Traffic issues categorization of Indian Cities using Word2Vec by Social Media Data

Mitesh Trivedi [1], Shilpa Serasiya[2]
[1]Student[2], Assistant Professor
[1,2]Department of Computer Engineering,
[1,2]Kalol Institute of Technology and Research Center, Kalol, Gandhinagar, Gujarat, India.

***Abstract*– In traditional ways, traffic related and road and transport related issued are highlighted and detected by mostly physical place visit in most of the world. New research area aims to use Social Media data and Classification techniques to detect Road and Transport Related issues. As current work has been done using Twitter data using Word2Vec classification algorithm to detect various issues at major cities of India. During our research work we plan to use Social Media Data and News Data with Word2Vec classification algorithm. We also intend to include tweets from local languages .We also planned to apply Word2Vec classification for Multi-Class Classification Problem and NLP for feature extraction. We have used cosine similarity function to find similarity among tweets and after find similarity we have obtained result from tweets into positive, negative, neutral with the help of semantic analysis.**

***Index Terms*: *Data Mining, Python,, Word2Vec, Social Media, Event Detection, Semantic Analysis, Cosine Similarity*.**

I. INTRODUCTION

During our Research Work we have use Social Media Data with Word2Vec classification algorithm. We also include tweets from local languages and apply Word2Vec classification for Multi-Class Classification problem and NLP (Natural Language Processing) for feature Extraction.

Data Mining & Business Intelligence have different concept, but there is a one big major role of generating the output in a efficient way that contribute in success of a business.[3][4] The term Data mining which used for mining data is actually opposite to the term Business Intelligence, when it refer to the cleansing, standardizing & Utilizing Business Data. Business Intelligence main role is to focus on monitoring of datasets & Key Performance Indicators (KPI).[4]

In the Previous Research some researchers have found their time to identify the traffic incident by developing an algorithm to spot the eventin real time by using the physical sensors[1],[2]. 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.[5]

II. EXISTING WORK

In the paper: "Faceoff: Travel Habits, Road Conditions and Traffic City Characteristics Bared Using Twitte "AMIT AGARWAL AND DURGA TOSHNIWAL", they have used data of twitter for finding the issues of peoples through tweets & for that theu have user various keywords and expand the tweets into many other categories. They are discussed below.

Step:1 Read Data Set Using Twitter API
Step:2 Data Preprocessing

- Plaintext Extraction
- Hash Tag Removal
- URL Removal
- Handle Removal
- Stop Word Removal
- Typo Correction

Step:3 Category Wise Extraction of Data

- Accident
- Potholes
- Congestion

Step:4 Semantic Analysis Using Word2Vec

Step:5 Remove Pragmatic Ambiguity from Segregated Data

Step:6 Detect Content Based Location from Textual Tweet Data

Step:7 Display Result

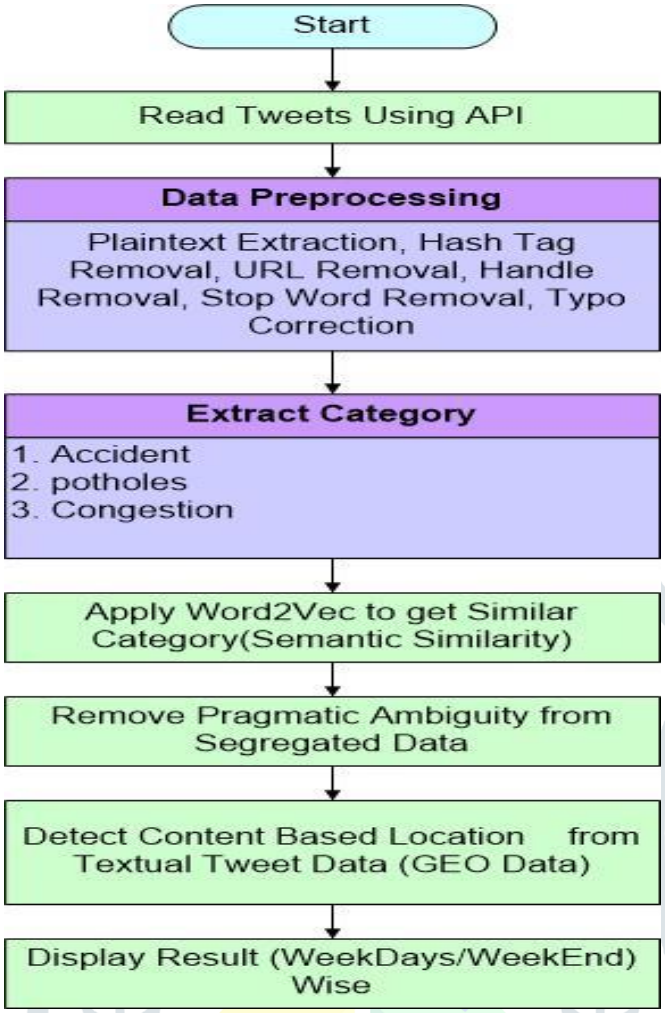


Figure 1. Existing Flow

III. Result Analysis Existing Flow

	Bengaluru 254823(H) 433133 (L)		Chennai 201435 (H) 172262 (L)		Kolkata 98624 (H) 205907 (L)		Mumbai 833574 (H) 820313 (L)	
Keywords	SD_KD	W2V_KD	SD_KD	W2V_KD	SD_KD	W2V_KD	SD_KD	W2V_KD
Pothole (H) (L)	5508 290	6115 391	1030 14	1115 37	831 22	907 56	43865 1345	87512 1922
Accident (H) (L)	8431 400	7553 1378	4711 92	7010 382	2284 138	7326 625	16581 841	29252 3112
Traffic (H) (L)	32464 2170	35273 2658	22988 4125	63020 9125	8534 258	8783 491	84815 5145	86096 5612

IV. PROPPOSED METHODOLOGY

I have gone through Face-off: Travel Habits, Road Conditions and Traffic City Characteristics Bared Using Twitter[5][6] and found there is still scope in this domain by converting local tewwts into the English and expand the tweets categories by add traffic diversion.[8]

I have perform cosine similarity on the tweeter data set to get the similarities between the tweets to get the better accurate output as compare to the previous research and then I have obtain the result into three mainly categories positive, negative, neutral with the help of semantic analysis and this is the advance feature which I have used and I have implemented completely analysing the limitation of previous research.[9][10]

Step:1 Read Data Set Using Twitter API

Step:2 Convert Tweets of Local Language to English using Google API

Step:3 Data Preprocessing

- Plaintext Extraction
- Hash Tag Removal
- URL Removal
- Handle Removal
- Stop Word Removal
- Typo Correction

Step:4 Category Wise Extraction of Data

- Accident
- Potholes
- Congestion
- **Traffic Diversion**

Step:5 Semantic Analysis using Word2Vec CBOW Model

Step:6 Remove Pragmatic Ambiguity from Segregated Data

Step:7 Detect Content Based Location from Textual Tweet Data

Step:8 Apply Fusion Model for Event Detection

Step:9 Display Result

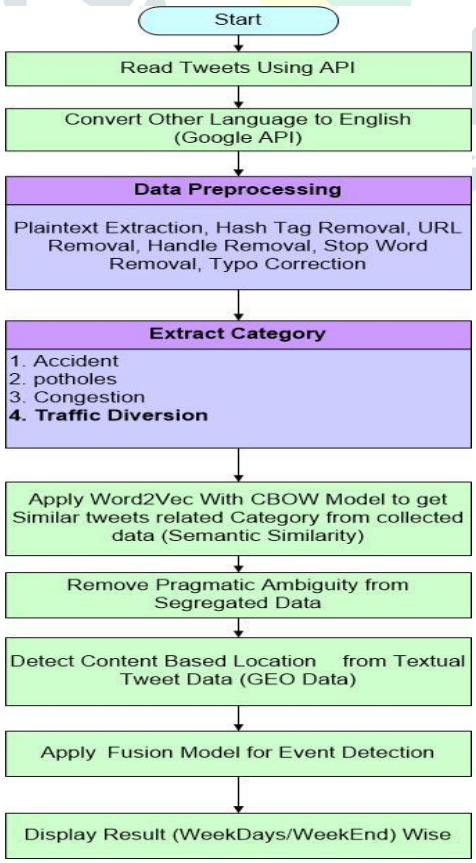


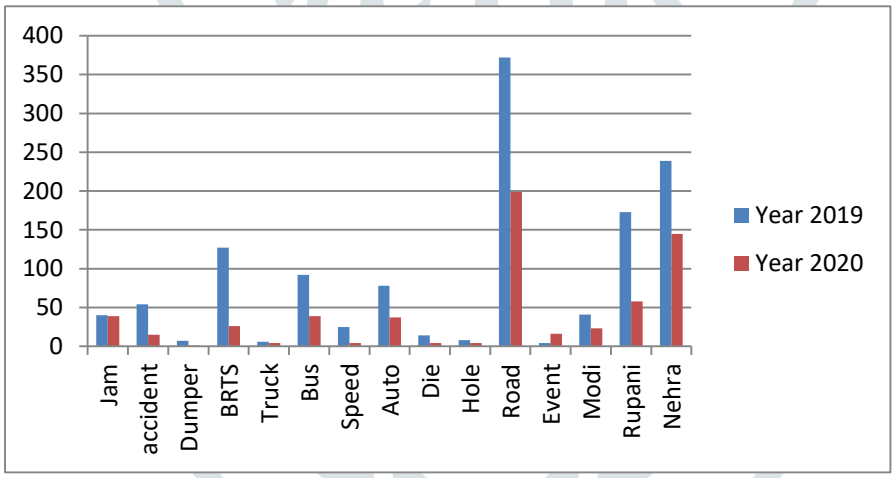
Figure2. Proposed Flow

V. RESULT ANALYSIS

We have obtained result to year 2019 to 2020 of Ahemdabad Location Traffic Related Tweets.

Tweets	Year 2019	Year 2020
Jam	40	39
accident	54	15
Dumper	7	1
BRTS	127	26
Truck	6	4
Bus	92	39
Speed	25	4
Auto	78	37
Die	14	4
Hole	8	4
Road	372	199
Event	4	16

Location	Year 2019	Year 2020
RiverFront	6	6
Paldi	7	3
RingRoad	9	11
Naranpura	2	3
Highway	14	13
Bopal	7	8
Airport	7	3
Nikol	5	3
Naroda	9	1
Maninagar	5	5



VI. CONCLUSION

In our research we have used Twitter data to analyses traffic related issues and we have used Different categories to analyze the tweets accurately and we have used different types of technique Wor2Vec, Fusion Model Etc.

We selected ahemdabad location for our research and after go to all things our research came to a final conclusion that peoples of ahemdabad are not to much traffic problem on twitter. Peoples of west Ahemdabad are sharing more tweets as compare to tweets of east ahemdabad. Areas like kalupur, laldarwaja are also not discussed in Social Media Regarding traffic issue. Still there is scope to Work on this Domain in terms of Flexibility and Efficiency by analyzing the tweets Deeply.

VII.REFERENCES

[1] 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.jan
[2] S. Grosenick, ‘‘Real-time traffic prediction improvement through semantic mining of social networks, ’’Ph.D. dissertation, Univ. Washington, Seattle, WA, USA, 2012.
[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] A. Agarwal, B. Gupta, G. Bhatt, and A. Mittal, ‘‘Construction of a semiautomated model for FAQ retrieval via short message service,’’ in Proc. 7th Forum Inf. Retr. Eval., Dec. 2015, pp. 35–38.
[5] L.V.Subramaniam,S.Roy,T.A.Faruquie,andS.Negi,‘‘Asurveyoftypes of text noise and techniques to handle noisy text,’’ in Proc. 3rd Workshop Anal. Noisy Unstructured Text Data, Jul. 2009, pp. 115–122.

[6] 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.
[7] 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.
[8] D. M. Blei, A. Y. Ng, and M. I. Jordan, ‘‘Latent Dirichlet allocation,’’ J. Mach. Learn. Res., vol. 3, pp. 993–1022, Mar. 2003.
[9] Y. Gu, Z. S. Qian, and F. Chen, ‘‘From Twitter to detector: Real-time traffic incidentdetection usingsocial mediadata,’’Transp. Res. C, Emerg. Technol., vol. 67, pp. 321–342, Jun. 2016
[10] D. Wang, A. Al-Rubaie, S. S. Clarke, and J. Davies, ‘‘Real-time traffic eventdetectionfromsocialmedia,’’ACMTrans.InternetTechnol.,vol.18, no. 1, p. 9, Dec. 2017.