to road construction are increased by 50% from 2016.

TABLE 7. Category wise tweet segregation by using (SD_KD) and (W2V_KD) over differently collected dataset methods i.e Hashtags and Handles (H) and by using Bounding box (L).

Keywords	Bengaluru 254823(H) 433133 (L)		Chennai 201435 (H) 172262 (L)		Kolkata 98624 (H) 205907 (L)		Mumbai 833574 (H) 820313 (L)		Delhi 911539 (H) 757885 (H)		Hyderabad 120728 (H) 226436 (L)	
	SD_KD	W2V_KD	SD_KD	W2V_KD	SD_KD	W2V_KD	SD_KD	W2V_KD	SD_KD	W2V_KD	SD_KD	W2V_KD
Pothole												
(H)	5508	6115	1030	1115	831	907	43865	87512	5927	6612	1957	2193
(L)	290	391	14	37	22	56	1345	1922	159	610	80	238
Accident												
(H)	8431	7553	4711	7010	2284	7326	16581	29252	15798	27192	3937	4710
(L)	400	1378	92	382	138	625	841	3112	819	2956	295	776
Traffic												
(H)	32464	35273	22988	63020	8534	8783	84815	86096	58663	65913	10805	11523
(L)	2170	2658	4125	9125	258	491	5145	5612	3389	4995	1026	1769

- **Potholes Tweets Distribution over Weekdays & Week- ends:** The ministry of road transport and highways releases statistics of fatalities due to killer-potholes



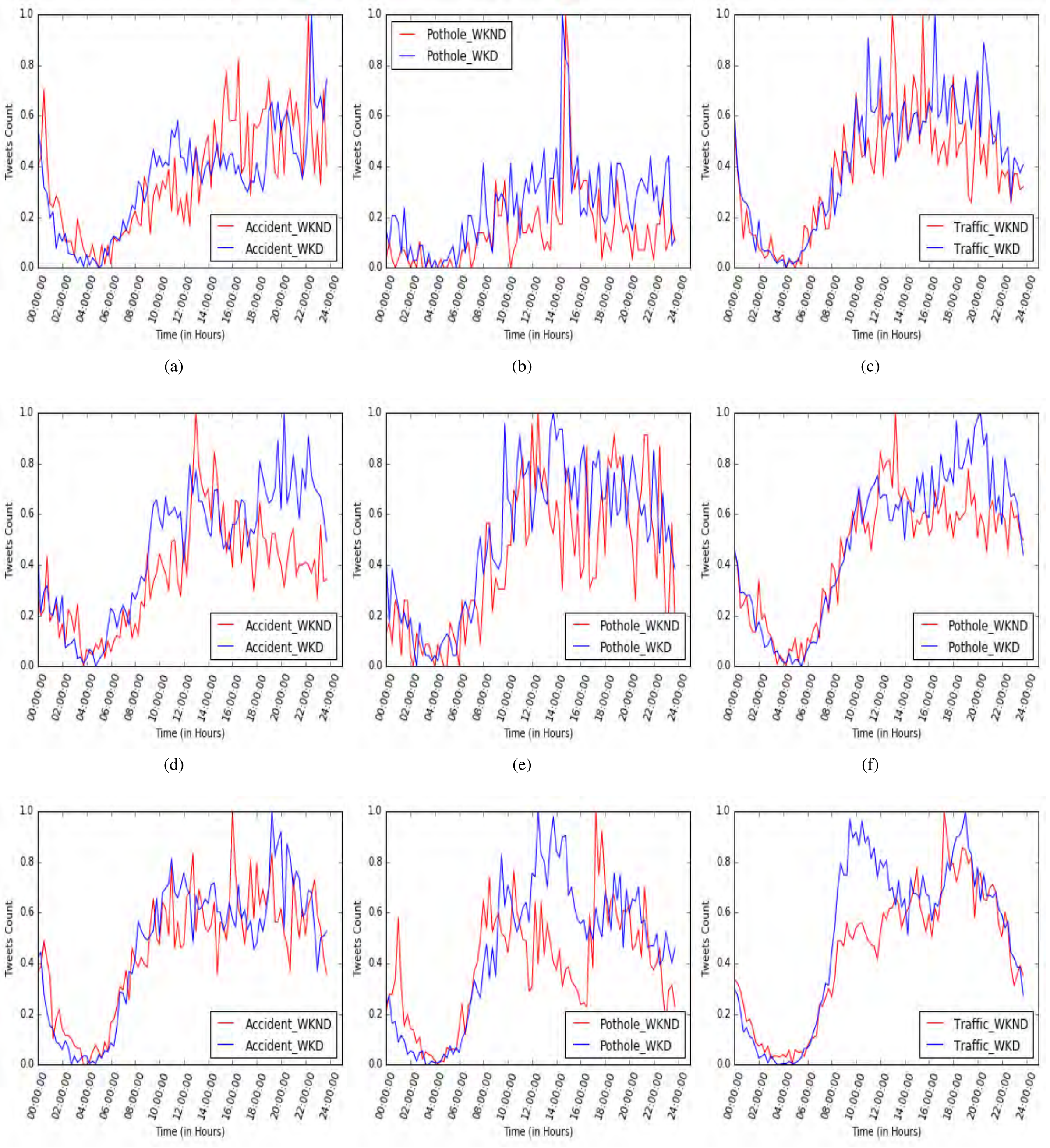
(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIGURE 5. Variations in tweets frequency w.r.t time over different category (Accident, Pothole, and Traffic) for Mumbai, Chennai, and Delhi city. (a) Accident_Tweets_Mumbai. (b) Potholes_Tweets_Mumbai. (c) Traffic_Tweets_Mumbai. (d) Accident_Tweets_Chennai. (e) Potholes_Tweets_Chennai. (f) Traffic_Tweets_Chennai. (g) Accident_Tweets_Delhi. (h) Potholes_Tweets_Delhi. (i) Traffic_Tweets_Delhi.

on roads from the last five year is 14,926 which is much higher than the fatalities due to terrorist or Naxal attacks. So it is important to handle this issue in a priority manner. Tweet distribution corresponding to potholes related issues as shown in figure 5 {(b), (e), (h)} and figure 6 {(b), (e), (h)}. From that we have found out, that people from Chennai, Delhi, Hyderabad, and Kolkatta follow the same tweeting pattern on WKD as well as on WKND. While users from Mumbai and Bengaluru have different tweeting pattern w.r.t to WKD and WKND. Individual from Mumbai on WKD posts issue during *(10 A.M to* 2 *P.M)* whereas on WKND they usually tweet in the afternoon and evening session *(1 P.M to* 3 *P.M and* 6 *P.M to 10 P.M).* Whereas users from Bengaluru on WKD tweets mostly during the morning session (8 A.M to 10 A.M) and an



(a) (b) (c)

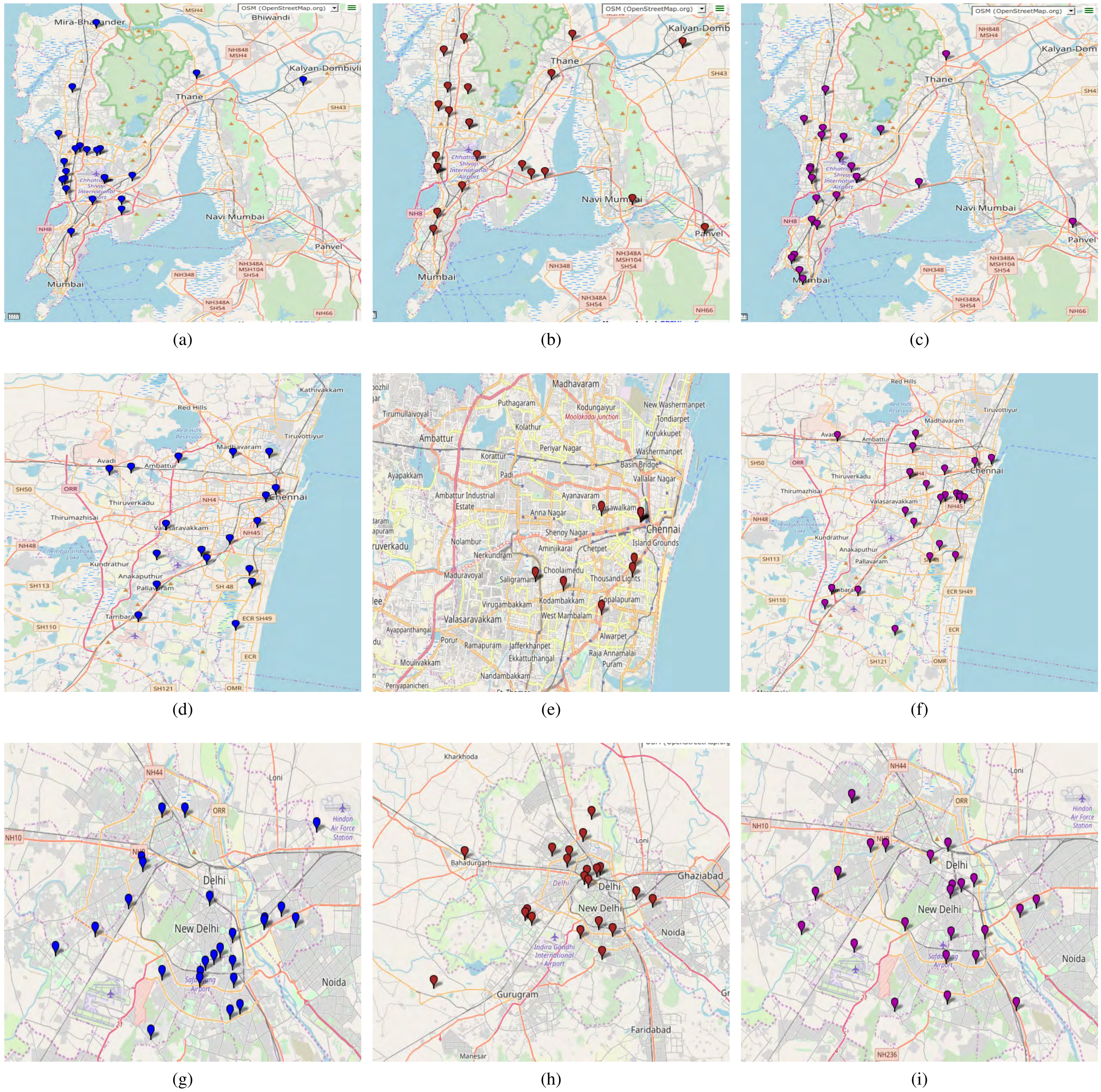
(d) (e) (f)

(g) (h) (i)

FIGURE 6. Variations in tweets frequency w.r.t time over different category (Accident, Pothole, and Traffic) for Kolkata, Hyderabad (HYB), and Bengaluru (BLR) city. (a) Accident_Tweets_Kolkata. (b) Potholes_Tweets_Kolkata. (c) Traffic_Tweets_Kolkata. (d) Accident_Tweets_HYB. (e) Potholes_Tweets_HYB. (f) Traffic_Tweets_HYB. (g) Accident_Tweets_BLR. (h) Potholes_Tweets_BLR. (i) Traffic_Tweets_BLR.

evening session (1 P.M to 4 P.M) and during WKND (4 P.M to 8 P.M). Futhermore, we have concluded that individual talks about the potholes related issues through out the day. At last, by using the tf-idf we able to identify the reasons which are mainly responsible for potholes such as *(waterlogging, constructions, inadequate drainage, overloaded vehicles, and poor maintenance)*.

- **Traffic Tweets Distribution over Weekdays & Weekends:** Four Tier-1 cities *(Mumbai, Delhi, Kolkata, and Bengaluru)* losses more than 22 billion dollars annually due to congestion and commuters has to spend more than one-and-a-half hour longer to travel during peak hours as compared to non-peak hours. To curtail this issue, we have performed analysis on tweets distribution related to congestion over Tier-1 cities in India as shown



(a) (b) (c)

(d) (e) (f)

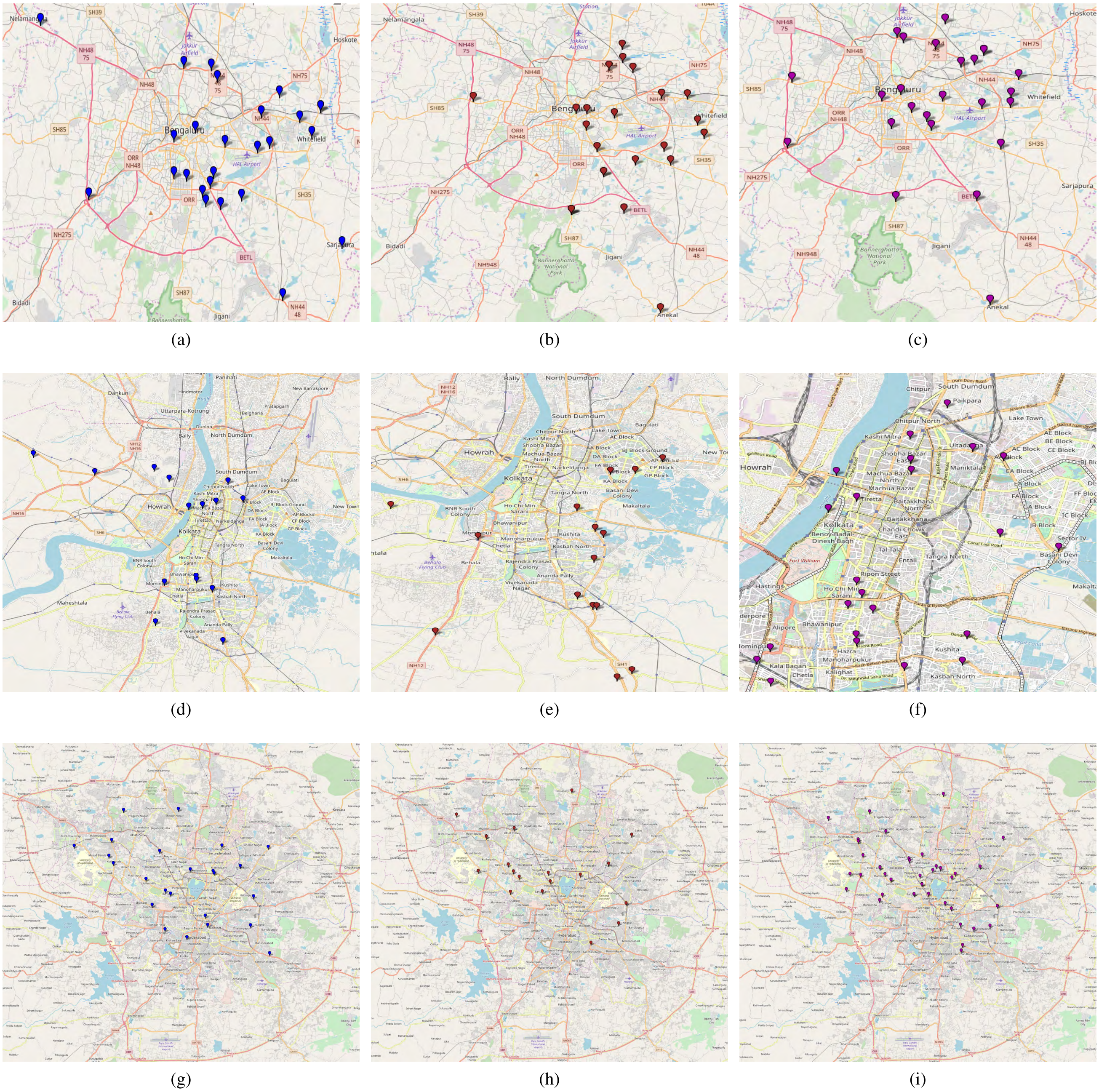
(g) (h) (i)

FIGURE 7. **Top twenty five hotspot with recpect to each category over different city Mumbai, Chennai, and Delhi. (a) Accident_Tweets_Mumbai. (b) Potholes_Tweets_Mumbai. (c) Traffic_Tweets_Mumbai. (d) Accident_Tweets_Chennai. (e) Potholes_Tweets_Chennai. (f) Traffic_Tweets_Chennai. (g) Accident_Tweets_Delhi. (h) Potholes_Tweets_Delhi. (i) Traffic_Tweets_Delhi.**

figure 5 {(c), (f), (i)} and figure 6 {(c), (f), (i)}. The individual from Chennai, Hyderabad, Bengaluru, Kolkata, and Delhi follows the same pattern, i.e. they tweet during traveling. Expect Mumbai, in which users tweets mostly in the working hour on WKD. While on WKND users more often tweets during the evening session *(1 P.M to 4 P.M)*. We have concluded that during the peak hour, i.e. (8 A.M to 10 A.M) and (5 P.M to 7 P.M) people face a lot of problems due to congestion. We have also discovered some of the reasons of traffic congestion such as *water logging, heavy rain, the absence of sign board, street light not working, violation of traffic rules, scattered garbage on roads, road constructions, and bad road conditions, a lot of potholes, and wrongly placed vehicles*.

B. HOTSPOT DETECTION ANALYSIS

This section shows the geospatial analysis of civic complaints like *{"Accident (C1), Traffic (C2), Potholes (C3)"}*. To identified the location from the above complaints we used a proposed method which is an amalgamation of NER, POS, and set of Regular Expression rules as explained in



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

FIGURE 8. Top twenty five hotspot with recpect to each category over different city Bengaluru (BLR), Kolkata, and Hyderabad (HYB). (a) Accident_Tweets_BLR. (b) Potholes_Tweets_BLR. (c) Traffic_Tweets_BLR. (d) Accident_Tweets_Kolkata. (e) Potholes_Tweets_Kolkata. (f) Traffic_Tweets_Kolkata. (g) Accident_Tweets_HYB. (h) Potholes_Tweets_HYB. (i) Traffic_Tweets_HYB.

section IV to retrieve the location information from textual content. After that, we have used Google Geocoding API to convert extracted location into latitude and longitude. After that, we segregated the tweets corresponding to the same area and calculate the frequency w.r.t location for all consecutive days. We have chosen the threshold value 5 (i.e., if the particular location frequency is 5 in a day then that place known to be a hotspot). Figure 7 and 8 shows the top 25 hotspots identified with respect to different category over tier-1 cities in India as well as table 8 shows the top ten areas which comparatively receive high complaints regarding the above category.

We have also identified some of the critical locations that are present in all category for a particular city.

- **Bengaluru:** Whitefield Main Rd, Outer Ring Rd, and Sarjapur Main Rd.
- **Mumbai:** Andheri East & West, Thane, Sion Panvel Expy and Bandra.
- **Delhi:** Dwarka, Punjabi Bagh road, and shastri park.
- **Chennai:** St.Thomas Mount & porur,
- **Hyderabad:** Banjara Hills, Jubliee hills, Uppal road, & Gachibowli

TABLE 8. Top Ten hotspot detected corresponding to tier-1 cities in India.

Cities	Transportation Isuue		
	Accident	**Potholes**	**Traffic**
Bengaluru	Sarjapur, Whitefield Main Rd, Outer Ring Rd, Chandapura, Bommanahalli, Koramangala, Hebbal Flyover, NH275, Jayanagar, Harlur Main Rd	Magadi Main Rd, Outer Ring Rd, Bannerghatta Main Rd, Madiwala, #26, Gundappa Building, Church Street, Varthur, Siddapura, Whitefield Main Rd, No.64, Lashkar-Hosur Rd	Hebbal Flyover, Outer Ring Rd, Lashkar-Hosur, Sarjapur Main Rd, HAL Old Airport Rd, Main Rd, Jayanagar Shopping Complex, Anekal, Koramangala, ITPL Main Rd,
Mumbai	Andheri, Andheri East, Khar (West), GK Gokhale Bridge, Andheri Station Rd, Andheri West, Bandra West, Santa Cruz, Thane, Gandhi Bazar	Sion, Prabhadevi, Bandra West, Sion - Panvel Expy, Andheri, Belapur Flyover, Goregaon East, Thane, Veera Desai Road, Jawahar Nagar,	Andheri West, Andheri East, Kennedy Bridge, GK Gokhale Bridge, Prabhadevi, Sion, Kurla East, Churchgate, Andheri Subway, Khar West,
Delhi	Jor Bagh Rd, Safdarjung Flyover, Greater Kailash, Pandav Nagar, West Vinod Nagar, Safdarjung Enclave, RK Puram Rd, Pandav Nagar, Wazirpur, INA market	Shastri Park, Sarojini Nagar, Near Tau devilal park, Manesar - Palwal Expy, Shakurpur, Anand Parbat, New Ashok Nagar, Dwarka, Sector 10 Dwarka, East Patel Nagar,	Mayur Vihar, Tilak Nagar, Dhaula Kuan I, Patel Chowk, Punjabi Bagh, Ansari Nagar East, Sarita Vihar, Dwarka, Lajpat Nagar, Karol Bagh,
Chennai	St. Thomas Mount, Pazhavanthangal, Avadi, Porur, Sunnambu Colony, Abdul Razzak St, Sankar Nagar, Tambaram, Old Mahabalipuram Rd, Teynampet,	Vadapalani, sivan koil street, Woods Road, Arcot Road, Kodambakkam. Sridevi Kuppam Road, Pillayar Koil Street, Velachery, Tambaram, Pallikaranai	St. Thomas Mount, Anna Salai, eldams road, Kaveri St East Tambaram, pallavaram, Porur, mudichur road, Ranganathan street, Medavakkam, Rajaji Rd Chennai
Kolkata	Howrah Bridge, BamangachiSalkia, EM Bypass, Diamond Harbour Road, Basanti Highway, Vivekananda flyover, Circular Garden Reach , sarat bose road, Strand Road, CIT Road	Taratala Flyover, Mejerhat flyovers , ruby Crossing, VIP bazar, Hyatt crossing, Metiabruz road, Nicco Park crossing Patuli, Metiabruz, Kalikapur	Howrah Bridge, Gariahat Road, Rabindra Sarani, Strand Road, Diamond Harbour, sarat bose road, Hazra Road, Chingrighata Flyover, Majherhat Bridge, Elgin Road
Hyderabad	Jeedimetla road, Hafeezpet flyover, uppal road Gandhi street, Jubilee Hills, Abid road, Amberpet T Junction, Gadribagh lane, Kalikabher, Suchitra bridge	Panjagutta, Kukatpally, kothapet village, 107-Madhapu, Banjara Hills, Jubilee hill, Gachibowli, hafeezpet flyover, Uppal, Ameenpur road	Begumpet Airport, Banjara Hills, Jubilee Hills, Panjagutta, Rasoolpura Junction, nizampet road, Hafeezpet flyover, Himayath Nagar, Malakpet Gunj, gachibowli route

- **Kolkata:** Howrah bridge, Mejerhat flyover, Diamond Harbour, Strand Road, & sarat bose road.

The identified location can be used to curtail the incident in the future.

VI. MODEL FEASIBILITY

We obtain top twenty-five hotspots for tier-1 cities in India w.r.t each category (Traffic, Potholes, and Accident). Due to lack of ground truth data availability, we crawl accident, potholes, and traffic-related reports from reputed news an article such as *(Times of India, Hindustan Times, etc.)* within the same time period i.e. (May-03-2018 to August-23-2018) and also collected reports from government agencies like *SP Traffic office and Municipal Corporation* of different cities. But in this study, we are not able to match the time of occurrence of any issues due to lack of ground truth data. So in this work, we match incident hotspot location names from the spots reported in different news articles and reports published from government agencies. We have extracted the location names from a news article in the same manner as we have extracted from the tweet posts as explained in section IV. After that calculate the feasibility ratio of our proposed model. The precision of the model is feasible only when it's feasibility ratio (f) between identified location from articles (α set) and the intersection of them with specified class of incident (β set) is closer to 100. We define the feasibility ratio f(w) as:

$$f(w) = \frac{|\alpha(w) \cap \beta(w)|}{|\alpha(w)|} \times 100 \quad (1)$$

Tweets related to the different incident was divided into different sets. The ratio constant obtained from these sets has significant feasibility i.e. (90% location overlap). We have found that most of the accident reports does not report in the news article, so the feasibility ratio in accident sets comes out be 65%. Our proposed method can identify a few new hotspots as shown in table 8 and also there is a number of posts which complaints about the road condition, street light not working, etc. These type of post can be useful for the government official to take proactive measure before any incident happens.

VII. CONCLUSION

In this paper, we introduced a framework that identifies incidents caused by non-recurrent events (accident, potholes, and traffic) from the social media platform. The proposed framework can be divided into five major components which include collecting data from multiple sources (i.e., hashtags, handle, and bounding box), data preprocessing, identification of similar semantic keywords corresponding to the different categories, removing the pragmatic ambiguity and content based location identification for finding the vulnerable areas. The major findings of this work are as follows:

- Introduce a robust method to classify the tweets into different categories by leveraging dense vector embedding to generate similar semantic keywords. The shortcoming of keyword-based approach (pragmatic ambiguity) was efficiently handled by using the word2vec model. Besides, the proposed method is capable to classify more than 40% tweets as compared to keyword-based approach.
- The location information from textual content was efficiently extracted by implementing a hybrid approach which is an amalgamation of NER, POS and Regular Expression (RE). The location information extracted from the RE might contain spelling errors. To subjugate this, employ edit distance is employed to match from the collected dataset which consists of street names, colony, and POI names for each tier-1 cities.
- The temporal and spatial analysis have been performed to determine user mobility patterns through their tweeting behaviour which can be used effortlessly by the government traffic agencies. Moreover, it was also observed that users often tweet when they are stuck in traffic congestion whereas some of them tweet during the working hours as well. From the experimental analysis, two peak tweet times were identified which were common to all the cities, i.e. in the morning between 8 A.M to 10:30 A.M and in the evening between 4:30 P.M to 6 P.M. The study reveals that tweet frequency is relatively higher during weekends than week days.
- Furthermore, we also identified the top 25 hotspots with respect to each category and listed some of the reasons (waterlogging, heavy rain, absence of sign boards, non functional street lights, traffic rules violation, littering on roads, road construction and ill-maintained roads, potholes and improperly parked vehicles) due to which people largely face traffic congestion and accidents.

The suggested approach can be put to use by the government agencies to take proactive action before any untoward incident takes place. It can also help in locating the most vulnerable places which should be taken care off on priority basis. However, the current work focuses only on English language tweets, so we plan to extend it for other languages also, in order to make the classification more robust and to improve the coverage of detected issues corresponding to different categories.

REFERENCES

[1] H. Abdelhaq, C. Sengstock, and M. Gertz, "Eventweet: Online localized event detection from twitter," *Proc. VLDB Endowment*, vol. 6, no. 12, pp. 1326–1329, Aug. 2013.

[2] F. Abel, C. Hauff, G.-J. Houben, R. Stronkman, and K. Tao, "Twitcident: Fighting fire with information from social web streams," in *Proc. 21st Int. Conf. World Wide Web*, Apr. 2012, pp. 305–308.

[3] M. Krstajic, C. Rohrdantz, M. Hund, and A. Weiler, "Getting there first: Real-time detection of real-world incidents on Twitter," in *Proc. 2nd Workshop Interact. Visual Text Anal., Task-Driven Anal. Social Media*, Washington, DC, USA, 2012.

[4] J. Weng and B.-S. Lee, "Event detection in Twitter," in *Proc. ICWSM*, vol. 11, 2011, pp. 401–408.

[5] A. Schulz, P. Ristoski, and H. Paulheim, "I see a car crash: Real-time detection of small scale incidents in microblogs," in *Proc. Extended Semantic Web Conf.* Berlin, Germany: Springer, 2013, pp. 22–33.

[6] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 4, pp. 2269–2283, Aug. 2015.

[7] Y. Gu, Z. S. Qian, and F. Chen, "From Twitter to detector: Real-time traffic incident detection using social media data," *Transp. Res. C, Emerg. Technol.*, vol. 67, pp. 321–342, Jun. 2016.

[8] Z. Zhang, Q. He, J. Gao, and M. Ni, "A deep learning approach for detecting traffic accidents from social media data," *Transp. Res. C, Emerg. Technol.*, vol. 86, pp. 580–596, Jan. 2018.

[9] D. Wang, A. Al-Rubaie, S. S. Clarke, and J. Davies, "Real-time traffic event detection from social media," *ACM Trans. Internet Technol.*, vol. 18, no. 1, p. 9, Dec. 2017.

[10] Y. Chen, Y. Lv, X. Wang, and F.-Y. Wang, "A convolutional neural network for traffic information sensing from social media text," in *Proc. IEEE 20th Int. Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2017, pp. 1–6.

[11] J. Cui, R. Fu, C. Dong, and Z. Zhang, "Extraction of traffic information from social media interactions: Methods and experiments," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 1549–1554.

[12] C. Gutiérrez, P. Figuerias, P. Oliveira, R. Costa, and R. Jardim-Goncalves, "Twitter mining for traffic events detection," in *Proc. Sci. Inf. Conf. (SAI)*, Jul. 2015, pp. 371–378.

[13] P. Tejaswin, R. Kumar, and S. Gupta, "Tweeting traffic: Analyzing Twitter for generating real-time city traffic insights and predictions," in *Proc. 2nd IKDD Conf. Data Sciences*, Mar. 2015, p. 9.

[14] A. Kurkcu, E. F. Morgul, and K. Ozbay, "Extended implementation method for virtual sensors: Web-based real-time transportation data collection and analysis for incident management," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2528, no. 1, pp. 27–37, Sep. 2015.

[15] S. Zhang, J. Tang, H. Wang, and Y. Wang, "Enhancing traffic incident detection by using spatial point pattern analysis on social media," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2528, pp. 69–77, Sep. 2015.

[16] J. He, W. Shen, P. Divakaruni, L. Wynter, and R. Lawrence, "Improving traffic prediction with tweet semantics," in *Proc. IJCAI*, Jun. 2013, pp. 1387–1393.

[17] M. Ni, Q. He, and J. Gao, "Using social media to predict traffic flow under special event conditions," in *Proc. 93rd Annu. Meeting Transp. Res. Board*, Jan. 2014, pp. 1–23.

[18] S. Grosenick, "Real-time traffic prediction improvement through semantic mining of social networks," Ph.D. dissertation, Univ. Washington, Seattle, WA, USA, 2012.

[19] J. Lafferty, A. McCallum, and F. C. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. 18th Int. Conf. Mach. Learn. (ICML)*, MA, USA, 2001, pp. 282–289.

[20] L. Ratinov and D. Roth, "Design challenges and misconceptions in named entity recognition," in *Proc. 13th Conf. Comput. Natural Lang. Learn.*, Jun. 2009, pp. 147–155.

[21] A. Ritter, S. Clark, O. Etzioni, "Named entity recognition in tweets: An experimental study," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, Jul. 2011, pp. 1524–1534.

[22] P. F. Brown, P. V. deSouza, R. L. Mercer, V. J. D. Pietra, and J. C. Lai, "Class-based n-Gram models of natural language," *J. Comput. Linguistics*, vol. 18, no. 4, pp. 467–479, 1992.

[23] X. Liu, S. Zhang, F. Wei, and M. Zhou, "Recognizing named entities in tweets," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol.*, Jun. 2011, pp. 359–367.

[24] X. Liu, F. Wei, S. Zhang, and M. Zhou, "Named entity recognition for tweets," *ACM Trans. Intell. Syst. Technol.*, vol. 4, no. 1, p. 3, Jan. 2013.

[25] J. Lingad, S. Karimi, and J. Yin, "Location extraction from disaster-related microblogs," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 1017–1020.

[26] S. Malmasi and M. Dras, "Location mention detection in tweets and microblogs," in *Proc. Conf. Pacific Assoc. Comput. Linguistics*. Singapore: Springer, 2015, pp. 123–134.

[27] J. Gelernter and S. Balaji, "An algorithm for local geoparsing of microtext," *GeoInformatica*, vol. 17, no. 4, pp. 635–667, Oct. 2013.

[28] C. Li and A. Sun, "Fine-grained location extraction from tweets with temporal awareness," in *Proc. 37th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2014, pp. 43–52.

[29] C. Li and A. Sun, "xtracting fine-grained location with temporal awareness in tweets: A two-stage approach," *J. Assoc. Inf. Sci. Technol.*, vol. 68, no. 7, pp. 1652–1670, Jul. 2017.

[30] W. Zhang and J. Gelernter, "Geocoding location expressions in twitter messages: A preference learning method," *J. Spatial Inf. Sci.*, vol. 2014, no. 9, pp. 37–70, Dec. 2014.

[31] J. Gelernter and W. Zhang, "Cross-lingual Geo-parsing for non-structured data," in *Proc. 7th Workshop Geographic Inf. Retr.*, Nov. 2013, pp. 64–71.

[32] T. Mikolov, K. Chen, G. Corrado, and J. Dean. (2013). "Efficient estimation of word representations in vector space." [Online]. Available: https://arxiv.org/abs/1301.3781

[33] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2013, pp. 3111–3119.

[34] A. Agarwal, B. Gupta, G. Bhatt, and A. Mittal, "Construction of a semi-automated model for FAQ retrieval via short message service," in *Proc. 7th Forum Inf. Retr. Eval.*, Dec. 2015, pp. 35–38.

[35] L. V. Subramaniam, S. Roy, T. A. Faruquie, and S. Negi, "A survey of types of text noise and techniques to handle noisy text," in *Proc. 3rd Workshop Anal. Noisy Unstructured Text Data*, Jul. 2009, pp. 115–122.

[36] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, Mar. 2003.

[37] X. Yan, J. Guo, Y. Lan, and X. Cheng, "A biterm topic model for short texts," in *Proc. 22nd Int. Conf. World Wide Web*, May 2013, pp. 1445–1456.

[38] H. Li, C. A. Calder, and N. Cressie, "Beyond Moran's I: Testing for spatial dependence based on the spatial autoregressive model," *Geograph. Anal.*, vol. 39, no. 4, pp. 357–375, Oct. 2007.

[39] A. Karlström and V. Ceccato, "A new information theoretical measure of global and local spatial association," in *Proc. Western Regional Sci. Assoc. Meeting Palm Springs*, Feb. 2001, pp. 1–31.

[40] D. Wang, W. Ding, H. Lo, T. Stepinski, J. Salazar, and M. Morabito, "Crime hotspot mapping using the crime related factors—A spatial data mining approach," *Appl. Intell.*, vol. 39, no. 4, pp. 772–781, Dec. 2013.

AMIT AGARWAL received the bachelor's degree from Uttrakhand Technical University, in 2012, and the master's degree from Graphic Era University, Dehradun, in 2015. He is currently pursuing the Ph.D. degree with IIT Roorkee, India. He is a recipient of the Visvesvaraya Ph.D. Scholarship for Doctoral Research. He is also a recipient of the Google scholarship to attend the LxMLS-2017, the SIGIR Scholarship to present his work in ESSIR-17, and the VLDB Scholarship to attend the VLDB conference, in 2016.

DURGA TOSHNIWAL received the bachelor's degree in engineering and the M.Tech. degree from the National Institute of Technology, Kurukshetra, and the D.Phil. degree from IIT Roorkee, India, where she is currently an Associate Professor with the Department of Computer Science and Engineering. She has published her research work in several international journals and conferences. She has attended, chaired sessions, and presented her work in several reputed international conferences in USA, U.K., Australia, and Europe. She has received various awards and honors, including the IBM Faculty Award, in 2012 and 2008, respectively, an Award from the UNESCO Chair in Data Privacy 2010, and the very prestigious IBM Shared University Research Award 2009 for her research projects. Her research work has also been featured in DataQuest, the leading IT Magazine in India in the DataQuest issue of Feb 15, 2010, in the article titled *Towards a Greener Planet*.