OPEN ACCESS INTERNATIONAL JOURNAL OF SCIENCE & ENGINEERING

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

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

Department of Computer Engineering, JSPM's Imperial College of Engineering, Wagholi, Pune-51
Assistant Professor, Department of Computer Engineering, JSPM's Imperial College of Engineering, Wagholi, Pune-51
ashupict@gmail.com , ppnimbalkar1@gmail.com

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: *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 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.

II PROBLEM STATEMENT

This implementation is aimed at a real time usage of Twitter for extract User Travel Habits, Road Conditions and Road Traffic.

III LITERATURE SURVEY

Yong-Ju Lee, Myungcheol Lee, Mi-Young Lee, Sung Jin Hur, Okgee Min, "Design of a Scalable Data Stream Channel for Big Data Processing", 2015.

This paper outlines big data infrastructure for processing data streams. Our project is distributed stream computing platform that provides cost-effective and large-scale big data services by developing data stream management system. This research contributes to advancing feasibility of big data processing for distributed, real-time computation even when they are overloaded. [1]

Babak Yadranjiaghdam, Seyedfaraz Yasrobi, Nasseh Tabrizi, "Developing a Real-time Data Analytics Framework for Twitter Streaming Data", 2017.

The proposed framework includes data ingestion, stream processing, and data visualization components with the Apache Kafka messaging system that is used to perform data ingestion task. Furthermore, Spark makes it possible to perform sophisticated data processing and machine learning algorithms in real time. [2]

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", 2018.

This research focuses on the detection and analysis of cyberbullying on pages and with pejorative terms in Spanish, taking advantage of the power of classification of feelings through specialized tools. For the detection of cyberbullying, first the efficiency of classification of each tool is measured, through a set of pejorative terms commonly used to hurt other people. [3]

Amir Hossein Akhavan Rahnama, "Distributed Real-Time Sentiment Analysis for Big Data Social Streams", 2014.

The real challenge with real-time stream data processing is that it is impossible to store instances of data, and therefore online analytical algorithms are utilized. To perform real-time analytics, pre-processing of data should be performed in a way that only a short summary of stream is stored in main memory. [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.

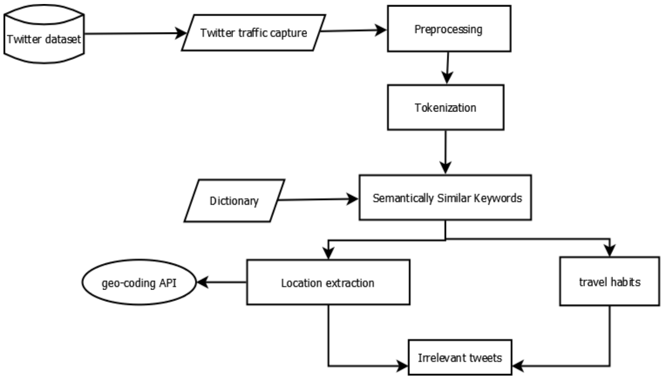

Figure 1: System Architecture

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.

V CONCLUSION

We propose a novel 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] Yong-Ju Lee, Myungcheol Lee, Mi-Young Lee, Sung Jin Hur, Okgee Min, “Design of a Scalable Data Stream Channel for Big Data Processing”, IEEE, 2015.

[2] Babak Yadranjiaghdam, Seyedfaraz Yasrobi, Nasseh Tabrizi, “Developing a Real-time Data Analytics Framework For Twitter Streaming Data”, IEEE, 2017.

[3] 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.

[4] Amir Hossein Akhavan Rahnama, “Distributed Real-Time Sentiment Analysis for Big Data Social Streams”, IEEE, 2014.