International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

Extract User Travel Habits, Road Conditions and Road Traffic, Identify Accident Location using Twitter

Ashwini A. Gaikwad[1], Prof. Pravin Nimbalkar[2]

Department of Computer Engineering, JSPM's Imperial College of Engineering and Research, Wagholi, Pune, India[1]

Assistant Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering and Research, Pune, India[2]

ABSTRACT: Twitter is an online social networking service. It have communication administration with in excess of 300 million clients, creating a colossal measure of data consistently. Twitter's most significant trademark is its capacity for clients to tweet about occasions, circumstances, emotions, conclusions, or in any event, something absolutely new, progressively. The internet based life tweet text have been mined in order to recognize the grievances in regards to different street transportation issues of traffic, mishap, and potholes. So as to recognize and isolate tweets identified with various issues, keyword based methodologies have been utilized already, yet these strategies are exclusively subject to seed catchphrases which are physically given and these arrangement of seed keywords are not adequate to cover all tweets posts. So, to overcome this issue, a methodology has been recommended that catches the semantic setting through dense word inserting by utilizing word2vec model. Nonetheless, the procedure of tweet segregation based on semantic similar keywords may experience suffer from the problem of pragmatic ambiguity. To deal with this Word2Vec model has been applied to match the semantically similar tweets with respect to each category. Furthermore, the hotspots have been recognized comparing to every class. In any case, because of the shortage of geo-labeled tweets, we have proposed a mixture technique which amalgamates Named Entity Recognition (NER), Part of Speech (POS), and Regular Expression (RE) to extricate the area data 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: Twitter, Social Media, Tweet, Travel Habits, Road Condition, Road Traffic

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 cities(Mumbai, Delhi, Kolkata, and Bengaluru) every year losses 22 billion dollar due to congestion. It for the most part incited from non-repetitive occasions, for example, mishap, unfavorable street conditions, development on streets, potholes, unfriendly climate condition, and insufficient seepage. Because of this person has to spend more than one a-half hour longer during the peak hour to cover the same distance as on non-peak hour

II.PROBLEM STATEMENT

The internet based life tweet text have been mined in order to recognize the grievances in regards to different street transportation issues of traffic, mishap, and potholes. So as to recognize and isolate tweets identified with various issues, keyword based methodologies have been utilized already, yet these strategies are exclusively subject to seed catchphrases which are physically given and these arrangement of seed keywords are not satisfactory to cover all tweets posts.

III.LITERATURE SURVEY

Z. Zhang, Q. He, J. Gao, and M. Ni "A deep learning approach for detecting traffic accidents from social media data," 2018

In this paper they utilize profound learning model, for example Deep Belief Network (DBN) and Long Short-Term Memory (LSTM) to distinguish auto collision from web based life. The creator isolated their work into three

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

stages, for example right off the batinclude determination, also arrangement, and thirdly approval. From that point forward, they have played out the arrangement calculation over individual and matched tokens. Their test results show that over combined token DBN accomplish higher exactness (i.e., 85%) at that point LSTM. [1]

D. Wang, A. Al-Rubaie, S. S. Clarke, and J. Davies, "Real-time traffic event detection from social media", 2017

They proposed a tweet-LDA to identify traffic related tweets from internet based life which is the steady methodology of the watchword based methodology and furthermore the creator handle the sober minded equivocalness. Their exploratory outcomes show that their model accomplishes preferred exactness over customary techniques like SVM [2]

A. Schulz, P. Ristoski, and H. Paulheim , "I see a car crash: Real-time detection of small scale incidents in microblogs,", 2013.

This paper contributes a methodology that use data gave in microblogs to location of small scale incidents. They indicated how AI and semantic web technologies can be consolidated to recognize incident related microblogs. With 89% recognition exactness, and beat cutting edge draws near. Moreover, their methodology can absolutely limit microblogs in existence, consequently, empowering the continuous identification of incidents. [3]

E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis,", 2015..

They have exploited available software packages and state-of the-art techniques for text analysis and pattern classification. These technologies and techniques have been analyzed, tuned, adapted and integrated in order to build the overall system for traffic event detection. Among the analyzed classifiers, They have shown the superiority of the SVMs, which have achieved accuracy of 95.75%, for the 2-class problem, and of 88.89% for the 3-class problem, in which they have also considered the traffic due to external event class. [4]

IV.SYSTEM DESIGN

We have proposed a hybrid method which amalgamates Named Entity Recognition, Part of speech, and Regular Expression 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). The evaluation results signify that the stated approach identifies few additional hotspots as compared to the existing reports while analyzing the tweets.

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.

To segregate the tweets, we proposed semantically similar adaptive keyword generative method by leveraging the semantic context through dense word embedding using Word2vec model.

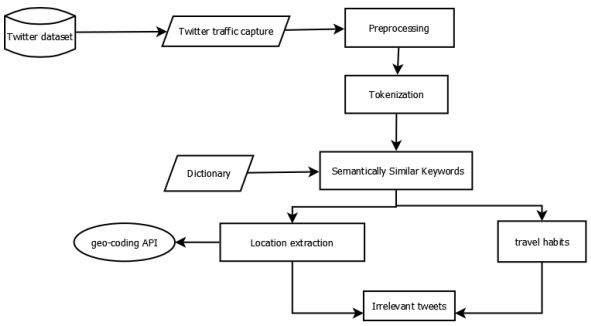

Fig 1: System Architecture

The proposed approach overcomes the shortcoming of traditional methods i.e. keyword based segregation, and

classification by using a machine learning algorithm. This project 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.

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.

V.ALGORITHM

Algorithm: Semantically Extended Keywords Generation

Input: Clean tweets Ƈ = {δ1, δ2, δ3, . . , δn}

Output: Expended keyword list W2V_KD = E(SD_KD)

Step 1: res1=[' ']

Step 2: res3=[' ']

Step 3: res2=[' ']

Step 4: for each tweet in Ƈ 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

VI.RESULT

In proposed system Word2Vec and WordNet dictionary are used. Word2Vec model has been applied to match the semantically similar tweets with respect to each category. WordNet requires more time compared to the Word2Vec.

So Word2Vec is better than WordNet as it requires less time than WordNet. Graph below shows performance comparision of both model (Word2Vec and WordNet).

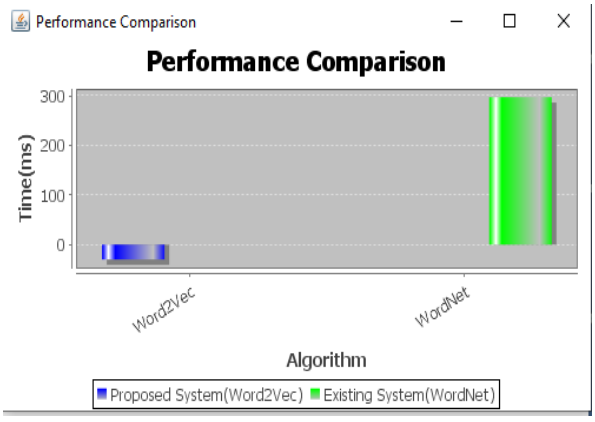

Figure 3: Performance Comparison

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.488 |

|| Volume 8, Issue 7, July 2020 ||

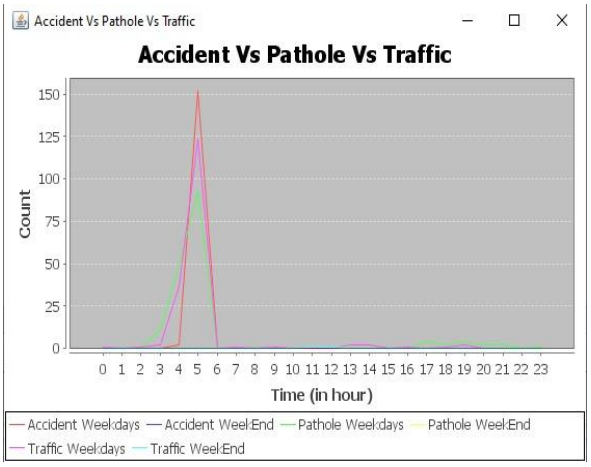

Figure 2: Accident vs Pothole vs Traffic

VII.CONCLUSION

We propose a approach for early detection and identification of user travel habits, road conditions, and road traffic and road accident.

- Semantic Similar keywords
- Handling Pragmatic Ambiguity
- Data enrichment
- Mention Based Location Extraction
- Hotspot & Critical location Identification
- Temporal Analysis over Weekends (WKND) and Weekday (WKD)

REFERENCES

[1] Z. Zhang, Q. He, J. Gao, and M. Ni "A deep learning approach for detecting traffic accidents from social media data,'' IEEE 2018.

[2] D. Wang, A. Al-Rubaie, S. S. Clarke, and J. Davies, "Real-time traffic event detection from social media", IEEE, 2017.

[3] A. Schulz, P. Ristoski, and H. Paulheim , "I see a car crash: Real-time detection of small scale incidents in microblogs,", 2013.

[4] E. D'Andrea, P. Ducange, B. Lazzerini, and F. Marcelloni, "Real-time detection of traffic from twitter stream analysis,", IEEE, 2015.

[5]. Freddy Tapia, Cristina Aguinaga y Roger Luje, "Detection of Behavior Patterns through Social Networks like Twitter, using Data Mining techniques as a method to detect Cyberbullying", IEEE, 2018.

[6] C. Gutiérrez, P. Figuerias, P. Oliveira, R. Costa, and R. Jardim-Goncalves, ''Twitter mining for traffic events detection'' in Proc.Sci.Inf.Conf.(SAI), IEEE 2015