International Journal of Engineering and Technical Research (IJETR)
ISSN: 2321-0869 (O) 2454-4698 (P), Volume-6, Issue-3, November 2016

Analysis and Implementation of Twitter Stream by Real-Time Traffic Detection

Prof. KL Chugh, B Madhuravani, Y Madan Reddy, P Naveen Kumar

***Abstract*— In late time we have seen than online networking is functions as method of data. We get loads of data furthermore critical news on online networking. Interpersonal organizations have been as of late utilized as a wellspring of data for occasion discovery, with specific notice to street movement clog and auto crashes. In this paper, we show an ongoing checking framework for activity occasion location from Twitter stream examination. The framework gets tweets from Twitter as indicated by a few inquiry criteria; forms tweets, by applying content mining systems; lastly plays out the arrangement of tweets. The point is to relegate the fitting class mark to every tweet, as identified with a movement occasion or not. The movement discovery framework was utilized for continuous checking of a few regions of the Italian street system, taking into consideration identification of activity occasions practically progressively, regularly before online movement news sites. We utilized the bolster vector machine as a grouping model, and we accomplished an exactness esteem by taking care of a parallel arrangement issue (movement versus no activity tweets). We were additionally ready to segregate if activity is brought about by an outside occasion or not, by tackling a multiclass characterization issue and getting precision esteem which demonstrates exactness. We can say that twitter is new proficient media of news and data alongside activity news.**

***Index Terms*— Traffic event detection, Tweet classification, Text mining, Social sensing.**

I. INTRODUCTION

In this time everyone needs online networking account, additionally called small scale blogging administrations (e.g., Hike, Twitter, Facebook, Google+, Instagram), have spread as of late, turning into another sort of constant data system. Their acknowledgment originates from the attributes of transportability. Individuals strongly utilize interpersonal organizations to appear (individual or open) genuine occasions happening around them or just to express their sentiment on a given point, through an open message, remarks. Informal communities permit individuals to make a character and let them offer it with a specific end goal to fabricate a group. The subsequent informal organization is then a premise for saving social connections, discovering clients with comparative interests, and finding substance and information entered by different clients [3].

The client message partook in interpersonal organizations is called Status Update Message (SUM), and it might contain, aside from the content, meta-data, for example, timestamp, geographic directions (scope and longitude), and name of the client, connections to different assets, hash labels, and says. A few SUMs alluding to a specific theme or identified with a constrained geographic region may ace vide, if effectively broke down, incredible arrangement of important data around an occasion or a subject. Indeed, we may see interpersonal organization clients as social sensors [4], [5], and SUMs as sensor data [6], as it happens with conventional sensors.

Presently radio is likewise inclines toward for news and activity subtle elements, a few stations like radio Mirchi which additionally gives constant movement news. Interpersonal organizations and media stages have been broadly utilized as a wellspring of data for the location of occasions, for example, movement blockage, episodes, common catastrophes like seismic tremors, or different occasions. An occasion can be characterized as a true event that happens in a particular time and space [1], [7]. Specifically, with respect to activity related occasions, individuals frequently share by method for SUM data about the present movement circumstance around them while driving. Thus, occasion discovery from informal communities is likewise regularly utilized with Intelligent Transportation Systems (ITSs). An ITS is a base which, by coordinating ICTs (Information and Communication Technologies) with transport systems, vehicles and clients, permits enhancing wellbeing and administration of transport systems. ITSs give, e.g., continuous data about climate, activity clog or control, or plan effective courses.

Be that as it may, occasion recognition from interpersonal organizations investigation is a more difficult issue than occasion discovery from customary media like web journals, messages, and so forth., where writings are all around arranged [2]. Truth be told, SUMs are unstructured and sporadic writings, they contain casual or contracted words, incorrect spellings or linguistic mistakes [1]. Because of their inclination, they are normally extremely short, in this way turning into an inadequate wellspring of data [2]. Moreover, SUMs contain a colossal measure of not valuable or futile data, which must be separated. As indicated by Pear Analytics, it has been assessed that more than 40% of all Twitter SUMs is pointless with no valuable data for the gathering of people, as they allude to the individual circle. For these reasons, with a specific end goal to investigate the data originating from interpersonal organizations, we misuse content mining systems, which utilize techniques from the fields of information mining, machine learning, measurements, and Natural Language..

II. RELATED WORK

With reference to current methodologies for utilizing online networking to separate helpful data for occasion discovery, we have to recognize little scale occasions and substantial

Prof. KL Chugh, B Madhuravani, Department of CSE, MLR Institute of Technology, Hyderabad, India
Y Madan Reddy, Department of CSE, MLR Institute of Technology, Hyderabad, India
P Naveen Kumar, Department of CSE, MLR Institute of Technology, Hyderabad, India

scale occasions. Little scale occasions for the most part have a little number of SUMs identified with them, have a place with an exact geographic area, and are moved in a little time interim. Then again, expansive scale occasions are portrayed by an enormous number of SUMs, and by a more extensive fleeting and geographic scope. Subsequently, because of the littler number of SUMs identified with little scale occasions, little scale occasion discovery is a non-paltry errand. A few works in the writing manage occasion location from informal organizations. Numerous works manage expansive scale occasion discovery [6] and just a couple works concentrate on little scale occasions.

As to scale occasion identification, Sakaki et al. [6] utilize Twitter's tweet to identify tremors and hurricanes, by observing uncommon trigger-catchphrases, and by applying a SVM as a double classifier of positive occasions and negative occasions. In the creators exhibit a strategy for distinguishing genuine occasions, for example, normal fiascos, by examining Twitter's tweet and by utilizing both Natural Language Processing and term-recurrence based systems. Bites investigate the substance of tweets shared amid the H1N1 (i.e., swine influenza) flare-up, containing catchphrases and hash-labels identified with the H1N1 occasion to decide the sort of data traded by online networking clients dissect geo-labeled tweets to recognize backwoods fire occasions and layout the influenced territory.

As to scale occasion location, concentrate on the discovery of flames in a plant from Twitter stream investigation, by utilizing standard Natural Language Processing methods and a Naive Bayes (NB) classifier. In data removed from Twitter's tweet is converged with data from crisis systems to identify and break down little scale occurrences, for example, fires. Wanichayapong et al. [12] remove, utilizing Natural Language Processing strategies and syntactic investigation, activity data from miniaturized scale web journals to recognize and arrange tweets containing place says and movement data. Propose a framework, called TEDAS, to recover episode related tweets. The framework concentrates on Crime and Disaster-related Events (CDE, for example, shootings, storms, and auto crashes, and expects to characterize tweets as CDE occasions by abusing a sifting in view of catchphrases, spatial and worldly data, number of devotees of the client, number of retweets, hash labels, connections, and notice. Sakaki et al. [9] extricate, taking into account catchphrases, realtime driving data by breaking down Twitter's SUMs, and utilize a SVM classifier to channel "loud" tweets not identified with street activity occasions. Identify little scale auto episodes from Twitter stream examination, by utilizing semantic web innovations, alongside Natural Language Processing and machine learning procedures.

Creators of[4] utilized twitter streams to identify tremors occasion which is an expansive scale occasion by observing exceptional trigger keywords. For recognition of seismic tremor creators utilized SVM as a parallel classifier of positive occasions and negative occasions [6]. In [7] shakki. et.al. exhibited a strategy is utilized for distinguishing certifiable occasion like characteristic catastrophes by breaking down Tweets by utilizing both term-recurrence based strategies and NLP. A novel framework is introduced in [2] for identification and examination of occasions from rich data of the twitter stream. The Authors introduced the accompanying three functionalities (1) New occasion identification, (2) Event positioning as per significance, and (3) gathering worldly and spatial examples for occasions [2]. The work was at first centered around terrible and Disaster related Events(CDE), eg. Shooting ,accidents and so forth [2]. Activity, flames, accidents and neighborhood indications are little scale occasions. They have a little number of SUMs identified with that occasions. Little number of SUMs classifications them to the little scale occasions. Little scale occasions have a place from little time interim. Huge scale occasions like a quakes, tornados, decisions are portrayed by a substantial number of SUMs. Such occasion has more extensive worldly and geographic scope.

In reference [13]authors displayed a framework which examine movement and its causes. Twitter is an online networking which permits individuals to share and read tweets identified with all events[13]. The framework can read the tweets from twitter and this framework utilizes regular dialect handling procedure on them. At that point framework arranges the tweets identified with movement and advises the enlisted clients about it. The normal dialect preparing concentrates on creating proficient calculations to process content. The NLP likewise center to change over content into dialect [13].

In paper [14] Vikram Singh et. al. proposed a compelling tokenization strategy in view of preparing. The introduced technique results in the better tokens utilizing tokenization alongside preprocessing. On the off chance that less number of token produced then least storage room is required, This encourages more exactness in results recovery [14]. Proposed calculation takes duty regarding lessening the season of recovery data [14].

Maximilian Waltheret. al.in [15] distinguish Geo-spatial occasion utilizing twitter SUM's. The proposed approach creators assembled tweets for target occasions that can be characterized by a client through catchphrases [15]. Grouping and molecule sifting techniques are utilized for discovering this geo-spatial occasions [15]. Creators utilized regular topic as though individuals are tweeting from the same spot or territory utilizes comparable words which, implies that these clients are discussing that thing only[15].

III. ARCHITECTURE OF TRAFFIC DETECTION SYSTEM

The architecture of traffic detection system based on Twitter's tweet analysis is presented. The system architecture is service-oriented and eventdriven, and is composed of three main modules, namely: i) "Fetch of SUMs and Pre-processing", ii) "Elaboration of SUMs", iii) "Classification of SUMs". The purpose of the proposed system is to fetch SUMs from Twitter, to process SUMs by applying a few text mining steps, and to assign the appropriate class label to each SUM. Finally, as shown in Fig. 1, by analyzing the classified SUMs, the system is able to notify the presence of a traffic event.

The main tools we have exploited for developing the system are: 1) Twitter's API, which provides direct access to the public stream of tweets; 2) Twitter4J, a Java library that we used as a wrapper for Twitter's API; 3) the Java API provided by Waikato Environment for Knowledge Analysis, which we mainly employed for data pre-processing and text mining elaboration.

They perform the experiments using SVM, NB, and RIPPER classifiers. In this paper, we focus on a particular small-scale

event, i.e., road traffic, and we aim to detect and analyze traffic events by processing users' SUMs belonging to a certain area and written in the Italian language. To this aim, we propose a system able to fetch, elaborate, and classify SUMs as related to a road traffic event or not. To the best of our knowledge, few papers have been proposed for traffic detection using Twitter stream analysis. However, with respect to our work, all of them focus on languages different from Italian, employ different input features and/or feature selection algorithms, and consider only binary classifications. In addition, a few works employ machine learning algorithms [9], while the others rely on Natural Language Processing techniques only. The proposed system may approach both binary and multi-class classification problems. As regards binary classification, we consider traffic related tweets, and tweets not related with traffic. As regards multi-class classification, we split the traffic-related class into two classes, namely traffic congestion or crash, and traffic due to external event. In o ur paper, with external event we refer to a scheduled event, or to an unexpected event in this way we aim to support traffic and city administrations for managing Scheduled or unexpected events in city.

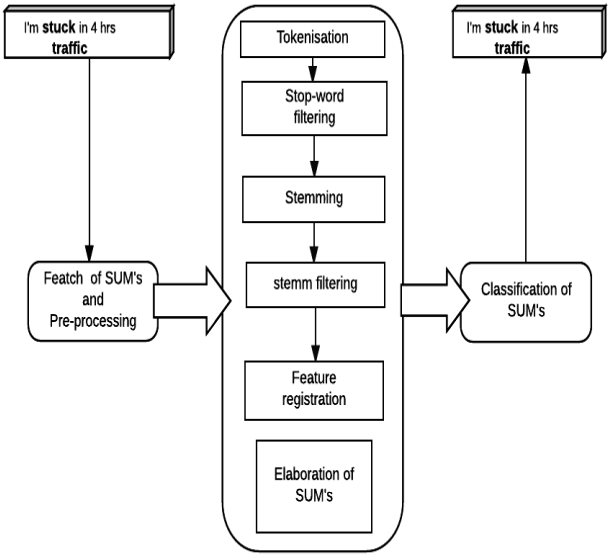

Fig 1. System Architecture

IV. REAL TIME DETECTION OF TRAFFIC EVENTS

Continuously identification of movement occasions we created framework was introduced and tried for the constant checking of a few zones of the Italian street system, by method for the examination of the Twitter stream originating from those territories. The point is to play out a constant observing of every now and again bustling streets and interstates keeping in mind the end goal to identify conceivable movement occasions continuously or even ahead of time as for the conventional news media. The framework is executed as an administration of a more extensive administration arranged stage to be produced with regards to the SMARTY venture. The administration can be called by every client of the stage, who yearnings to know the activity conditions in a specific zone. In this area, we plan to demonstrate the adequacy of our framework in deciding movement occasions in brief time. We simply introduce a few results for the 2-class issue.

V. PROPOSED SYSTEM

Proposed system architecture is service oriented, event-driven. The systems has following main functionality.

1) Extraction of SUMs and Preprocessing
2) Elaboration of SUMs
3) Classification of SUMs

At first client tweets are gotten from the twitter stream utilizing twitter API's. In the meantime the activity occasion related preparing must be given to the framework utilizing prepared classifier. Once the tweets are brought further preparing, for example, tokenization, stemming, separating are connected on the tweets. After these undertaking the component determination is performed. At that point utilizing chose highlights the tweets are grouped in activity related and non-movement related tweets.

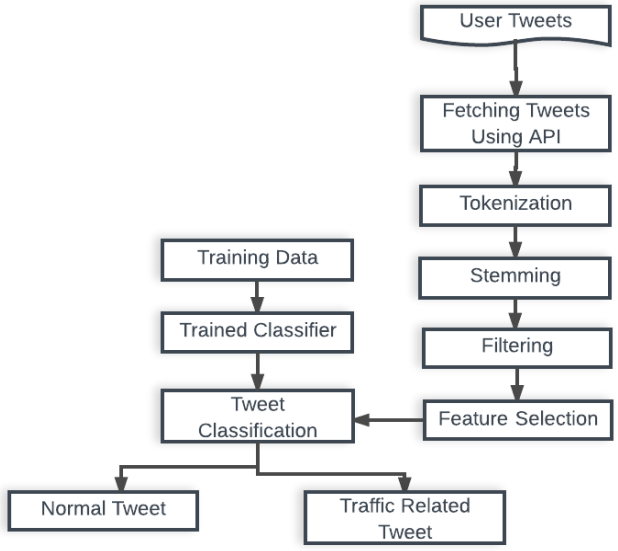

Fig 2. Classification of tweets.

At first client tweets are brought from the twitter stream utilizing twitter API's. In the meantime the activity occasion related preparing must be given to the framework utilizing prepared classifier. Once the tweets are brought further handling, for example, tokenization, stemming, sifting are connected on the tweets. After these undertaking the element determination is performed. At that point utilizing chose highlights the tweets are ordered in activity related and non-movement related tweets.

1) Fetch of SUMs and Pre-Processing

Entirety's from the client profile are brought and the further handling is connected on it to separate the crude data. Brought crude tweet contains data like: client id, timestamp, geographic direction, re-tweet banner and content of the tweet [1].

2) Elaboration of SUMs

In preprocessing module elaboration of SUMs is finished. This module changes the arrangement of strings into an arrangement of numeric vectors are explained by the Classification of Status Update Message module [1]. Diverse content mining strategies are utilized as a part of arrangement to the pre-prepared SUMs to accomplish this. The content mining steps done in this module are

- Tokenization is ordinarily the initial step of the content mining process which is utilized for changing a flood of words into a surge of preparing

units called tokens. In this progression diverse operations utilized like evacuation of other non-content characters and accentuation, and standardization of images [1].

- In stop-word separating expel pointless words which do not give any data to the content examination.
- Stemming is the way toward finding the root word from particular word. It evacuates its postfix.
- Stem sifting is accustomed to decreasing the quantity of stems of every SUM. Every SUM is separated by expelling from the arrangement of stems the ones not from the arrangement of significant stems [1].

3) Classification of SUMs

Grouping of SUMs module is utilized to doles out every SUM a class name identified with activity occasions. This module yields a gathering of N marked SUMs [1].

VI. PROPOSED ALGORITHM

Input: Training Dataset T, Test dataset D,
Output: Clustered Tweet set.
1) Initially train the classifier by using semi-supervised traffic related training dataset.
2) Fetch tweets of user from tweeter account.
3) Store it in DB
4) For each tweet in DB finding the similarity using Euclidean distance by trained data.
5) If (similarity > Threshold)
6) Add that tweet to traffic related tweet set
7) Else
8) Add to normal (not related to traffic) tweet set.
9) Return classified tweets

VII. RESULTS

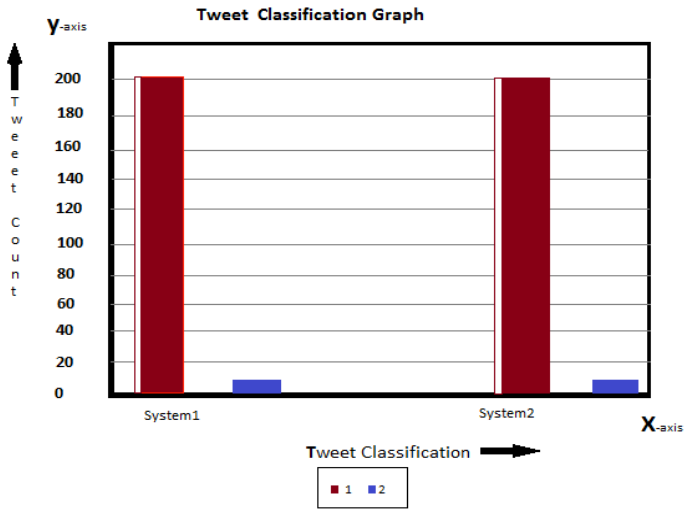

Fig 3.tweet Classification

The above diagram demonstrates the outcome and examination of existing framework SVM and proposed semi-directed strategy. Framework 1 is SVM and framework 2 is semi-administered strategy. Review, exactness qualities are more noteworthy for proposed semi-managed approach than existing framework.

VIII. CONCLUSION

In this paper, we have proposed a framework for realtime location of activity related occasions from Twitter stream examination. The framework, based on a SOA, can bring and characterize floods of tweets and to tell the clients of the nearness of movement occasions. Besides, the framework is additionally ready to separate if a movement occasion is because of an outside cause, for example, football match, parade and indication, or not. We have abused accessible programming bundles and best in class procedures for content examination and example characterization. These advances and strategies have been dissected, tuned, adjusted and incorporated keeping in mind the end goal to manufacture the general framework for activity occasion location. Among the broke down classifiers, we have demonstrated the prevalence of the SVMs, which have accomplished precision for the 2-class issue, and the 3-class issue, in which we have likewise considered the activity because of outer occasion class.

REFERENCES

[1] P. Ruchi and K. Kamalakar, *"ET: Events from tweets,"* in Proc. 22nd Int. Conf. World Wide Web Comput., Rio de Janeiro, Brazil, 2013, pp. 613–620.

[2] G. Anastasi et al., *"Urban and social sensing for sustainable mobility in smart cities,"* in Proc. IFIP/IEEE Int. Conf. Sustainable Internet ICT Sustainability, Palermo, Italy, 2013, pp. 1–4.

[3] A. Rosi et al., *"Social sensors and pervasive services: Approaches and perspectives,"* in Proc. IEEE Int. Conf. PERCOM Workshops, Seattle, WA, USA, 2011, pp. 525– 530.

[4] T. Sakaki, M. Okazaki, and Y. Matsuo, *"Tweet analysis for real-time event detection and earthquake reporting system development,"* IEEE Trans. Knowl. Data Eng., vol. 25, no. 4, pp. 919–931, Apr. 2013

[5] K. Perera and D. Dias, *"An intelligent driver guidance tool using location based services,"* in Proc. IEEE ICSDM, Fuzhou, China, 2011, pp. 246–251.

[6] T. Sakaki, Y. Matsuo, T. Yanagihara, N. P. Chandrasiri, and K. Nawa, "*Real-time event extraction for driving information from social sensors,"* in Proc. IEEE Int. Conf. CYBER, Bangkok, Thailand, 2012, pp. 221–226.

[7] B. Chen and H. H. Cheng, *"A review of the applications of agent technology in traffic and transportation systems,"* IEEE Trans. Intell. Transp. Syst., vol. 11, no. 2, pp. 485–497, Jun. 2010.

[8] S. Weiss, N. Indurkhya, T. Zhang, and F. Damerau, *Text Mining: Predictive Methods for Analyzing Unstructured Information*. Berlin, Germany: Springer-Verlag, 2004.

[9] V. Gupta, S. Gurpreet, and S. Lehal, *"A survey of text mining techniques and applications,"* J. Emerging Technol. Web Intell., vol. 1, no. 1, pp. 60–76, Aug. 2009.

[10] M. W. Berry and M. Castellanos, *Survey of Text Mining*. New York, NY, USA: Springer-Verlag, 2004.

[11] H. Takemura and K. Tajima, *"Tweet classification based on their lifetime duration,"* in Proc. 21st ACM Int. CIKM, Shanghai, China, 2012, pp. 2367–2370.

[12] J. Yin, A. Lampert, M. Cameron, B. Robinson, and R. Power, *"Using social media to enhance emergency situation awareness,"* IEEE Intel. Syst., vol. 27, no. 6, pp. 52–59, Nov./Dec. 2012.

[13] Harshita Rajwani, Srushti Somvanshi, AnujaUpadhye*," Dynamic Traffic Analyzer Using Twitter"*, International Journal of Science and Research (IJSR) 2014.

[14] Vikram Singh and Balwinder *Saini "An Effective Tokenization Algorithm for Information Retrieval System"* CS and IT-CSCP 2014

[15] Maximilian Walther and Michael Kaisser,*"Geo-spatial Event Detection in the Twitter Stream"*, P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 356367, 2013 .springer Verlag Berlin Heidelberg 2013.

[16] M. Krstajic, C. Rohrdantz, M. Hund, and A. Weiler, *"Getting there first: Real-time detection of real-world incidents on Twitter"* in Proc. 2nd IEEE Work Interactive Vis. Text Anal.-Task-Driven Anal. Soc. Media IEEE VisWeek," Seattle, WA, USA, 2012.