ISSN 2395-1621

EXTRACT USER TRAVEL HABITS, ROAD CONDITIONS AND ROAD TRAFFIC USING TWITTER

Ms. S. N. Darade, Ms. P. S. Kamble, Ms. A. J. kulthe, Mr. S. C. Dahake, Prof. P.S. Hase

shivamdahake39@gmail.com

Department of Information Technology Amrutvahini College of Engineering, Sangamner, Maharashtra
Department of Information Technology Amrutvahini College of Engineering, Sangamner, Maharashtra

ABSTRACT

Twitter is an online social networking service with more than 300 million users, generating a huge amount of information every day. Twitter's most important characteristic is its ability for users to tweet about events, situations, feelings, opinions, or even something totally new, in real time. 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.

KEYWORDS: Named Entity Recognition (NER), Part of speech (POS), Regular Expression (RE), Natural Language Processing (NLP).

ARTICLE INFO

Article History

Received: 28th September 2020

Received in revised form :

28th September 2020

Accepted:2nd October 2020

Published online :

2nd October **2020**

I. INTRODUCTION

In social media, posts analysis has always been considered as the most challenging task for twitter analyst/data scientist. 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 and 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 project, 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 monthly active users over the globe, where users interact with each other through a textual/visual post that is 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. However, these algorithms work well over the highways, but not on local 10 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. 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.

II. OBJECTIVE

- Tweets before and after executing the pre-processing steps i.e. hash tag & handle removal, URL removal, typo correction, abbreviation, and redundant consecutive character removal (RCCR).
- Designing a system that can extract the user travel habits using semantically extended keywords generations technique.
- Designing a system that can identify the road condition using semantically extended keywords generations technique.
- Designing a system that can identify the road traffic using semantically extended keywords generations technique.
- Designing a system that can identify the road accident using semantically extended keywords generations technique.

III. RELATED WORK

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 VOLUME 7, 2019 66537 A. Agarwal, D. Toshniwal: Face off: Travel Habits, Road Conditions, and Traffic City Characteristics Bared Using Twitter 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. 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.
Gu et al. [6] 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. [7] 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. [8] 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.

IV. PROBLEM STATEMENT

Propose system implementing a real time usage of Twitter for extract User Travel Habits, Road Conditions and Road Traffic.

STATEMENT OF SCOPE
This project scope is 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.

V. SYSTEM REQUIREMENTS

HARDWARE RESOURCES REQUIRED

Sr. No.	Parameter	Minimum Requirement
1	Processor	Intel Core I3
2	RAM	3 GB
3	HD	100 GB (min)

Table 1: Hardware Requirements

Database Requirements

MySQL Database MySQL is on open source database which is mainly a RDBMS i.e. relational database management system. As a database server, primary function of this software is to storing and retrieving data as requested by other from end software applications like java which may Or may not run either on the same computer or on different computer. This can be across the network either in internet or intranet.

Sr. No.	Parameter	Minimum Requirement
1	Operating System	Microsoft Windows 7
2	Programming Language	Java
3	Database	MySQL

Table 2: Software Requirements

VI. SYSTEM ARCHITECTURE

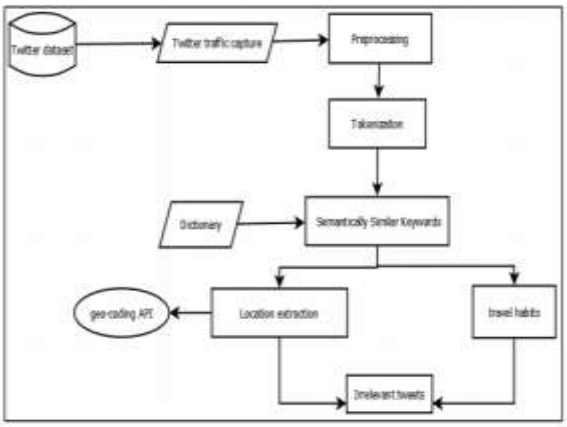

Figure 1: Architecture Diagram

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.

ALGORITHM

Semantically Extended Keywords Generation

• Input: Clean tweets $\zeta = \delta1,\delta2,\delta3,....,\delta n$

Step 1: res1 = [' ']
Step 2: res3 = [' ']
Step 3: res2 = [' ']
Step 4: for each tweet in ζ do
Step 5: T = nltk.tokenize(tweet)
Step 6: end for
Step 7: model = models.gensim.Word2Vec(T)
Step 8: SD KD = s1, s2, ..., sn
Step 9: res3 = SD KD
Step 10: for each keyword in SD KD do
Step 11: res1 = model.most similar(Keyword, topn = 5)
Step 12: res2.append (res1)
Step 13: res2 = unique (res2)
Step 14: if (res2 - SD KD) = φ then
Step 15: Stop 27
Step 16: Exit
Step 17: else
Step 18: res3 = res2-SD KD
Step 19: SD KD = SD KD U res3
Step 20: goto step 10
Step 21: end if
Step 22: end for

• Output: Expended keyword list W2V KD = E(SD KD)

VII.RESULTS

Figure 2: Select Road or Hotel

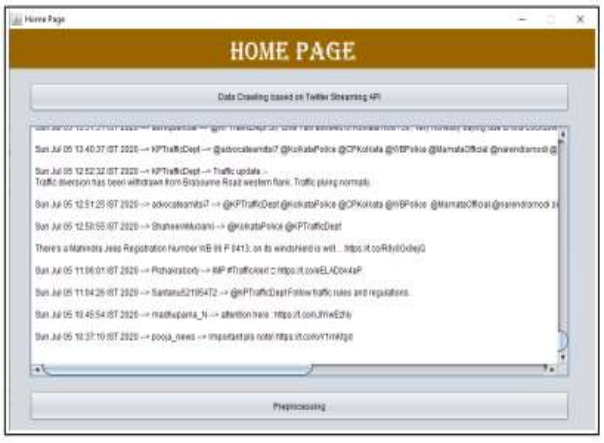

Figure 3: Home Page

Figure 4: Processing

Figure 5: Semantic Similar Automatic Keywords Generation

Figure 6: Mentioned based Location Identification

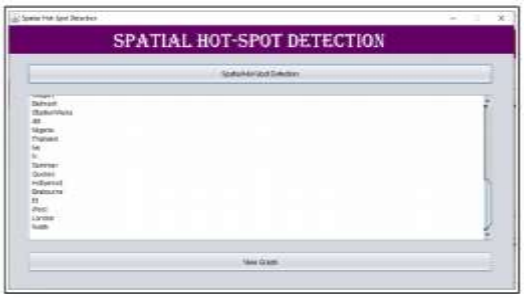

Figure 7: Spatial Hot-spot Detection

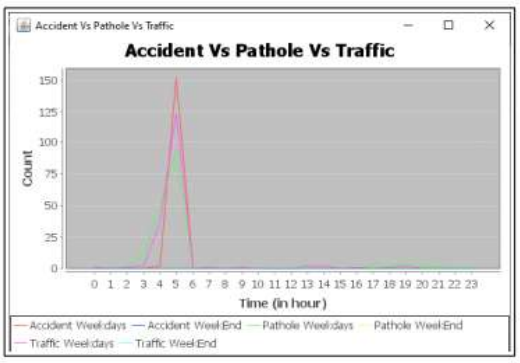

Figure 8: Accident Vs Pothole Vs Traffic

VIII. CONCLUSION

In this paper, we introduced a framework that identifies incidents caused by nonrecurrent 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 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 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.

• Furthermore, we also identified the top 25 hotspots with respect to each category and listed some of the reasons (water logging, 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.

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.

[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.

