Social Media Data Preprocessing (Tweet)

Tweet database T:
T_1: I saw a traffic accident in front.
T_2: Car damage on Route 1.
……

Tokenization

Tokens:
i saw a traffic in accident front
car damage on route 1
……

Stop-word filtering

Tokens:
~~I~~ saw ~~a~~ traffic ~~in~~ accident front
car damage ~~on~~ route 1 ……

Stemming

Stemmed tokens:
see traffic accid route ……
saw see accident accidents
traffic route……

Each tweet T_i can be summarized with a set of stemmed **tokens**.

Topic 1 :Inferring Traffic Condition during Inclement Weather using Social Media Data

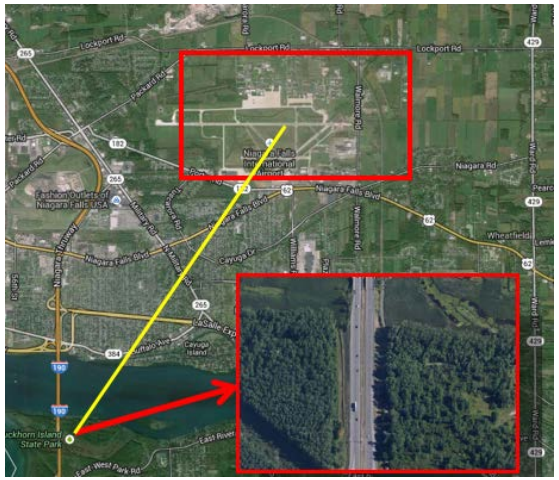
Aims:

- 1. Examine if real world weather events can be inferred from social media data, and
- 2. Determine whether including weather variables, *extracted from social media data*, can improve the predictive accuracy of models developed to quantify the impact of inclement weather on freeway traffic speed.

Study Locations and Datasets

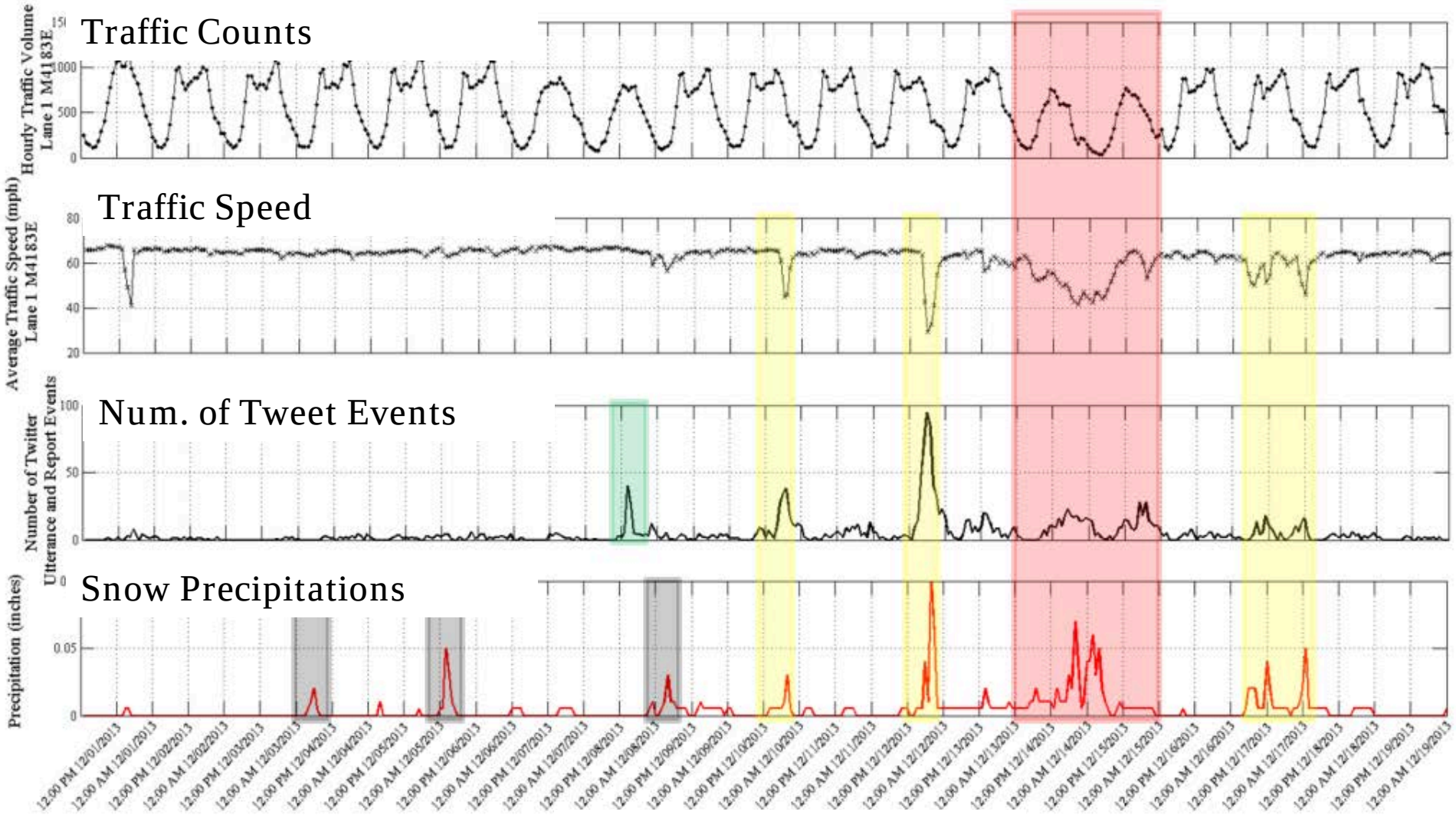
- Weather stations near two Airports: Buffalo Airport (BFIA) and Niagara Falls Airport (NFIA)
- A total of 360,112 tweets with longitude and latitude located within the Buffalo area, and within the time interval from 12/01/2013 to 12/19/2013
- Traffic hourly counts and speeds data for each lane

BFIA



NFIA

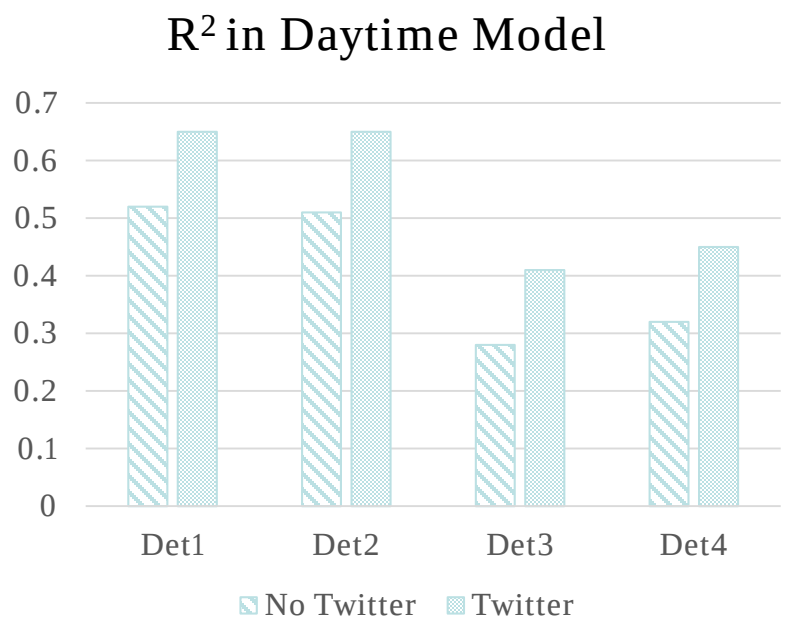
Tweet Events Indicating Weather Events

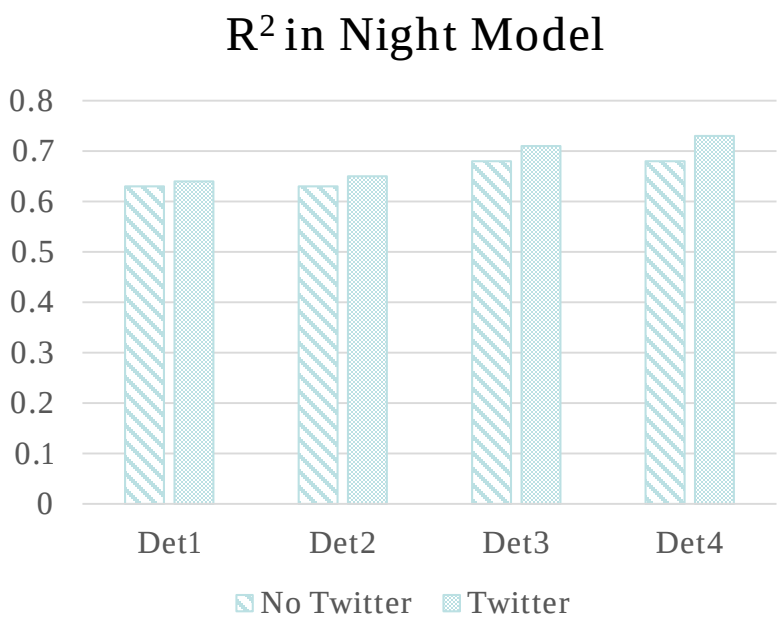


Time 12/01/13 - 12/19-13

Regression Results: Comparing R-Squared

Target variable: Traffic speed on freeway detectors Det1-Det4
Input variables: weather data + twitter data



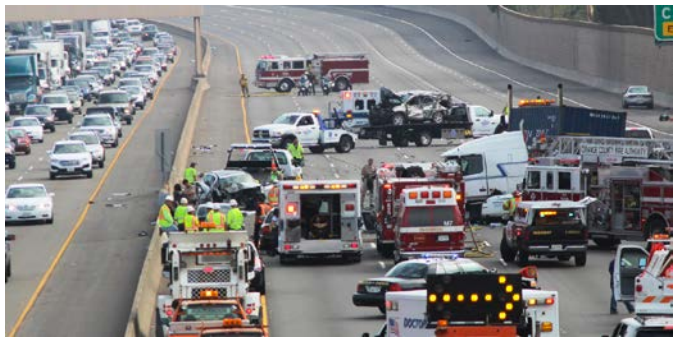


More improvement observed in Daytime model

Topic 2: Traffic Accident Detection with Social Media and Traffic Loop Detectors

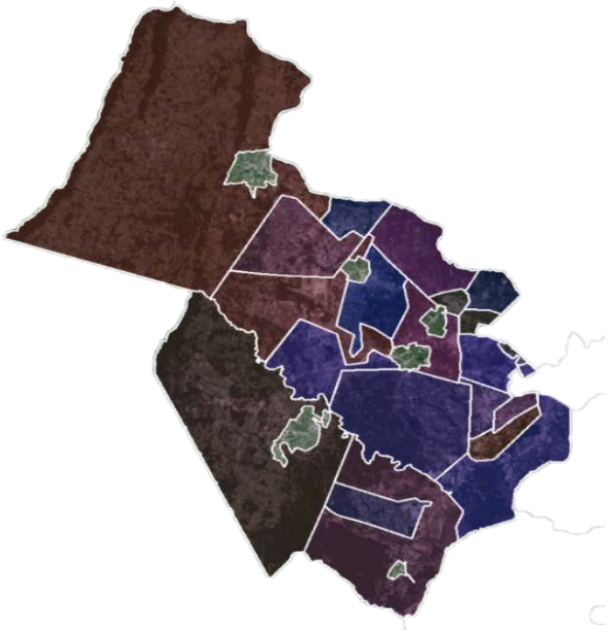
Aims

1. Extract social media data related to traffic accidents
2. Use social media to detect a traffic accident.



Data Description at Northern Virginia (NOVA)

- Time: 12 months in 2014
- Social Media: 584,000 tweets
- Traffic: 15,000 loop detector data from, reporting volume, speed, and occupancy in 15 min interval
- VDOT Accidents: 10,000 traffic incidents



Keyword-based Filtering- Example for “Traffic Accidents”

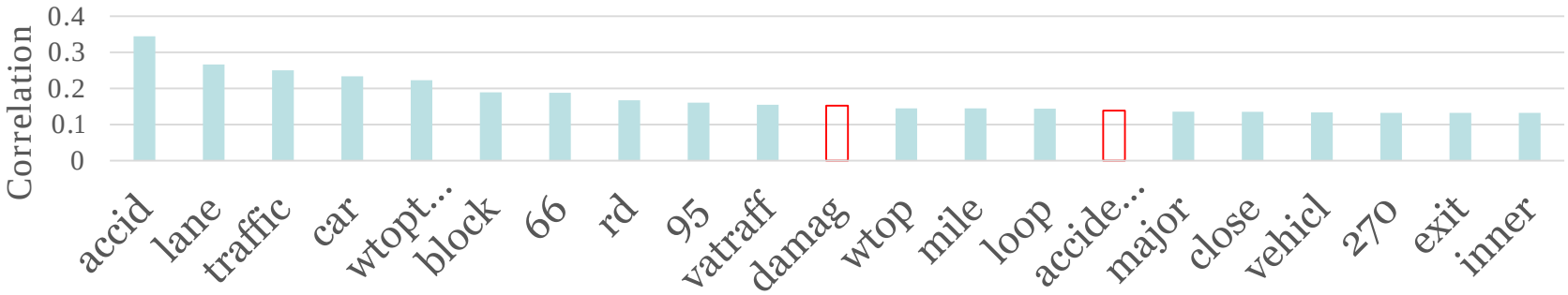
- We first examine 100 articles of news about traffic accidents. In all these articles, we select the words that appear the most frequently, except for common words.

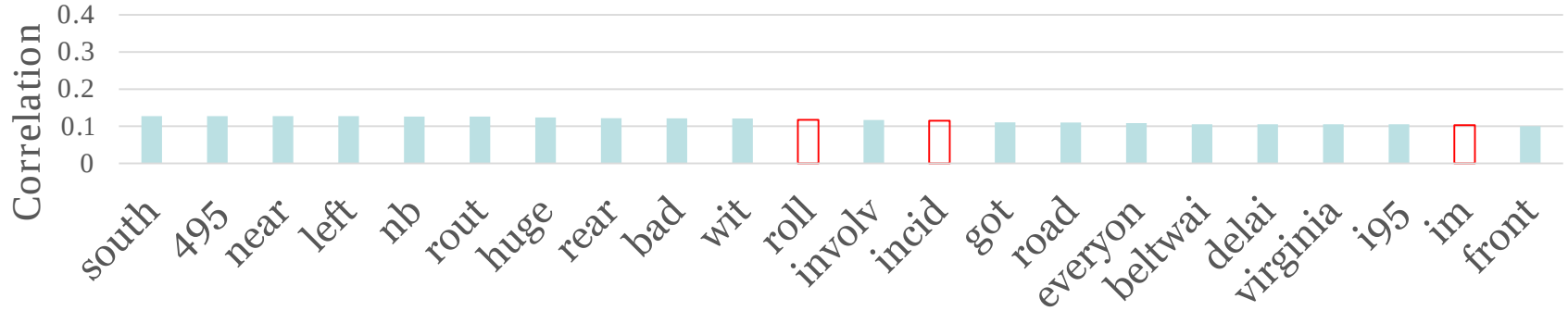
“accident”, “incident”, “crash”, “collision”, “head on”, “damage”, “pile up”, “rear end”, “rear-end”, “sideswipe”, “lost control”, “rolled over”, “roll over”, “tailgating”, “police”

- We obtained filter tweets as inputs for training the accident identification model.
- Each tweet is manually labeled as “Accident” or “Non-accident”.

Token-based Method: individual tokens

27 individual tokens whose correlation $|\phi|$ is higher than 0.1 : notice that the “accid” refers to “accident”; “accident” refers to “accidently”

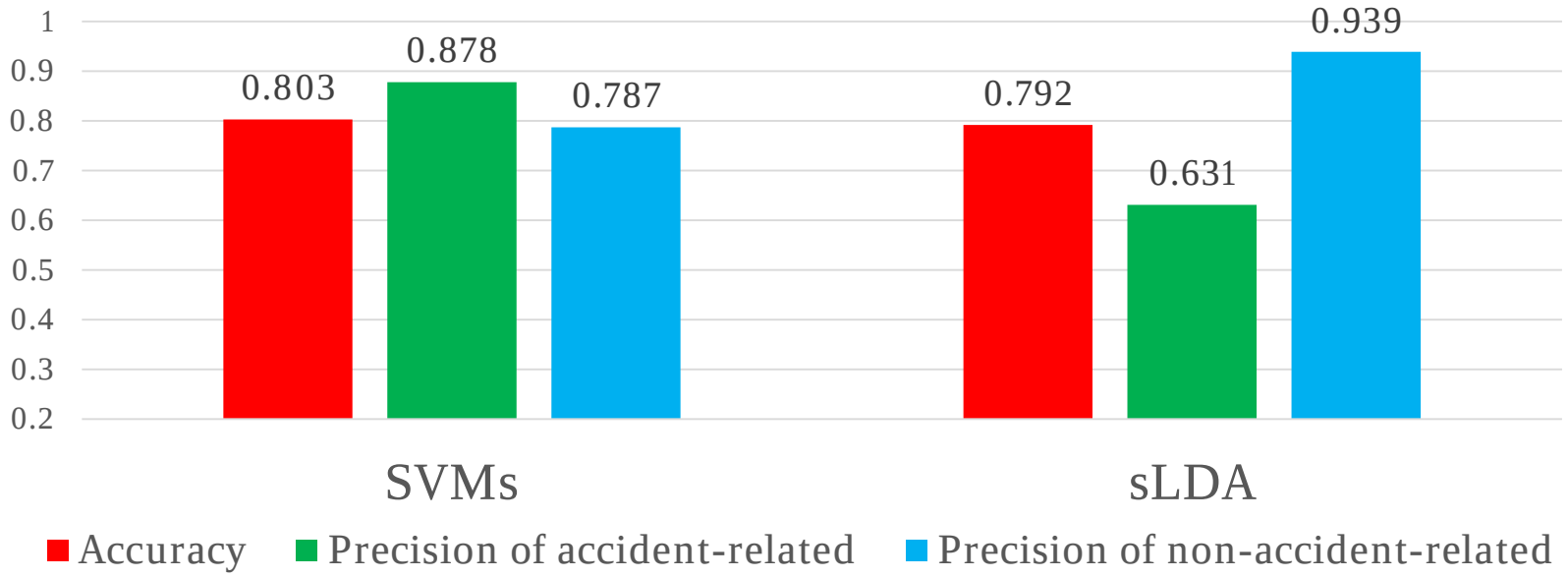




Positive correlation Negative correlation

Paired tokens VS Topic Modeling

- Compare the results from sLDA, a popular topic modeling method.
- As compared, the paired tokens can have competing results.



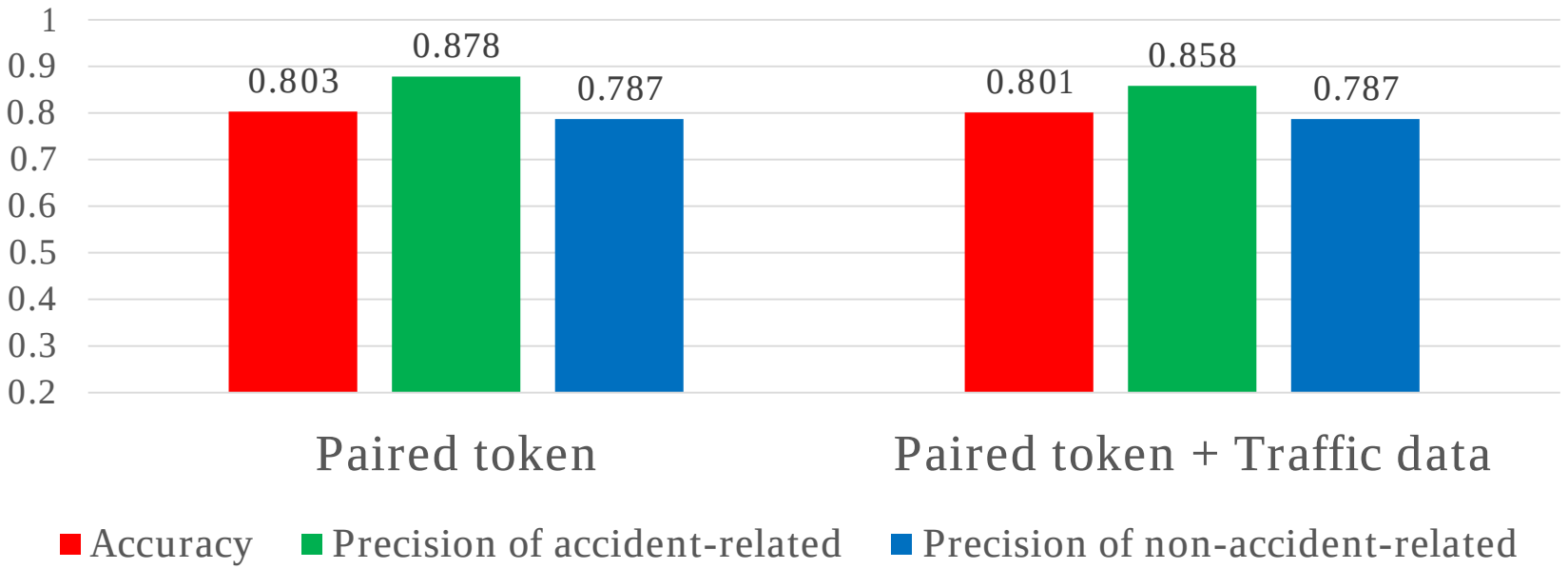
sLDA has higher False Positive.

Results with/without traffic data

- Traffic-related features do not necessarily improve the prediction results.

Without traffic-related features

With traffic-related features

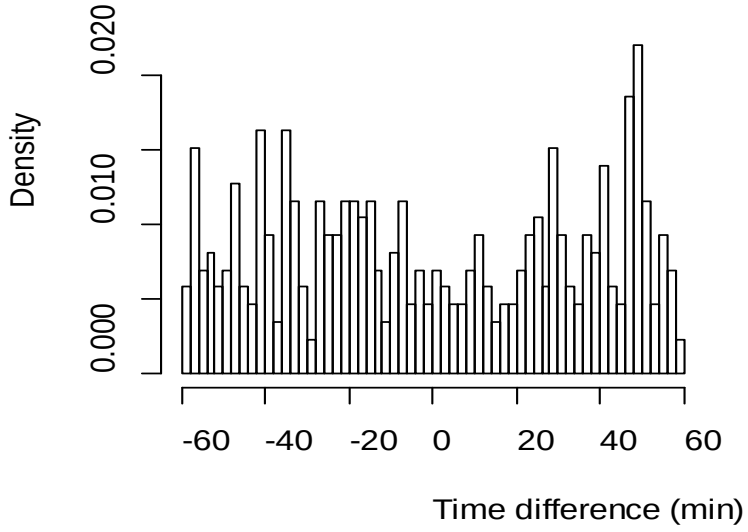


Possible reasons:

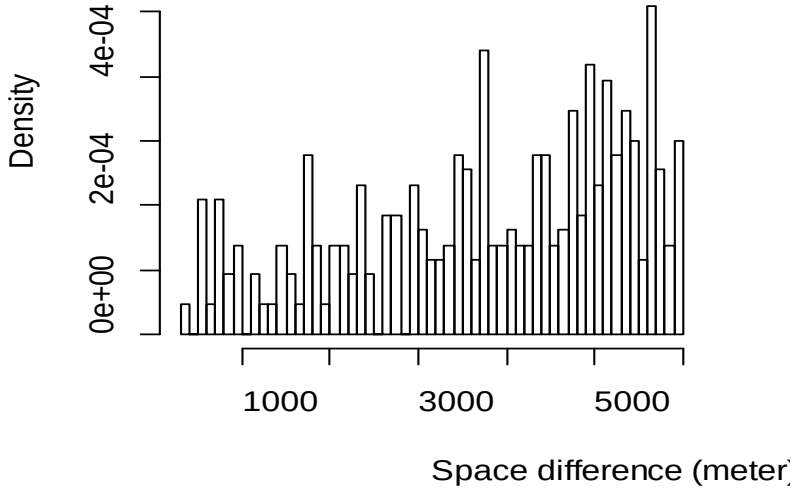
- Accidents may not generate impact on surrounding traffic, or
- The locations of the tweets do not match that of the accident sites.

Validation with DOT accident data

- Distributions of (a) time and (b) space difference between the accident-related tweets and the true accident records by VDOT.



(a)



(b)

About 300 of 422 accident-related tweets can be traced to an accident record by VDOT.