

Trouble on the Road: Finding Reasons for Commuter Stress from Tweets

Reshmi Gopalakrishna Pillai, Mike Thelwall and Constantin Orasan

Research Institute in Information and Language Processing

University of Wolverhampton, UK

reshmi.g85@gmail.com, {m.thelwall, c.orasan}@wlv.ac.uk

Abstract

Intelligent Transportation Systems could benefit from harnessing social media content to get continuous feedback. In this work, we implement a system to identify reasons for stress in tweets related to traffic using a word vector strategy to select a reason from a predefined list generated by topic modeling and clustering. The proposed system, which performs better than standard machine learning algorithms, could provide inputs to warning systems for commuters in the area and feedback for the authorities.

1 Introduction

Transportation systems connect hubs of human settlements and facilitate the movement of goods and people, with limiting congestion and accidents being a key design goal for cities. Continuous updates about traffic bottlenecks and accidents can help with this. The social web is a potential feedback source for transportation systems, providing insights about the experiences and mental states of commuters. Stress is a key factor to monitor since transportation problems of all kinds are likely to increase stress.

In this study, we implement a framework for finding the causes of stress expressed in tweets related to traffic. Our system identifies specific reasons for stress of commuters (accidents, congestion, etc.) that could be used to generate automated warnings for other travelers and feed in to context-aware GPS devices. Commuters can then take informed decisions to opt for alternate routes, avoiding traffic bottlenecks. Urban planning authorities

can also leverage the analysis of stress reasons to take remedial actions as part of their Intelligent Transportation Systems strategy.

In this study we collected tweets about traffic in London during July 2018 and analyzed them to understand the reasons for the stress expressed by the commuters. As a pre-processing step, a list of potential stressors in the traffic domain was found by topic modelling and k-means clustering. Three different word-vector based methods were then applied to tweets to select a stressor from the stressors list. The output is evaluated by comparing it with the stress reason selected by human annotators.

The contributions of this work are as follows:

1. This is the first study detecting reasons for stress expressed in traffic-related tweets.
2. A dataset of traffic-related tweets annotated with reasons for stress.

2 Related Work

2.1 Stress Detection from Social Media

Social media has become a source of data for mental health analysis and evaluation. Lin et al. (2015) proposed a factor graph model combined with CNN based on linguistic, visual and social interaction data to detect stress from social media content. There have been several studies on detecting mental health disorders from social media data. De Choudhury et al. (2013a) leveraged behavioural patterns from social media, such as decreased activity, increased negative sentiment, religious involvement and clustered ego networks to build a classifier to proactively find the risk for depression in individuals before the onset. De Choudhury et al. (2013b) introduced a statistical model for predicting the onset of post-partum depression in mothers

with an accuracy of 71% using prenatal data and 80-83% when utilizing postnatal data as well. Also, the anonymity of mental health related postings in Reddit and online forums is an important factor which is discussed by Umashanthi (2015). Copper-smith et al. (2014) analysed linguistic features of Tweets of individuals with Post Traumatic Stress Disorder and built a classifier based on it. Stress can manifest around specific incidents, such as gun violence (Saha, 2017), and student deaths (Saha, 2018).

Thelwall et al. (2016) introduced TensiStrength, a novel lexicon-based system to detect stress/relaxation scores from social media based on linguistic resources such as LIWC (Tausczik and Pennebaker, 2010), General Inquirer (Stone et al, 1986) and the sentiment analysis software SentiStrength (Thelwall et al, 2010; Thelwall et al, 2012). Incorporation of word sense disambiguation (Gopalakrishna Pillai et al, 2018) improved the performance of TensiStrength.

Detecting reasons for stress from social media is a significant challenge. Lin et al. (2016) introduced a comprehensive scheme for identifying stressor event and subject and finding the stress reason and level based on it. This study is limited to personal stressor events, such as divorce, death, and relationships.

2.2 Natural Language Generation from Tweets

Traffic is a relatively new domain for the application of Natural Language Generation. However, with the proliferation of social media and accurate location devices, there is tremendous potential for data-driven, automatically generated messages about traffic incidents. Tran and Popowich (2016) introduced a novel generation model to provide location-relevant information. This traffic notification system trains a model to generate natural language texts and deliver real time warnings in case of traffic events.

Generating route descriptions is another application of NLG in traffic. Dale, Geldoff and Prost (2005) used GIS data as input and applies NLG principles to produce output texts. It used discourse structure to understand route structure and aggregation techniques to form fluent and natural sentences.

2.3 Analysis of traffic-related tweets

Lenormand et al. (2014) analyzed geo-located tweets from roads and railways in European countries. This study showed a positive correlation between the number of tweets and Average Annual Daily Traffic (AADT) on highways in France and the UK, especially in long highway segments.

Kurniawan et al. (2016) proposed tweets as an alternate source to detect traffic anomalies. Support Vector Machines achieved a classification accuracy of 99.77% in the task of detecting traffic events in Yogyakarta province in Indonesia. In a similar study, D'Andrea et al. (2015) implemented an SVM classification model to distinguish traffic from no-traffic tweets with an accuracy of 95.8% and to distinguish externally caused traffic events with an accuracy of 88.9%.

Long-term traffic prediction through tweet analysis has been shown to be effective by He et al. (2014) using traffic and Twitter data originating from San Francisco Bay in California. A cloud based system proposed by Sinnott and Yin (2015) has identified and verified accident black spots in the Australian city of Melbourne.

Gu, Cian and Chen (2016) discussed mining tweets as an inexpensive and novel way to find traffic incidents. Using a keyword dictionary, traffic incident tweets were identified, geocoded and then classified into one of five incident categories.

Cottrill and Gault (2017) analyzed a case study of a Twitter account @GamesTravel2014 used during the Commonwealth Games in Glasgow 2014, to share and respond to traffic related information. This Twitter account, in collaboration with the leading transportation providers in the city, was instrumental in detecting traffic hotspots. This case study establishes social media as a powerful tool to share trusted traffic information.

Salas and Georgakis (2017) created a collection of Tweets which mention traffic events in the UK. A key contribution of this work is a methodology with 88.27% accuracy for crawling, preprocessing and classifying traffic tweets using Natural Language Processing and Support Vector Machines. The dataset was analyzed to find the temporal and linguistic features of the Tweets.

Lv and Chen (2016) summarized the main research topics in transportation research using social media. However, though traffic-related tweets have been studied for identification of traffic events, the sentiments of the commuters using transportation systems have been largely unexplored. Cao and

Zeng (2014) proposed Traffic Sentiment Analysis as a new tool to analyze sentiments expressed in traffic related social media content. In the context of the ‘yellow light rule’ and fuel prices in China, they demonstrated the architecture, data collection and process of their methods. Our system is the first attempt to identify stress reasons in the social media outputs of commuters, providing feedback about the improvement areas for transportation systems.

3 Methodology

We use a two-step method developed in our earlier research on finding stress reasons for airlines and politics (Gopalakrishna Pillai et al., 2018b). High-stress tweets belonging to the traffic domain were first analyzed by topic modelling to find the most frequently occurring topics. A list of potential reasons for stress in traffic was constructed from it. In the second step, tweets were analyzed by word-vector methods to find the reason for stress from the potential list.

3.1 Overview

Construction of Potential Stressors list: Our study was limited to tweets discussing traffic in the city of London. The tweets collected were pre-processed by removing duplicates. They were assigned stress scores by TensiStrength, on a scale of -1 to -5, (-1 denote the least stress and -5 the highest stress).

The higher stress scored Tweets (with score -5 or -4) constituted the corpus for performing batch-wise LDA topic modeling with hashtag pooling. The topics from this step were grouped into 5 clusters (the number of clusters was chosen based on the coherence of the collection).

A label was assigned to each cluster using the most similar word vector method. These cluster labels encasing the most frequent topics in high-stress tweets constituted the potential stressors for the traffic domain (Table 1).

Example topics in the cluster	Stressor
Air cleanair smoke fog burning noise emissions mess	Pollution
Accident damage collision break-down fatality	accident
Race hour rush delay late hours slow jam busy mayhem chocker	Congestion
Attack assault gunman fear chaos kill crime murder terrorist blast stab	Violence
Awareness counsel protest march rally public crowd	Campaign

Table 1: Clusters and stressors (London Traffic)

Finding stress reasons for tweets: As previously introduced for politics and airlines (Gopalakrishna Pillai et al., 2018b), we employed 3 word-vector based methods to find causes for stress in Tweets.

Method 1 (maximum word similarity): Select the stressor with the highest cosine similarity with any of the content words in the tweets.

Method 2 (context vector similarity): Select the stressor with the highest cosine similarity with the context vector representing the tweet (average of vectors corresponding to the content words).

Method 3 (cluster vector similarity): This is the same as Method 2, but the stressor is represented by a cluster vector found by averaging vectors of all words in the topic cluster. Select the cluster with highest cosine similarity with the context vector.

3.2 Dataset and Annotation

We first collected 23249 tweets from 1st to 31st July 2018 with the Twitter API. The search queries were strings and hashtags related to traffic in London and two key motorways (‘London’ AND ‘traffic’, #london AND #traffic, #londontraffic, #m25, #m40). Removing tweets consisting only of URLs, and duplicate tweets, left 13321 tweets. Using TensiStrength, these tweets were given scores in a scale of -1 to -5 to indicate their stress level. There were 2334 tweets with a stress score of -4 or -5. A high-stress traffic tweet corpus was created by randomly selecting the 1000 tweets from this set. This was divided into 5 groups with 200 tweets each and each group was subjected to LDA-based topic modelling with hashtag pooling. A list of potential

stressors in the traffic domain was created as described in the previous section.

Out of the 4410 tweets with -5, -4 or -3 stress scores, the 1000 tweets which were used for finding the potential stress reasons were excluded and from the remaining tweets, 2000 were randomly chosen for evaluation of the word vector methods. We included Tweets with stress score -3 too in this dataset because this is a sufficiently high stress score for the annotators and the word vector analysis to assign a stress reason.

These 2000 tweets were annotated individually and independently by three human coders, marking them with the most appropriate reason for stress from the list of potential stressors. The annotators had engaged in a similar stress-reason annotation experiment, and their reliability was further assessed with Krippendorff’s α inter-coder agreement scores. The agreement rates were sufficiently high to claim that the annotations are coherent and usable (Table 2).

Agreement Between Coders	Krippendorff’s α
A and B	0.732
B and C	0.786
A and C	0.721

Table 2: Inter-coder agreement for stressor annotation (London Traffic)

3.3 Experimental Setup

The experiments performed were similar to our earlier research on finding stress reasons for airlines and politics (Gopalakrishna Pillai et al., 2018b). To train the word vectors for finding stressors, we used a Word2Vec model trained on 400 million tweets, from an ACL WNUT task (Godin et al, 2015).

The default Weka 3.6 configurations of three machine learning algorithms (AdaBoost: An adaptive boosting algorithm, Logistic Regression and Support Vector Machines) were run using 10-fold cross validation to serve as comparison baselines for our methods. The feature selection was adapted from a similar task of assessing the stress and relaxation strengths expressed in tweets (Thelwall, 2017). Term unigrams, bigrams and trigrams and their frequencies were used as features. Punctuation was included as a term, with consecutive punctuation treated as a single term.

4 Results

4.1 Results Summary

The stress reasons for the traffic tweets were found using the word vector processing methods. It is summarized in Table 3. Cluster vector method gives the best performance in terms of precision, recall and accuracy.

Figure 1 shows a distribution of stress reasons in the traffic Tweets.

Method	Percentage Correct	Precision	Recall
max. word	48.3%	47.6%	46.3
context vector	53.9%	52.1%	52.5%
cluster vector	63.8%	62.3%	61.4%
SVM	50.1%	52.2%	55.7%
AdaBoost	52.5%	54.3%	52.8%
Logistic	49.7%	43.6%	48.7%

Table 3: Accuracy of stress reason detection methods applied to London Traffic tweets

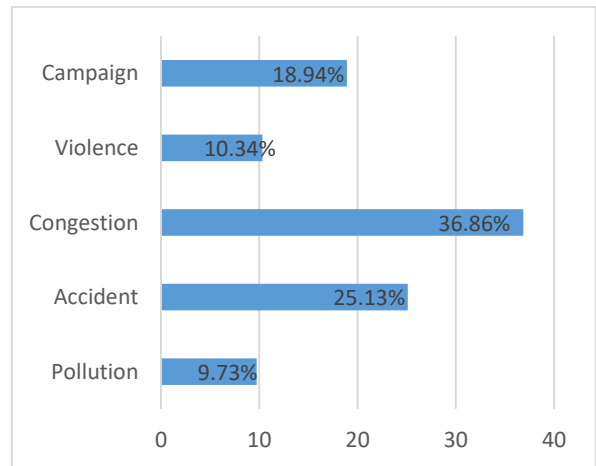


Figure 1: Stress reasons detected in London Traffic Tweets by the cluster vector method

4.2. Error Analysis

Indirect/Sarcastic Expressions: Tweets with an indirect expressions of stress pose a challenge to our methods. An example is “The joys of London traffic; not moving for last one hour is killing me”, “Lea Bridge rd E10 traffic is murder really London counsel” in which the reason for stress is Congestion, but is detected as violence.

Multiple stressors: Tweets in which there are multiple reasons for stress. E.g.: “Vehicle emissions rise during rush hours, making the traffic jams hellish” has two stressors, “pollution” and “congestion”. A possible solution will be to expand the methods to accommodate multiple stressors.

5 Conclusion and Future Work

This paper proposed and implemented word vector based methods to find the reasons for stress expressed in Traffic domain. A dataset containing 23249 Tweets about London traffic were collected and after preprocessing, 2000 tweets were randomly chosen to be annotated by human coders for the reasons for stress. The performance of our proposed methods to detect the reasons for stress in this dataset was compared to that of standard machine learning algorithms. As a future work, we propose to extend this study to content in other social media and to modify our methods to accommodate Tweets with multiple stressors.

This automated stress detection from tweets can identify traffic bottlenecks and accident-prone areas. These reasons identified for traffic-related stress can be used as inputs in the form of pictograms or short texts to an automated warning system for the commuters and feedback for the traffic policymakers. Further research is required to use the findings of this stress reason detection method for improving Intelligent Transportation Systems.

Acknowledgement

This research was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No 636160-2, the Optimum project www.optimumproject.eu.

References

Jianping Cao et al., "Web-Based Traffic Sentiment Analysis: Methods and Applications," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 15, no. 2, pp. 844-853, April 2014. doi: 10.1109/TITS.2013.2291241

Glen A. Coppersmith, Craig T. Harman, and Mark Dredze. 2014. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Caitlin Cottrill, Paul Gault, Godwin Yeboah, John D. Nelson, Jillian Anable, Thomas Budd, Tweeting Transit: An examination of social media strategies for transport information management during a large event, *Transportation Research Part C: Emerging Technologies*, Volume 77, 2017, Pages 421-432, ISSN 0968-090X, <https://doi.org/10.1016/j.trc.2017.02.008>.

Robert Dale and Sabine Geldof and Jean-Philippe Prost. (2005). Using Natural Language Generation in Automatic Route Description. *Journal of Research and Practice in Information Technology*. 37.

Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzarini, Francesco Marcelloni "Real-Time Detection of Traffic From Twitter Stream Analysis," in *IEEE Transactions on Intelligent Transportation Systems*, vol. 16, no. 4, pp. 2269-2283, Aug. 2015. doi: 10.1109/TITS.2015.2404431

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013a. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.

Munmun De Choudhury, Scott Counts, Eric Horvitz, 2013b. Predicting Postpartum Changes in Behavior and Mood via Social Media. In *Proc. CHI 2013*, to appear.

Reshmi Gopalakrishna Pillai, Mike Thelwall, Constantin Orasan. 2018a. Detection of Stress and Relaxation Magnitudes for Tweets. In *The 2018 Web Conference Companion (WWW'18 Companion)*, April 23-27, 2018, Lyon, France, ACM, New York, NY, 8 pages. DOI: <https://doi.org/10.1145/3184558.3191627>

Reshmi Gopalakrishna Pillai, Mike Thelwall, Constantin Orasan, 2018b. What Makes you Stressed? Finding Reasons From Tweets. *Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, EMNLP-2018*.

Yiming Gu, Zhen Qian, Feng Chen From twitter to detector: real-time traffic incident detection using social media data. *Transport. Res. Part C: Emerging Technol.*, 67 (2016), pp. 321-342

Jingrui He, Wei Shen, Phani Divakaruni, Laura Wynter and Rick Lawrence. “Improving Traffic Prediction with Tweet Semantics.” *IJCAI* (2013).

Dwi Aji Kurniawan, Sunu Wibirama and Noor Akhmad Setiawan, "Real-time traffic classification with Twitter data mining," 2016 8th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta-ta, 2016, pp. 1-5. doi: 10.1109/ICITEED.2016.7863251

- Maxim Lenormand, Antonia Tugores, Pere Colet, Jose J Ramasco (2014) Tweets on the Road. PLoS ONE 9(8): e105407. <https://doi.org/10.1371/journal.pone.0105407>
- Huijie Lin, Jia Jia, Jiezhong Qui, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tan, Ling Feng, Tat-Seng Chua. Detecting stress based on social interactions in social networks. 2017. IEEE Transactions on Knowledge and Data Engineering, 2017.
- Huijie lin, Jia Jia, Lexing Xie, Guangyao Shen, Tat-Seng Chua. 2016. What does social media say about your stress? In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, 3775–3781.
- Yisheng Lv, Yuanguang Chen, Xiqiao Zhang, Yanjie Duan and Naiqiang L. Li, "Social media based transportation research: the state of the work and the networking," in IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 1, pp. 19-26, Jan. 2017. doi: 10.1109/JAS.2017.7510316
- Umashanthi Pavalanathan and Munmun De Choudhury. 2015. Identity Management and Mental Health Discourse in Social Media. In Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion). ACM, New York, NY, USA, 315-321. DOI: <https://doi.org/10.1145/2740908.2743049>
- Koustuv Saha, and Munmun De Choudhury. 2017. Modeling stress with social media around incidents of gun violence on college campuses.
- Koustuv Saha, Ingmar Weber and Munmun De Choudhury. 2018. A Social Media Based Examination of the Effects of Counseling Recommendations After Student Deaths on College Campuses.
- Angelica Salas, Panagiotis Georgakis, C. Nwagboso, Ahmad Ammari and Ionnis Petalas, "Traffic event detection framework using social media," 2017 IEEE International Conference on Smart Grid and Smart Cities (ICSGSC), Singapore, 2017, pp. 303-307. doi: 10.1109/ICSGSC.2017.8038595
- Richard Sinnott, Yikai Gong & Fengmin Deng (2015). Identification of (near) Real-time Traffic Congestion in the Cities of Australia through Twitter. 7-12. 10.1145/2811271.2811276.
- Philip J. Stone, Dexter C. Dunphy, Marshall S. Smith and Daniel M. Ogilvie. 1966. The general inquirer: A computer approach to content analysis. Cambridge, MA: The MIT Press.
- Yla R. Tausczik and James W. Pennebaker, 2010. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.
- Mike Thelwall, Kevan Buckley Georgios Paltoglou, D. Cai and A. Kappas. 2010. Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), 2544–2558.
- Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. 2012. Sentiment strength detection for the social web. J. Am. Soc. Inf. Sci. Technol. 63, 1 (January 2012), 163-173. DOI=<http://dx.doi.org/10.1002/asi.21662>
- Mike Thelwall. 2017. TensiStrength: stress and relaxation magnitude detection for social media texts. Journal of Information Processing and Management. 53: 106–121
- Khoa Tran and Fred Popowich, Automatic Tweet Generation From Traffic Incident Data, Proceedings of the 2nd International Workshop on Natural Language Generation and the Semantic Web (WebNLG 2016), 2016, Association for Computational Linguistics, 59-66.