American Journal of Engineering and Applied Sciences

Original Research Paper

Block-Matching Twitter Data for Traffic Event Location

Amal Shuqair and Samuel Kozaitis

Department of Electrical and Computer Engineering, Florida Institute of Technology, Melbourne, FL, USA

Article history
Received: 09-03-2017
Revised: 14-03-2017
Accepted: 08-04-2017

Corresponding Author:
Samuel Kozaitis
Department of Electrical and Computer Engineering, Florida Institute of Technology,
Melbourne, FL, USA
Email: kozaitis@fit.edu

Abstract: We used a block-matching approach that is data-driven and relies mostly on patterns of tagged speech in Twitter streams as a way to identify events in road traffic. Events are useful because their location may identify the status of road segments, especially when cross-street data are available. Basing a system on patterns that are not pre-defined has the advantage of flexibility for a variety of scenarios.

Keywords: Block-Matching, Parts of Speech, Social Media, Twitter

Introduction

Road traffic congestion is a problem for many areas around the globe. This can be for a variety of reasons and may occur even when it is not normally a problem, such as when natural disasters occur. Social networking services are widely used by people sharing personal information about status and events. Twitter is one example where such information is shared, is accessible to the public and has a large volume of use. Since there are many users sharing information within a congested situation, there have been several methods developed to extract information from Twitter for use in traffic analysis.

There is a commonality among methods in that tweets are usually detected, tokenized and pruned, but there are different specifics (Kulkarni *et al*., 2016; Shuqair and Kozaitis, 2015). It has been shown that tweets alone can be used to estimate road traffic congestion conditions as tweet density and hours of the day are useful attributes for building a congestion severity prediction model (Wongcharoen and Senivongse, 2016). In addition, using official and public tweets were correlated to identify mobility patterns (Rebelo *et al*., 2015).

A real-time monitoring system with particular reference to traffic congestion and car accidents from Twitter stream analysis has been developed (D’Andrea *et al*., 2015). It used text mining techniques and then classification to determine if a tweet is traffic related. The system was reported to identify issues, often before traffic news web sites. It was also able to discriminate whether traffic was caused by an external event or not by solving a multiclass problem.

There have been several other approaches to using Twitter and social media to provide real-time traffic information by way of text mining or Natural Language Processing (NLP). In one system, words were tokenized, matched to a database and classified into one of eight categories of words such as adjective, noun, verb, etc. (Sakaki *et al*., 2010).

When considering tweets that were official in capacity, they were tokenized and classified into 12 categories that were classified as either events or situations (Ribeiro Jr *et al*., 2012). Then, an exact string matching process was used to find street names in a large database and fuzzy string searching using gazetteers was performed to match the streets by crossroads and neighborhood names. Another method used an existing tokenizer called Lexto to analyze and classify Twitter data by keywords that described traffic conditions (Wanichayapong *et al*., 2011). This approach classified road data either into a point (e.g., an intersection) or a link (e.g., a road) then into eight subcategories such as place, verb, etc. Tweets were limited to traffic keywords such as *accident* and *traffic congestion* with a large dictionary created for each category.

We examined Twitter data for cross streets and points-of-interest to help more accurately determine the status of a road segment in congested traffic. We looked for patterns of text and grouped like patterns together in what we refer to as a block-matching approach that was also data-driven by examining groups of like patterns to extract information such as the condition of a road. This approach has the advantage that it can adapt to different syntaxes and potentially different languages.

System

The system initially acquired tweets from Twitter streams and preprocessed them before further analysis.

Science Publications

Preprocessing included removing unnecessary characters and retaining tweets that were only related to traffic. Then, the tweets were broken into tokens, tagged and grouped before being classified. During classification, rules were applied that clustered and/or separated blocks of tweets. A block diagram of the system is shown in Fig. 1.

Preprocessing

Tweets were first gathered by the Twitter Streaming API and unwanted symbols were removed. Tweets were limited to a geographical area specified by a user. Symbols were removed including punctuation marks, URLs, emoticons, etc. Retaining those symbols would not help the classification process and would add unnecessary complexity to later stages in the system.

The next part of the preprocessing step involved the removal of tweets that did not pertain to traffic. In order to determine if a tweet was useful, words within the tweets first had to be tagged. To perform this operation, each tweet was separated into tokens and each token was tagged as a Part Of Speech (POS). We then compared each POS that was most likely a name to a list of street names that were determined to be in the geographical area specified by a user. If a match was found, then that tweet was retained; otherwise it was discarded. If a tweet was found that was not considered to be traffic-related, then is was discarded and another tweet was acquired.

POS Tagging

We used a text parsing POS method by means of a Stanford parser (Endarnoto *et al*., 2011) and a Carnegie Mellon parser (de Marneffe *et al*., 2006), to tag tokens. This approach allowed us to develop algorithms based on tags to classify text. Although there are many tags, our work focused only on the most popular and simplest ones such as noun, verb, etc.

Block Matching

Once a collection of tweets had been acquired, we selected the first *N* POS tags and then searched the tweets for the same *N* tags, which we refer to as blocks. Each time we found the same block, we grouped the corresponding words together. We then looked at the next block of *N* tags in the tweet and repeated the process. Eventually, we built up a collection of blocks and associated words. We continued this process until all blocks of *N* consecutive tags were used in the matching process.

Results from a simple example of the block matching process are shown in Fig. 2. The process included nine tweets that were tagged using the convention in Fig. 3a adopted from, LG (2016).

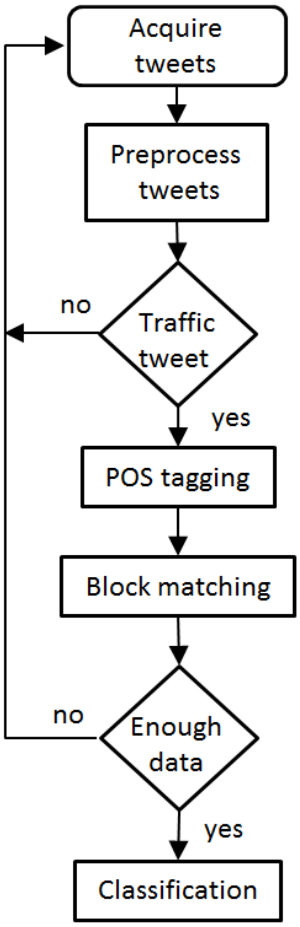
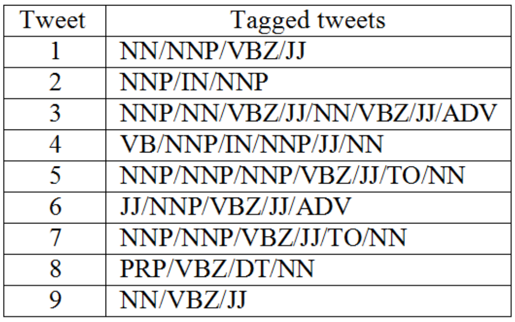


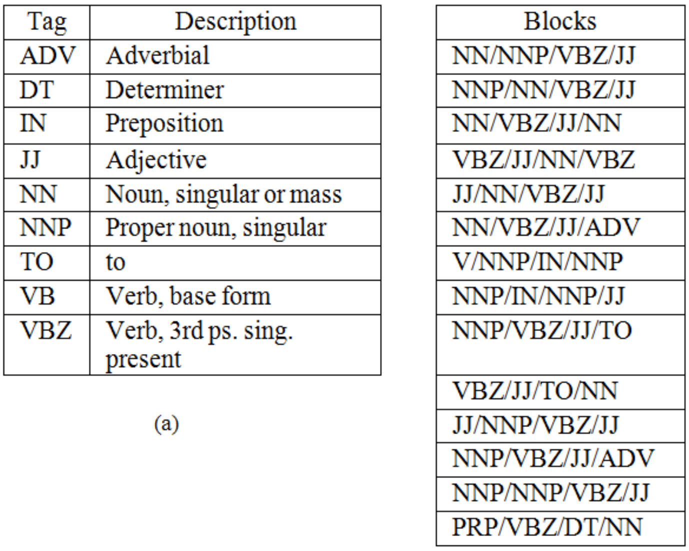
Fig. 1. Block diagram of system



Tweet	Tagged tweets
1	NN/NNP/VBZ/JJ
2	NNP/IN/NNP
3	NNP/NN/VBZ/JJ/NN/VBZ/JJ/ADV
4	VB/NNP/IN/NNP/JJ/NN
5	NNP/NNP/NNP/VBZ/JJ/TO/NN
6	JJ/NNP/VBZ/JJ/ADV
7	NNP/NNP/VBZ/JJ/TO/NN
8	PRP/VBZ/DT/NN
9	NN/VBZ/JJ

Fig. 2. Example of tagged text

Considering $N = 4$, the grouping results are shown in Fig. 3b. The first tweet contains four tags so only one grouping was possible and is shown in the first row of Fig. 3b. The second tweet only contained three tags, so that tweet was discarded because the number of tags was less than *N*. The third tweet contained eight tags, so five entries were possible and are shown in rows 2-6. The block matching results show 14 different blocks.



Tag	Description
ADV	Adverbial
DT	Determiner
IN	Preposition
JJ	Adjective
NN	Noun, singular or mass
NNP	Proper noun, singular
TO	to
VB	Verb, base form
VBZ	Verb, 3rd ps. sing. present

(a)

Blocks
NN/NNP/VBZ/JJ
NNP/NN/VBZ/JJ
NN/VBZ/JJ/NN
VBZ/JJ/NN/VBZ
JJ/NN/VBZ/JJ
NN/VBZ/JJ/ADV
V/NNP/IN/NNP
NNP/IN/NNP/JJ
NNP/VBZ/JJ/TO
VBZ/JJ/TO/NN
JJ/NNP/VBZ/JJ
NNP/VBZ/JJ/ADV
NNP/NNP/VBZ/JJ
PRP/VBZ/DT/NN

(b)

Fig. 3. Blocks resulting from POS tagging Fig. 2 with N = 4 (a) tagging convention (b) resulting blocks

Table 1. Extracting blocks of patterns

No	POS tagging	No	Resulting blocks
1	NN/IN/NNP/NN/CC/NNP/NN	1	NN/IN/NNP/NN/CC/NNP/NN
2	NN/IN/NNP/CC/NNP/JJ	2	NN/IN/NNP/CC/NNP/JJ
3	NN/IN/NNP/CC/NNP/VBZ/JJ	3	NN/IN/NNP/CC/NNP/VBZ/JJ
4	NN/IN/NNP/CC/NNP/JJ	4	NNP/VBZ/JJ/IN/NNP/CC/NNP/NN
5	NNP/VBZ/JJ/IN/NNP/CC/NNP/NN	5	NNP/VBZ/JJ/IN/NNP/CC/NNP
6	NN/IN/NNP/IN/NNP/NN/CC/NNP/NN		
7	NNP/VBZ/JJ/IN/NNP/CC/NNP		
8	NNP/JJ/IN/NNP/NN/CC/NNP/NN		
9	NNP/VBZ/JJ/IN/NNP/CC/NNP		

Rules can be applied at this level or within the classification stage to reject or manipulate blocks of tweets if necessary. Once a number of tweets have been collected, the system passed the groups of blocks to the next step for classification.

Example

Match Blocks

We chose an example to illustrate our approach that determined whether a street was possible or not. Specifically, we determined whether a street was open or closed between two cross streets. Our approach was to look for a pattern of tags that indicated a possible obstruction in a specific area, such as the intersection between two streets. We started by considering two different sets of blocks of tags. Then, we eliminated and/or combined different blocks before further classification, because they provided specific information about the location and condition of the road. Blocks used for block matching must satisfy one of the rules described below with $N = 8$.

Rule 1

- A block should consist of 8 tags if the tweet length is $N \geq 8$ tags
- A block may be less than 8 tags if the tweet length is $N \leq 8$ tags
- Blocks should start with the NNP tag
- Each block should contain at least 3 NNP tags
- Each block should have the tag IN

Rule 2

- All blocks should start with a NN noun tag
- A block should contain the pattern NN/IN/NNP
- Each block should have the tag IN

Applying these rules, we used the block matching process to create different blocks for further processing. Table 1 illustrates an example of the blocks created from example tweets. In left side of the table, several tagged tweets are shown and the right side shows the resulting blocks using the rules above.

Eliminate Blocks

We also used additional rules to elminate tweets from further processing. For example, we used the following rules for this case to eliminate tweets:

- A block starts with NNP tag is not a known street name
- A block that does not contain an NN tag with what we called traffic status nouns such as, *closed, open, accident, crash*, etc
- A block that had fewer than two street names in NNP tags

Classify Blocks

In this step, groups of blocks were formed that contained information on a particular road. Then, those blocks were examined to determine the status of a road. The conditions used to group blocks were as follows:

- Blocks that include same NNP tags in the same positions
- Blocks that include traffic nours and at least two NNP tags
- Blocks that refer to the same street will be group together

Results

The performance of the system can be altered by the user. For example, if the rules for block-matching are very specific, then we can easily determine the status of a road, but many tweets may have to be rejected, which is not necessarily practical. In general, most tweets are not about traffic; however, at the time of a significant natural disaster, weather event or crisis, traffic related tweets will be more probable.

To test our system, we distributed a map of a city that contained indications of closed roads, accidents, etc. to students without any knowledge of our system and asked them to send a traffic-related tweet. Using the example described, 39% of the tweets generated were retained as useful. Of those, 54% aided a decision to be made - if a road segment was closed or open.

A decision is ready to be made about a road segment when combing and separating groups has ceased. At this point a group may consist of a single entry or multiple entries. For multiple entries, a variety of methods can be used to identify the status of a road. A weighted average of probabilities assigned to the tweets is the most straightforward. Relative probabilities and a threshold for a closed/open decision can be assigned by the user.

Conclusion

By grouping blocks of tokenized POS Twitter tags in a data-driven approach, we were able to determine the condition of road traffic segments. This process allows for more detailed information about congested areas to help navigate away from the area. Furthermore, by identifying cross-streets near a traffic event, better paths can be found.

Funding Information

The authors have no support or funding to report.

Authors' Contributions

Amal Shuqair: Participated in all experiments, did programming, gathered results wrote draft of manuscript, did portion of literature search.

Samuel Kozaitis: Designed research plan, organized study, performed final writing of manuscript, did portion of literature search.

Ethics

This article mostly original and is an extension of a conference paper.

References

D'Andrea, E., P. Ducange, B. Lazzerini and F. Marcelloni, 2015. Real-time detection of traffic from Twitter stream analysis. IEEE Trans. Intell. Transport. Syst., 16: 2269-2283. DOI: 10.1109/TITS.2015.2404431

de Marneffe, M.C., B. MacCartney and C.D. Manning, 2006. Generating typed dependency parses from phrase structure parses. Proceedings of the International Conferences Language Resources and Evaluates, (LRE' 06).

Endarnoto, S.K., S. Pradipta, A.S. Nugroho and J. Purnama, 2011. Traffic condition information extraction and visualization from social media twitter for android mobile application. Proceedings of the International Conference on Electrical Engineering and Informatics, Jul. 17-19, IEEE Xplore Press, pp: 1-4. DOI: 10.1109/ICEEI.2011.6021743

Kulkarni, R., S. Dhanawade, S. Raut and D.S. Lavhakare, 2016. Twitter stream analysis for traffic detection in real time. Int. J. Adv. Res. Ideas Innovat. Technol., 2: 1-4.

LG, 2016. Link Grammar.

Rebelo, F., S. Soares and R.J.F. Rossetti, 2015. TwitterJam: Identification of mobility patterns in urban centers based on tweets. Proceedings of the IEEE 1st International Smart Cities Conference, Oct. 25-28, IEEE Xplore Press, pp: 1-6. DOI: 10.1109/ISC2.2015.7366156

Ribeiro Jr., S.S., C.A. Davis Jr., D.R.R. Oliveira, W. Melra Jr. and T.S. Concalves *et al*., 2012. Traffic observatory: A system to detect and locate traffic events and conditions using twitter. Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Social Networks, Nov. 06-06, ACM., Redondo Beach, California, pp: 5-11. DOI: 10.1145/2442796.2442800

Sakaki, T., M. Okazaki and Y. Matsuo, 2010. Earthquake shakes Twitter users: Real-time event detection by social sensors. Proceedings of the 19th International Conference on World Wide Web, Apr. 26-30, ACM., Raleigh, North Carolina, USA., pp: 851-860. DOI: 10.1145/1772690.1772777

Shuqair, A. and S.P. Kozaitis, 2015. Block-matching Twitter data for traffic analysis. Proceedings of the 2nd International Conference on Advances in Big Data Analytics, (BDA ’15), Part of WORLDCOMP, Las Vegas, NV, 155-159.

Wanichayapong, N., W. Pruthipunyaskul, W. Pattara-Atikom and P. Chaovalit, 2011. Social-based traffic information extraction and classification. Proceedings of the 11th International Conference on ITS Telecommunications, Aug. 23-25, IEEE Xplore Press, pp: 107-112. DOI: 10.1109/ITST.2011.6060036

Wongcharoen, S. and T. Senivongse, 2016. Twitter analysis of road traffic congestion severity estimation. Proceedings of the 13th International Joint Conference on Computer Science and Software Engineering, Jul. 13-15, IEEE Xplore Press, pp: 1-6. DOI: 10.1109/JCSSE.2016.7748850