Transportation Research Part C 86 (2018) 580–596

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

A deep learning approach for detecting traffic accidents from social media data

Zhenhua Zhang[a], Qing He[b,c,*], Jing Gao[d], Ming Ni[c]

[a] HERE North America LLC, Chicago, IL 60606, United States
[b] Department of Civil, Structural and Environmental Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States
[c] Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States
[d] Department of Computer Science and Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States

ARTICLE INFO

Keywords:
Traffic accident detection
Tweet
Social media
Association rules
Deep learning

ABSTRACT

This paper employs deep learning in detecting the traffic accident from social media data. First, we thoroughly investigate the 1-year over 3 million tweet contents in two metropolitan areas: Northern Virginia and New York City. Our results show that paired tokens can capture the association rules inherent in the accident-related tweets and further increase the accuracy of the traffic accident detection. Second, two deep learning methods: Deep Belief Network (DBN) and Long Short-Term Memory (LSTM) are investigated and implemented on the extracted token. Results show that DBN can obtain an overall accuracy of 85% with about 44 individual token features and 17 paired token features. The classification results from DBN outperform those of Support Vector Machines (SVMs) and supervised Latent Dirichlet allocation (sLDA). Finally, to validate this study, we compare the accident-related tweets with both the traffic accident log on freeways and traffic data on local roads from 15,000 loop detectors. It is found that nearly 66% of the accident-related tweets can be located by the accident log and more than 80% of them can be tied to nearby abnormal traffic data. Several important issues of using Twitter to detect traffic accidents have been brought up by the comparison including the location and time bias, as well as the characteristics of influential users and hashtags.

1. Introduction

Traffic accidents disturb the traffic operations, break down the traffic flow, and cause severe urban problems worldwide. Major traffic accidents can sometimes lead to irreparable damages, injuries, and even fatalities. National Highway Traffic Safety Administration (NHTSA), which publishes yearly reports on traffic safety facts, states that since 1988 more than 5,000,000 car crashes occur in the States each year and about 30% of them result in fatalities and injuries (NHTSA, 2015). After years of research, it has been widely accepted that significant reductions of accident impact can be achieved through effective detection methods and corresponding response strategies. As an essential component of traffic incident management, accurate and fast detection of traffic accidents are critical to modern transportation management (He et al., 2013).

Traditional detector-based methods usually give accurate location and time of the traffic accident and have been proved valid in many applications (Hall et al., 1993; Samant and Adeli, 2000; Sethi et al., 1995). Despite the adaptabilities of previous studies,

* Corresponding author at: Department of Civil, Structural and Environmental Engineering and Department of Industrial and Systems Engineering, University at Buffalo, The State University of New York, Buffalo, NY 14260, United States.
E-mail addresses: zhenhua.zhang@here.com (Z. Zhang), qinghe@buffalo.edu (Q. He), jing@buffalo.edu (J. Gao), mingni@buffalo.edu (M. Ni).

https://doi.org/10.1016/j.trc.2017.11.027
Received 30 May 2017; Received in revised form 27 November 2017; Accepted 29 November 2017

traditional detection methods with only traffic data still meet certain challenges. First, most of the previous research, which utilized the field data to detect the traffic accidents, build on an implicit assumption that the data is reliable. However, detector failures and communication errors are perennial problems in traffic operations. For example, Illinois Department of Transportation (IDOT) in Chicago reported that around 5 percent of their loops (detectors) are inoperative at any given time (Kell et al., 1990). The problem of malfunctioned sensors could cause even more troubles in accident detection in large regions. Second, the uncertainty nature of traffic patterns and non-recurrent events may undermine the potential of traffic metrics in justifying the traffic accidents. Besides traffic accidents, daily traffic operations may suffer breakdowns by other factors such as parades, road construction, running races, etc. Thus, the metrics including the traffic flow and occupancy inherently perform as an indirect support for traffic accidents instead of direct proof. To address these challenges, there are efforts in applying clustering or classification methodologies such as K-means (Münz et al., 2007) on large data collections to diminish the errors. Our counter-measure lies in extracting "direct report" from tweet users, and the applicability is fully discussed. At the same time, one of the traditional methods is also employed to validate our results in later sections.

In recent years, the accident-related studies have witnessed the power of data crowdsourcing in complementing the traditional methods and finding new knowledge. In this study, we explore the possibility of using Twitter to detect the traffic accidents. Twitter, the microblogging service that received increasing attention in recent years, has been gradually accepted as a direct user-contributed information source in event detection. Twitter establishes an online environment where the content is created, consumed, promoted, distributed, discovered or shared for purposes that are primarily related to communities and social activities, rather than functional task-oriented objectives (Gal-Tzur et al., 2014). Thus, each tweet acts as a data source of "We Media", and it is entirely possible to retrieve the wide-range information from the broad masses of people in a timely manner. Our preliminary examinations also demonstrate the potential of Twitter in delivering the accident-related information.

As we can see from Table 1, the Twitter information is both noisy and unstructured. An effective text mining method is necessary to extract the useful accident-related information from tweets. In this study, we employ and compare two deep learning methods: Deep Neural Network (DNN) and Long Short-Term Memory (LSTM), in training and classifying the accident-related tweets. Unlike classifiers such as logistic regression, Support Vector Machines (SVMs) or Artificial Neural Network (ANN) with a single hidden layer, deep learning does not seek direct functional relationships between the input features and the output classification results. Instead, it is a set of machine learning algorithms that attempt to learn in multiple levels, corresponding to different levels of abstraction (Deng and Yu, 2014). The training process of DNN is divided into multiple layers, and the output result is expressed as a composition of layers, where the higher level features are the composition of lower-level features, giving the potential of modeling complex data with fewer units than a similarly performing shallow network (Bengio, 2009). Our efforts can be detailed by three major contributions: First, we propose a systematic feature selection process in extracting both the individual and paired token features from social media. We unveil the language customs of the tweet users in describing the traffic accident detection. Second, we validate the effectiveness of the deep learning approach in classifying the social media data. The results show that deep learning outperforms other prevailing data mining methods. Third, the advantages and disadvantages of tweets in accident detection are verified and fully discussed by comparing tweets with both accident log from state Department of Transportation and traffic data from thousands of loop detectors.

The rest of the paper is organized as follows: Section 2 reviews the current studies in social media applications in transportation and the deep learning in language modeling. Section 3 introduces tweet data preprocessing for accident detection. Section 4 details the process of feature selection from tweet contexts including both the individual token features and paired token features. Section 5 introduces the two deep learning method: DBN and LSTM; and their performances in classifying the tweets as compared with SVMs and sLDA. Section 6 validates the accident-related tweets by the traffic accident log and loop-detector data. Section 7 concludes the paper with a few empirical findings and generalizations together with some thoughtful discussions.

2. Literature review

2.1. Review of social media in traffic-related studies

The newly emerged data source, social media data, has proved its capability in recent traffic studies including activity pattern identification (Hasan and Ukkusuri, 2014), special traffic-related events (Ni et al., 2014; Shirky, 2011), traffic flow prediction (Cottrill et al., 2017; Lin et al., 2015; Ni et al., 2017), transport information management (Cottrill et al., 2017), travel mode detection (Maghrebi et al., 2016), destination or route choice (Huang et al., 2017), etc. According to Rashidi et al. (2017), as social media data encompasses information that is revealed by users in realistic situations, such data is free from sampling, surveying or laboratory biases. The location effectiveness and timeliness features of Twitter can be proved in a recent accident detection study that uses the

Table 1
Tweet samples describing the general traffic information, general traffic incident and road accident, respectively.

General information	*"I am waiting at the silver line, exciting"*
	"Always hate the signals ahead of the hip-hop, making me sick"
General incident	*"standstill for 1 h, there must be accidents in front"*
	"this is typical NOVA traffic, what a bad day"
Traffic accident	*"major accident next to the sunoco near the parkway a car got flipped over"*
	"the worst car accident possible just happened in front of me"