Analysis of Tweets to Identify, Predict and Visualize Accidents

Lei Li, Xiao Du, Umesh Nair

1. Introduction

Over the last few years, with the proliferation of smartphones and tablets in the market, almost everyone has access to mobile devices that offer better processing capabilities and access to new information and services along with easy connectivity to the internet. The Web is undoubtedly the best place for finding and sharing content, especially through social media. One such social media platform is Twitter, which provides a very convenient way to share one's thoughts and opinions with the world, and has now become ubiquitous.

While social media has its fair share of negative aspects, it is also very useful because it makes all types of information readily available. People now have a tendency to tweet every small little interesting or unusual thing, which is used in the right manner, will be useful in many applications. One such application is gathering information about various traffic events and accidents. There have many research work done on social media and Twitter analysis to interpret and detect traffic events [1][2]. We plan to implement some of these methods and focus on accidents, and take these works forward by following up these analyses with descriptive and interactive visualizations.

In this project, we intend to analyze tweets to understand and identify accident spots, and visualize them along with important aspects about the tweets, and for research purposes, try and figure out if there is any relation between the sentiment score of a tweet and the seriousness of the accident.

2. One-sentence Description

Analysis of Tweets to identify, predict and visualize accidents

3. Project Type

Natural Language Processing,
Prediction, Interactive Visualizations

4. Audience

The visualizations would be useful for drivers, people travelling or planning to travel by road, to understand if there have been any accidents on their route based on real-time tweets, or be aware of accident-prone areas nearby where they should be careful, identified based on tweets collected over time. These visualizations could also be useful for traffic and road safety authorities in a particular region to continuously monitor accidents and traffic incidents, and to take precautionary measures in areas

identified as being highly prone to accidents.

The NLP and text analysis part of this project could be useful for researchers interested in knowing more about topic modeling and if there is any relation between the sentiment score of a tweet and the seriousness of an incident/accident.

5. Approach

5.1 Details

In this project, we focus on finding relationship between tweets and accidents by sentimental analysis.

First, we will apply NLP on the tweets to transform it into computable numbers and vectors for further analysis. In order to reduce the unnecessary waste on computing power and increase the accuracy of the final result, we apply a filter to the data set for the data preprocessing, which is removing the stop words. Stop words are commonly used words such as “the”, “a”, “an”, etc. The stop words package we will be used from NLTK.

Then, we will focus on L-LDA model. Labeled LDA is a probabilistic graphical model that describes a process for generating a labeled document collection. Labeled LDA models each document as a mixture of underlying topics and generates each word from one topic. In this case, we plan to summarize each tweet to a topic and link them to other similar topics, to discover is there any hidden accidents. Besides that, we plan to analysis if there are any relationship between sentimental score and seriousness of accident for those tweet mention any serious word.

As for evaluation part, we plan to crawl on the website to find news related to accidents. Comparing to these ‘truth’ to the results from L-LDA model, we will end up with the relationship between analyzing tweets and accidents.

At last, when it comes to visualization, we plan to use D3 to display the heat map for traffic accidents in USA using the model. In each accident location, it will display the word cloud with daily accident frequency analysis.

5.2 Evidence for Success

It will reveal some hidden accidents from tweets after comparing to the news. What’s more, using sentiment analysis, we can find certain way to display seriousness of accidents. In this way, users can have a direct impression of distribution of accidents. They can also try to prevent some future accidents from historical patterns.

6. Best-case Impact Statement

In the best-case scenario, for data analysis part, we can discover some hidden relationship between sentimental score and seriousness of accident for those tweet mention any serious word. What’s more, using tweets for predicting accidents can have a better performance when comparing to the news.

For visualization, linking different views of plots successfully can be important.Implementing map, word-cloud and line chart in appropriate location to show the results.

7. Major Milestones

- Successfully processing the tweets data and implement L-LDA model to discover hidden relationship with accidents, sentimental score and seriousness of accident.
- Adjusting parameters of Model according to testing results with found news.
- Successfully display predicted and true accidents in heat map, and using word-cloud with line chart for daily accident frequency analysis.
- If we have time, we will try to implement this in real-time.

8. Obstacles

8.1 Major obstacles

- Twitter data acquisition. Since we are using rest API of twitter. How to extract useful tweet content effective cleaning and processing the data within tolerable time (depends on browser cache and application processing speed).
- Designing our own stop words and unrelated words to filter and delete unrelated twitters.
- Implement real-time data. Since the purpose of this project is monitoring traffic, the data should be streaming data.

8.2 Minor obstacles

- Designing the multiple linked views, includes accident tweet word cloud(depends on the sentimental score, positive or negative), Daily accident frequency analysis, accident location using geomap API, heat map for traffic accident.
- UI design, makes it easy, effective and convenient for user to monitor the accident.

9. Resources Needed

Flask framework, Twitter REST API, AWS, Python

10. 5 Related Publications

- Marco Bonzanini post a simple approach for sentiment analysis which is intuitive and simple to understand, test, and most of all unsupervised so it doesn't required any labelled data for training[3].
- Eleonora D'Andrea post a real-time detection of traffic which is a real-time monitoring system for traffic event detection from Twitter stream analysis. The system design for streaming data could be different with offline batch data[4].
- Jonas Traub presents I^2, an interactive development environment that coordinates running cluster applications and corresponding visualizations such that only the currently depicted data points are processed and transferred. Besides that, they presents an algorithm for the real-time visualization of time series[5]. We would like to try this method in our project.
- While developing a web portal the appearance of web portal

makes a development more critical. The good appearance of a web can easily attract more number of visitors which is a success of web portal. This paper introduces the basic instruction of using Flask to build web application[6].

- Vital Design present an approach to filter twitter stream by split stream.

11. Define Success

A web application for accident monitoring in the USA. User can identify the specific spots, times and seriousness of accidents easily and effectively.

References:

[1] F.C. Albuquerque et al. , "A methodology for traffic-related Twitter messages interpretation", *Computers in Industry*, Volume 78, May 2016, pp. 57-69

[2] Y. Gu et al. , "From Twitter to detector: Real-time traffic incident detection using social media data", *Transportation Research*, Part C 67 (2016) pp. 321–342

[3] Marco Bonzanini. , “Mastering Social Media Mining with Python”, Chapter 6, July, 2016

[4] Eleonora D'Andrea. , “Real-Time Detection of Traffic From Twitter Stream Analysis”, IEEE Transactions on Intelligent Transportation Systems, Volume: 16 , Issue: 4 , Aug. 2015

[5] Jonas Traub, Nikolaas Steenbergen, “I2 : Interactive Real-Time Visualization for Streaming Data”

[6] Fankar Armash Aslam, “Efficient Way Of Web Development Using Python And Flask”, Volume 6, No. 2, March-April 2015 International Journal of Advanced Research in Computer Science