Final Project Report

Real-Time Transportation Social Media Analytics Using Pulse (Pulse-T)

Prepared for **Teaching Old Models New Tricks (TOMNET) Transportation Center**

By

Srinivasa S. Kandala
Email: srivatsav.kandala@asu.edu

Vikash Bajaj
Email: vbajaj2@asu.edu

Decision Theater Network
Arizona State University
Tempe, AZ

August 2021

TECHNICAL REPORT DOCUMENTATION PAGE

1. Report No. N/A	**2. Government Accession No.** N/A	**3. Recipient's Catalog No.** N/A	
4. Title and Subtitle Real-Time Transportation Social Media Analytics Using Pulse (Pulse-T)		**5. Report Date** August 2021	
		6. Performing Organization Code N/A	
7. Author(s) Srinivasa S. Kandala, https://orcid.org/0000-0002-9723-7658 Vikash Bajaj		**8. Performing Organization Report No.** N/A	
9. Performing Organization Name and Address Decision Theater Network Arizona State University 660 S. College Avenue, Tempe, AZ 85287-3005		**10. Work Unit No. (TRAIS)** N/A	
		11. Contract or Grant No. 69A3551747116	
12. Sponsoring Agency Name and Address U.S. Department of Transportation, University Transportation Centers Program, 1200 New Jersey Ave, SE, Washington, DC 20590		**13. Type of Report and Period Covered** Research Report (2020 – 2021)	
		14. Sponsoring Agency Code USDOT OST-R	
15. Supplementary Notes N/A			
16. Abstract As city planners and transportation system planners consider changes and upgrades to transportation systems and infrastructure, they require models that accurately reflect communities' needs. Planners need access to advanced activity-travel demand analysis models that are responsive and sensitive to emerging transportation technologies; models are needed that not only provide insights into communities' current travel demands and behaviors, but also help understand people's attitudes and expectations toward a change — or a proposed change — in a community's transportation infrastructure or transportation options. These models often rely on data collected from surveys or opinion polls. Surveys result in regimented answers to specific questions, only measure behavior at a single-point in time, and are often vulnerable to self-selection bias. These limitations prevent attitudes from being observed across the population and across time, thus hindering the ability to do real-time analysis and get the most accurate and recent public sentiment to inform policy decisions. In this project, Arizona State University's Decision Theater (DT) staff built PULSE-T to address these limitations. PULSE-T gathers data directly from Twitter, allowing for information to be gathered live about the impact of a policy change. Twitter offers a large volume of publicly available data in which people and groups broadcast their feelings and preferences far more widely than what a survey instrument could capture			
17. Key Words Responsive Modeling; Public Sentiment Analysis; Technology-Driven Data Collection		**18. Distribution Statement** No restrictions.	
19. Security Classif.(of this report) Unclassified	**20. Security Classif.(of this page)** Unclassified	**21. No. of Pages** 22	**22. Price** N/A

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

DISCLAIMER

The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the information presented herein. This document is disseminated in the interest of information exchange. The report is funded, partially or entirely, by a grant from the U.S. Department of Transportation's University Transportation Centers Program. However, the U.S. Government assumes no liability for the contents or use thereof.

ACKNOWLEDGMENTS

This study was funded by a grant from A USDOT Tier 1 University Transportation Center, supported by USDOT through the University Transportation Centers program. The authors would like to thank the TOMNET and USDOT for their support of university-based research in transportation, and especially for the funding provided in support of this project.

TABLE OF CONTENTS

LIST OF TABLES

No table of figures entries found.

LIST OF FIGURES

EXECUTIVE SUMMARY

As city planners and transportation system planners consider changes and upgrades to transportation systems and infrastructure, they require models that accurately reflect communities' needs. Planners need access to advanced activity-travel demand analysis models that are responsive and sensitive to emerging transportation technologies; models are needed that not only provide insights into communities' current travel demands and behaviors, but also help understand people's attitudes and expectations toward a change — or a proposed change — in a community's transportation infrastructure or transportation options. These models often rely on data collected from surveys or opinion polls. Surveys result in regimented answers to specific questions, only measure behavior at a single-point in time, and are often vulnerable to self-selection bias. These limitations prevent attitudes from being observed across the population and across time, thus hindering the ability to do real-time analysis and get the most accurate and recent public sentiment to inform policy decisions. In this project, Arizona State University's Decision Theater (DT) staff built PULSE-T to address these limitations. PULSE-T gathers data directly from Twitter, allowing for information to be gathered live about the impact of a policy change. Twitter offers a large volume of publicly available data in which people and groups broadcast their feelings and preferences far more widely than what a survey instrument could capture.

INTRODUCTION

As city planners and transportation system planners consider changes and upgrades to transportation systems and infrastructure, they require models that accurately reflect communities' needs. Planners need access to advanced activity-travel demand analysis models that are responsive and sensitive to emerging transportation technologies; models are needed that not only provide insights into communities' current travel demands and behaviors, but also help understand people's attitudes and expectations toward a change — or a proposed change — in a community's transportation infrastructure or transportation options.

These models often rely on data collected from surveys or opinion polls. Surveys result in regimented answers to specific questions, only measure behavior at a single-point in time, and are often vulnerable to self-selection bias. These limitations prevent attitudes from being observed across the population and across time, thus hindering the ability to do real-time analysis and get the most accurate and recent public sentiment to inform policy decisions.

In this project, Arizona State University's Decision Theater (DT) staff built PULSE-T to address these limitations. PULSE-T gathers data directly from Twitter, allowing for information to be gathered live about the impact of a policy change. Twitter offers a large volume of publicly available data in which people and groups broadcast their feelings and preferences far more widely than what a survey instrument could capture [1].

TOMNET (Center for Teaching Old Models New Tricks) is an Arizona State University center dedicated to understanding and modeling human mobility choices under a wide variety of conditions with a view to explicitly incorporate attitudes, values, perceptions, and lifestyle preferences in activity-travel demand forecasting models. PULSE-T exponentially expanded the access of TOMNET researchers and other organizations to an up-to-date, filtered dataset of public opinion and discussions around virtually any transportation research area. Researchers and organizations have user perceptions on transport demand at their fingertips, enabling them to take appropriate measures and actions and undertake planning projects much more effectively than was possible with previous models.

With successful analysis of social media data guiding outreach strategies that are useful for TOMNET, transport planning and necessary interventions can be done at the right time.

Objective 1: To increase TOMNET's visibility through a public-facing dashboard that provides real-time access to users' reactions or opinions about specific transportation policies or topics.

Objective 2: To give TOMNET researchers the ability to mine large amounts of data on different topics through filtering by keywords, location, user accounts, geographical attributes, and language of search. Researchers will be able to access topic-specific data and analyze it through an interactive dashboard.

Objective 3: To provide TOMNET researchers with crowd- sourced, location-specific enriched data on residents' attitudes, values, perceptions, and preferences around transportation options or changes, which can be employed in current and future research.

TECHNOLOGY USED

Keeping with the spirit of the TOMNET Center, and Teaching Old Models New Tricks, the foundation of PULSE-T was PULSE, a platform previously created by ASU's Decision Theater. PULSE was designed to search through social media posts, journal publications, databases, and more to capture key information and their implications, perform a variety of analyses, and visualize the results (Figure 1).

Users identify key terms, search parameters, and research objectives, then Decision Theater researchers help to determine the mechanisms for analysis and implement the appropriate algorithms to gather and sort information. These techniques uncover granular insights while filtering out irrelevant data.

Information is then presented on a custom-developed graphic dashboard and Decision Theater researchers work with users to identify appropriate visualization techniques to communicate new insights.

For PULSE-T this technology was used to specifically address questions about transportation policy. Decision Theater and TOMNET researchers collaborated to identify the appropriate key terms and research questions, as well as appropriate filtering techniques to ensure the final visualization only included relevant information.

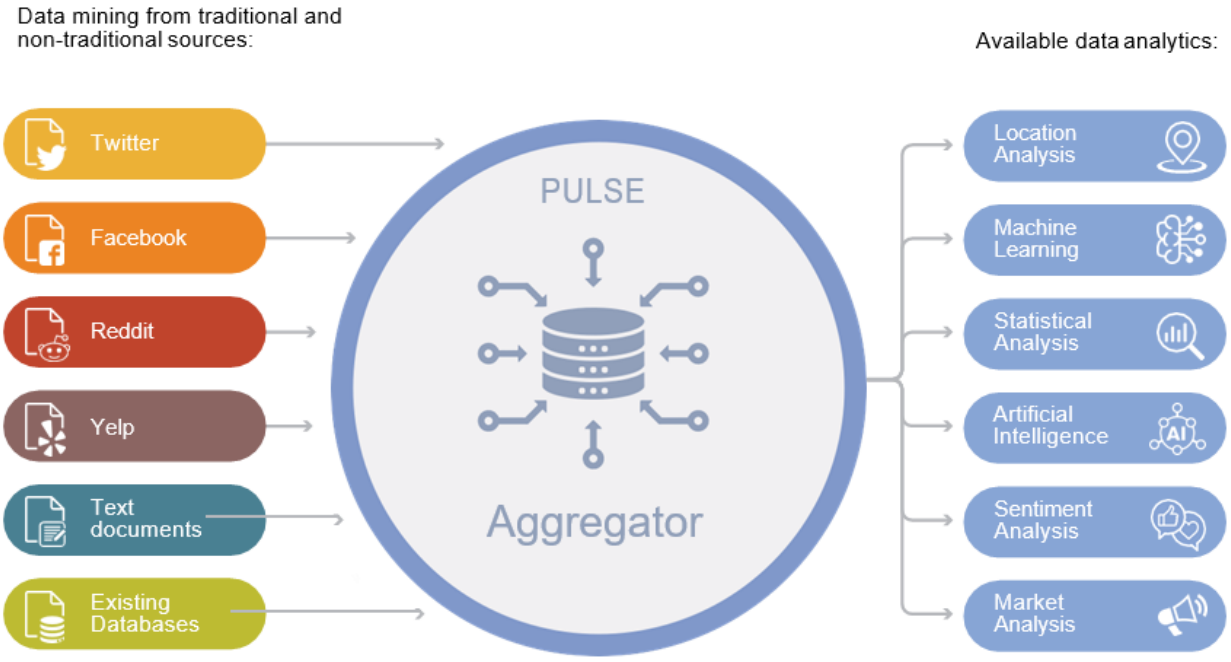

Figure 1 PULSE platform

METHODOLOGY

Step 1: Storage and processing. DT staff used distributed systems to collect data through Twitter streaming API and store it with its metadata in the database. The tweets were then processed and important attributes such as languages, hashtags, and URLs were stored. Relevant information from these stored attributes is later used for sentiment analysis, network analysis, and displaying visualizations.

Step 2: Extraction of sentiment. Sentiment analysis is a technique which seeks to identify the viewpoint(s) underlying a text span. Sentiment analysis is at the heart of PULSE. The model was built using the open source models DeepMoji [2] and Unsupervised Sentiment Neuron [3]. The DeepMoji model is used to predict emojis; this model is trained from 1.2B tweets that are filtered from 55B tweets. DT staff extended the model to extract sentiment and emotion and to detect sarcasm. Unsupervised Sentiment Neuron is trained to predict the next letter in Amazon reviews, and its makers assert that it has 91.8% accuracy in detecting sentiment.

Step 3: Network analysis. Analyzing data from all users is not necessary to find prominent conversational issues and extract meaningful sentiment [4]. DT staff used network analysis to understand influential Twitter users' statements, predictions, opinions, and impact. Picking the highly active users gave information about the most-talked-about issues in the Twittersphere, helping DT to access voices related to the topic of interest, i.e., transportation in the Twittersphere. First, DT made use of specific metrics to capture prominent users. Then, they analyzed the sentiment of tweets and the sentiment of the reactions to the influential tweets from prominent users and depicted the results in a multidimensional data representation [5].

Step 4: Indexing results. Lastly, DT staff indexed the data and results on a fully indexed distributed database. Usage of a distributed database provides high-speed query, search, and aggregation capabilities by serving parts of queries from the nodes in the distributed system along with providing redundancy in case of downtime [6].

Step 5: Development of dashboard. DT staff then built the user-facing real-time dashboard that includes data analytics, sentiment and network analysis, visualizations, and charts.

USE CASE: OVERVIEW

PULSE-T sought to look into public sentiment surrounding autonomous vehicles. The key terms identified for PULSE-T were: Autonomousvehicle, DriverlessTechnology, AVSafety, AVCrash, AVCollision, AVAdoption, AVPublicPerception, AVReliability, Waymo, Tesla, autopilot, selfdriving, AVOwnership. These keywords have been retrieved from a literature survey of research papers related to different aspects of autonomous vehicles including Crash, Safety, Market & Sales, Automobile Industry, Transportation, Perception.

Using these keywords and the steps outlined above, the following dashboard was produced.

Figure 2 Overview of PULSE dashboard

USE CASE: ELEMENTS

This section will provide a more in-depth tool walkthrough of the elements on the dashboard.

Filters Tab: The purpose of this tab is to see which filters have been selected. If there are no filters selected yet, nothing appears in this section. Filters can be added or removed as needed.

Tweet Feed: This section displays the tweets that have been retrieved based on the selected keywords. There is also a drop box that has three options for which type of tweets to show: those with just links, those containing photos, or those containing videos.

Time Series: The Time Series graph helps visualize data points at successive time intervals. Every data point on the chart has a corresponding value representing the time on the X-axis and a corresponding quantity representing the total number of tweets at that time on the Y-axis.

The Zoom feature allows the user to choose a particular time frame ranging from 1 hour, 1 day or 1 week. The user can also select a specific time frame by dragging a window on the time series, which will filter the entire dashboard.

Once a particular time frame is chosen, it appears in the filters section as well, and all the sections of the dashboard will be updated accordingly. The Time Series chart can be downloaded as an image or in other various formats and can be printed using the button on the top right corner of the section.

This chart helps identify trends in the number of tweets regarding autonomous vehicles at different times of the day and can be compared with the data of previous years as well to see if there is any trend with respect to a season or month.

The graph below represents the Time Series data from Feb 1 to Feb 14. The time represented is in Coordinated Universal Time (UTC). It can be observed that the graph follows a cyclic pattern with the number of tweets peaking at certain times and then decreasing.

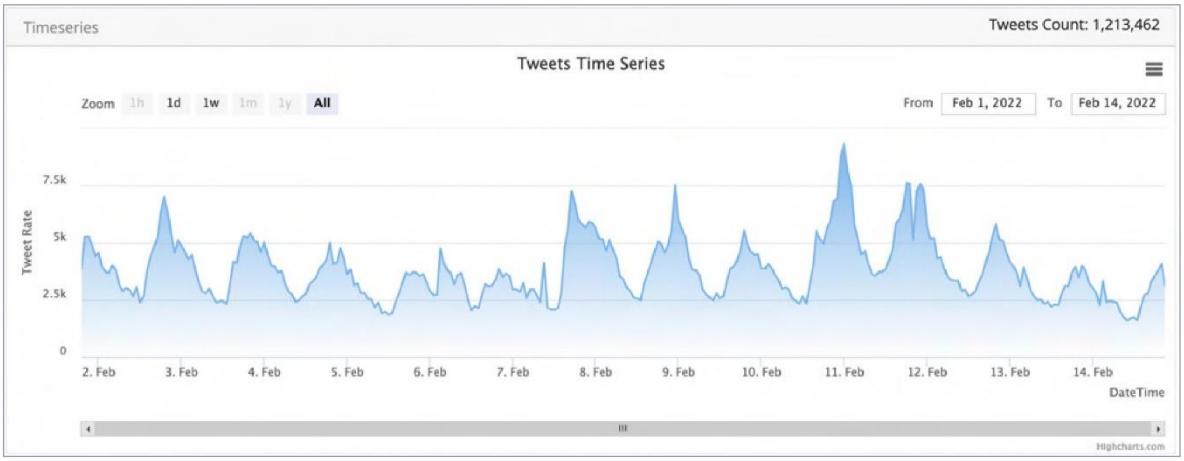

Figure 3 Tweets time series

USE CASE: BAR CHARTS

Vertical bar charts are used here to represent the categorical data, where the height of the bar on the Y-axis corresponds to the quantity of each data point on the X-axis. The bar charts are based on six parameters. Each bar chart gives the count of each parameter in the entire twitter dataset within a particular time frame. The bar chart changes dynamically corresponding to the filters applied. Each of these bars can be clicked individually to filter the data based on the selected data point. Upon choosing a particular value, the entire dashboard results will also be updated.

The six bar charts with the corresponding categorical variable are as follows:

Hashtags

A hashtag with '#' at the beginning of a word links a particular tweet where this hashtag has been used to all the other tweets with similar hashtags. '#' on Twitter is used to index key words or topics which allow people to follow topics of their interest.

This bar chart displays the most frequent hashtags that are related to autonomous vehicle tweets, extracted from Twitter. This can help identify the most popular hashtags to further filter the tweets in the future.

The bar chart shown below represents the count of each hashtag on the X-axis. It can be inferred from the below graph, for the entire data ranging from Feb 1 to Feb 14, that the following are the most frequent.

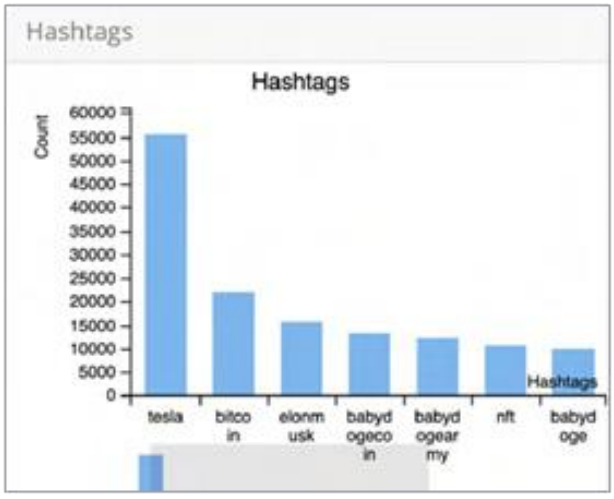

Figure 4 Tweet distribution by hashtag

Hashtags: Tesla, bitcoin, elonmusk, babydogecoin, babydogearmy, nft, babydoge, crypto, nftgiveaway, spacex, nfts

Tesla – Tesla is the most popular hashtag as can be seen from the bar chart. Tesla is an American manufacturing firm that manufactures electric vehicles, solar panels and other products.

Elon Musk – Elon Musk is the founder of SpaceX, Tesla and a few other companies that manufacture autonomous cars and space transportation services.

SpaceX – SpaceX is the manufacturer of space communication and transportation services including autonomous spacecrafts, drones and autonomous space landing for the future.

Dogecoin – Dogecoin is an open-source cryptocurrency started in 2013 by Jackson Palmer and Billy Markus. Dogecoin appears in autonomous vehicle tweets because there is speculation it could be accepted as a payment method by Tesla.

NFT – A NFT, or non-fungible token, is a unique unit of data that uses technology to log and authenticate digital content, such as films, songs, and photographs, on cryptocurrency blockchains, primarily Ethereum.

Every transaction from transfers to sales is recorded on-chain once content is logged onto the blockchain, producing an easily accessible trail of provenance and pricing history. The most significant impact of NFTs is that they make it simple to own and sell digital content.

Baby Dogecoin – It is also a cryptocurrency but with 420 quadrillion coins in existence.

Upon selection of a particular hashtag, a filter is applied which can be seen in the filters section. All the sections of the dashboard are updated to give the time series analysis of tweets, mentions, users, and a word cloud of that particular hashtag.

For example, in the screenshot below when the mouse is hovered over Elon Musk and that particular hashtag is chosen, each section of the dashboard is changed. Selective charts are shown in Figure 5.

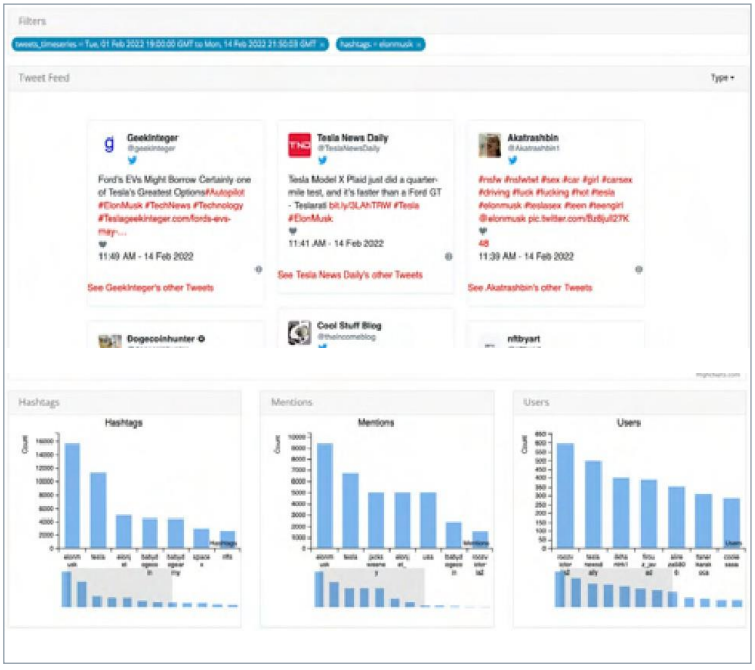

Figure 5 Updated dashboard (hashtag 'elonmusk' selected)

In the Filters section, the hashtag filter that has been applied can be seen and the tweet feed contains only those tweets that contain the hashtag ElonMusk.

Filtering in this manner can be done with any of the six bar charts.

Clicking a particular column in any of the charts filters the data by that parameter. This updates the analysis of the tweet data according to that filter.

Mentions

Mentions are those tweets which have another person's username anywhere in the tweet. In the tweets collected for this dashboard, it can be seen that Elon Musk and Tesla were mentioned the most.

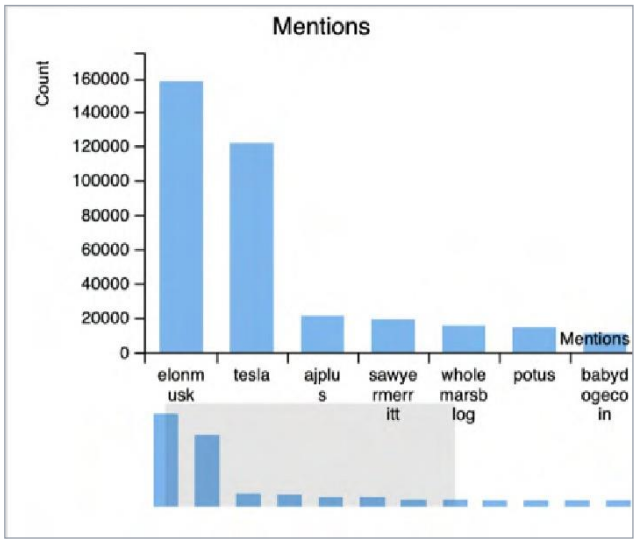

Figure 6 Tweet distribution by mention

Users

This bar chart shows the users who tweeted the most from the retrieved tweets. This is dynamic and changes based on the user's continued usage of Twitter.

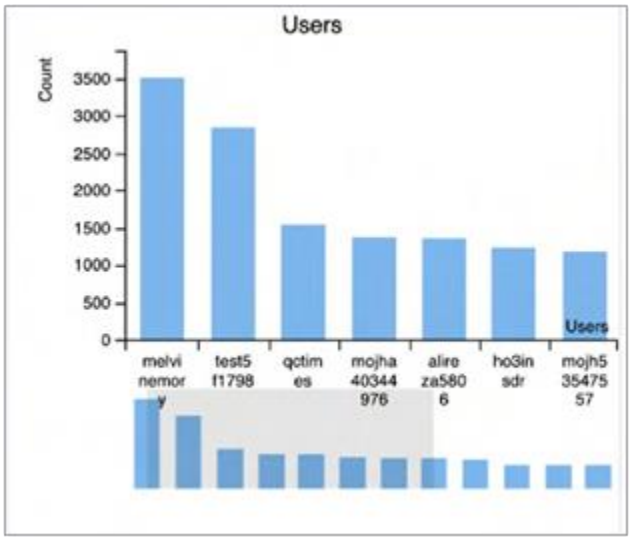

Figure 7 Tweet distribution by user

User Interface

Users use various interfaces to access Twitter. This bar chart shows which interfaces are used the most to tweet and which ones are used rarely. It is observed that the most used interfaces are Twitter for Android, Iphone, Twitter Web App, Social News Desk, Twitter for Ipad, and Wordpress. The least popular interfaces are Tweet Deck, dlvr, ifttt, testsforsearch, and buffer.

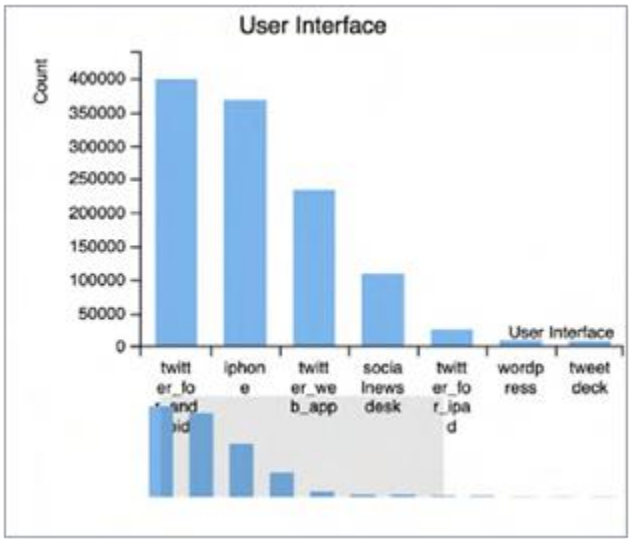

Figure 8 Tweet distribution by user interface

User Language

While the bar chart shows that English is the most popular language on Twitter, there are 34 other languages that Twitter supports. This source provides the full list of Twitter language abbreviations if needed for interpretation of the bar chart.

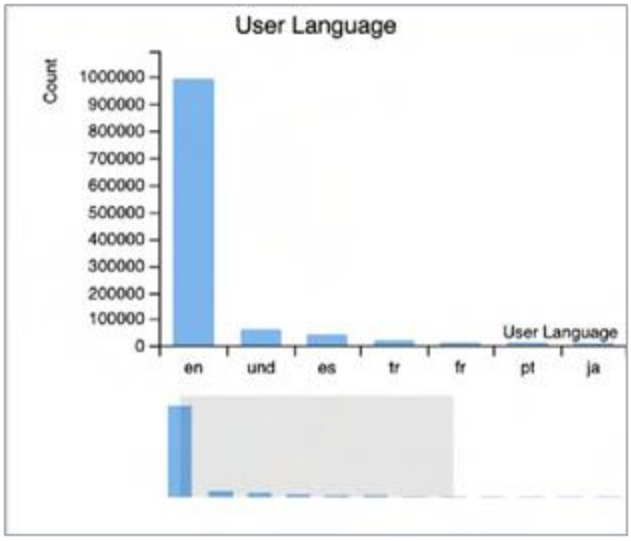

Figure 9 Tweet distribution by language

Hostname of URL
Twitter is the most used hostname.

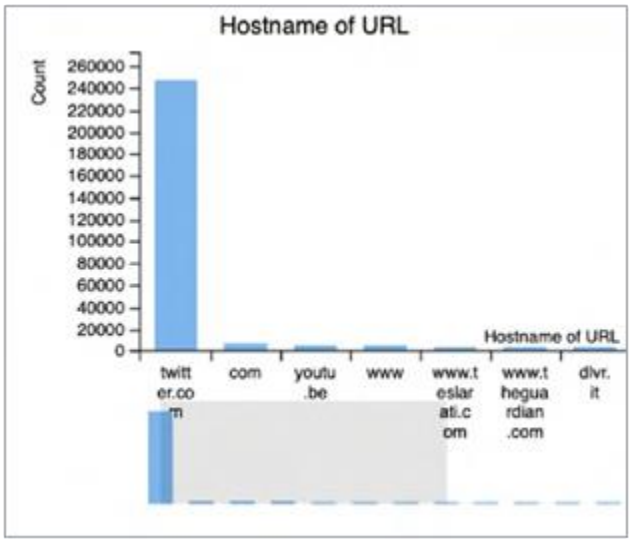

Figure 10 Tweet distribution by hostname

USE CASE: PIE CHARTS

This section has three pie charts. Each distribution in the pie chart can be selected separately as a parameter to filter the data.

Type of Tweet

This pie chart represents the following out of all the tweets that have been retrieved.

Tweet – The percentage of the tweets that are original tweets.

Reply – The percentage of replies to tweets.

Quote – The percentage of tweets which are quotes.

A quote is a retweet that allows a user to add their own comments to the tweet.

For example, upon choosing the "reply" section in the pie chart, the dashboard results and other graphs are updated as seen in selective charts in Figure 12.

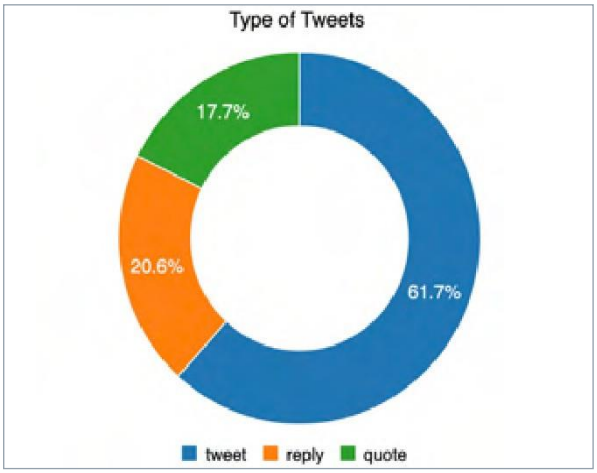

Figure 11 Tweet distribution by tweet type

The filter section now displays the tweet type that has been chosen as well and the results are produced accordingly. Filtering in this manner can be done on any of the three pie charts.

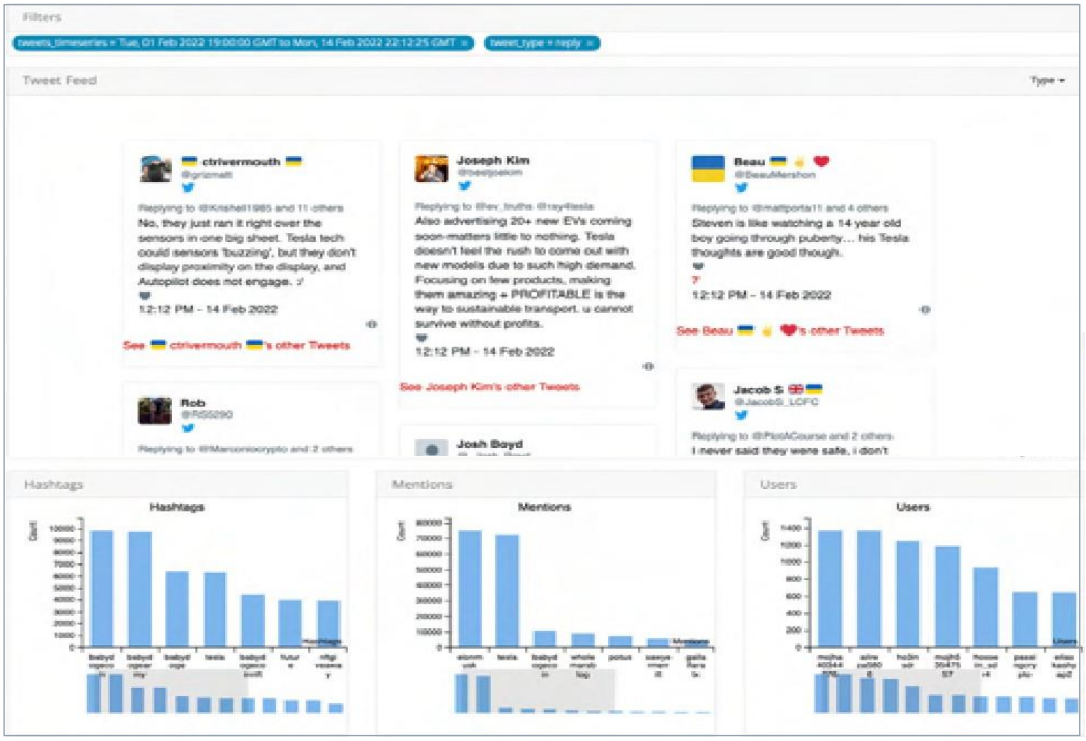

Figure 12 Updated dashboard (type 'reply' selected)

Tweet Summary

Upon application of the filters, Tweet Summary gives the percentage of the selected tweets, i.e. tweets that have been filtered out of the total tweets that have been retrieved. When there is no filter, the Tweet Summary is 100% as all the tweets are selected.

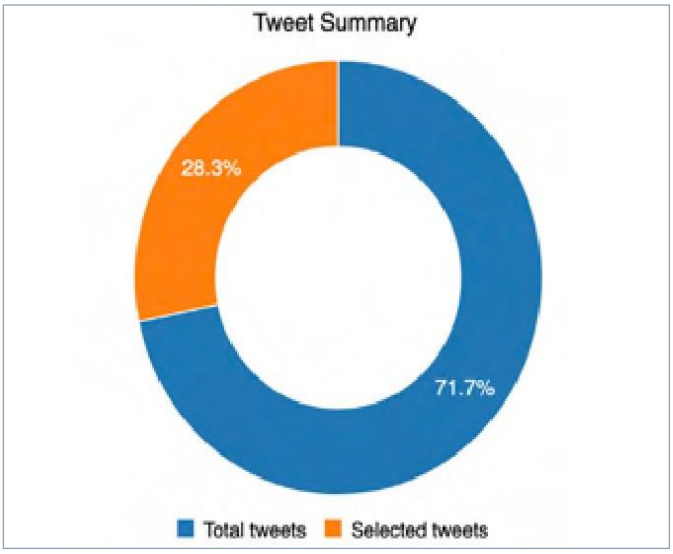

Figure 13 Tweet summary

Sentiments

The Sentiment Analysis results of the dataset are displayed in this pie chart. As mentioned in the Methodology section, algorithms have been used to identify the sentiment of the tweets.

Positive, Negative, and Neutral sentiments display the percentage of tweets that belong to each of these sentiments. If the tweet can't be clearly categorized into one of these using an algorithm, it is counted in the Confused category of the pie chart.

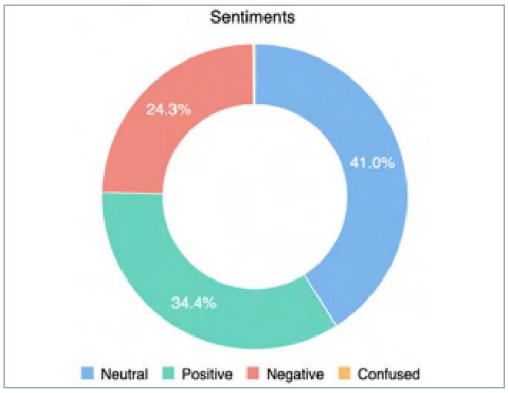

Figure 14 Tweet distribution by sentiment

USE CASE: MAPS AND CLOUDS

This section introduces three mapping features.

Network Map

A network diagram is a type of data visualization that allows users to quickly understand relationships in data. Nodes and edges make up a network diagram. Nodes are single data points with edges connecting them to other nodes. The relationship between two or more nodes is represented by edges. A network graph exposes trends and aids in anomaly detection. The Network Map updates to show data based on the filtering criteria.

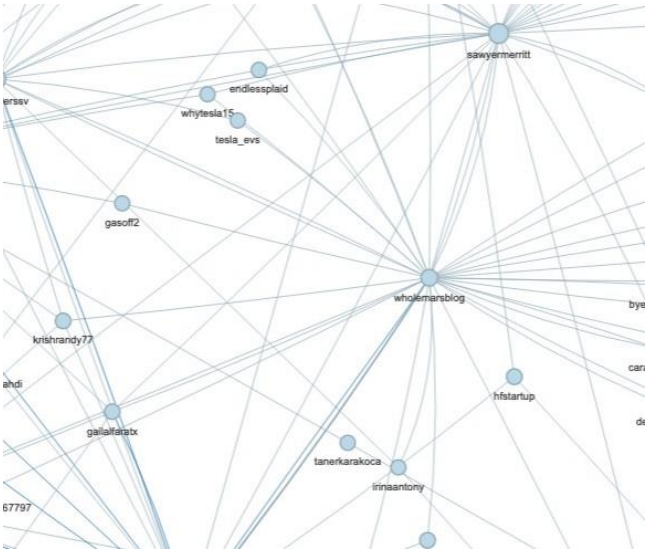

Figure 15 Network analysis and representation

World Map

Shows the location distribution of geo-coded tweets.

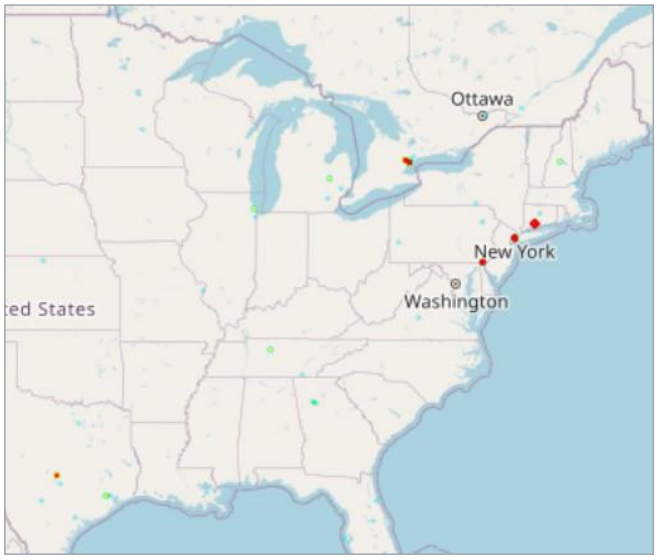

Figure 16 Geovisualization of geo-coded tweets

Word Cloud

A Word Cloud is a visualization that shows the most frequently used words in a text. The word size is larger for words that appear more frequently in the data and the size is smaller for words that appear less frequently. The Word Cloud updates to show data based on the filtering criteria.

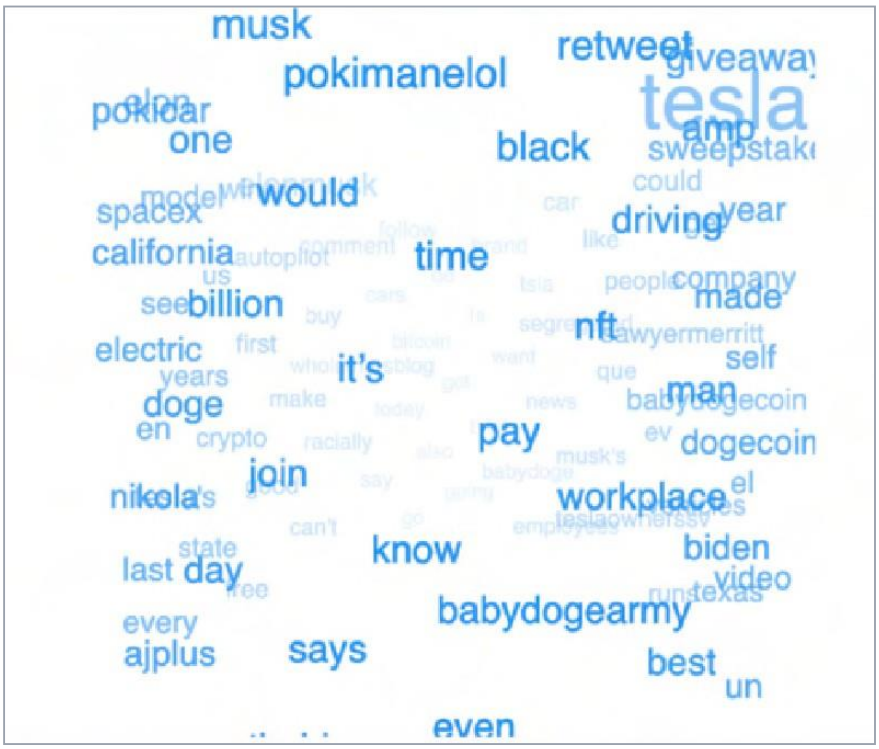

Figure 17 Word cloud

CONCLUSIONS

PULSE-T is a tool that was created to assist TOMNET researchers in gathering social media data. The tool gathers data directly from Twitter and provides real-time information, as well as sentiment analysis, network analysis, and other useful visualizations. The tool further allows researchersto download data in CSV format in order to use it for other research purposes. The successful accomplishment of the project objectives allows for TOMNET researchers to use the social media analysis insights for transport planning and interventions that are timely and based on significant and accurate public feedback.

REFERENCES

[1] Link, M., Muphy, J., Schober, M. F., Buskirk, T. D., Childs, J. H., Tesfaye, C. L., ... & Pasek, J. (2014). Social media in public opinion research.

[2] Felbo, B., Mislove, A., S gaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. arXiv preprint arXiv:1708.00524.

[3] Lengerich, B. J., Konam, S., Xing, E. P., Rosenthal, S., & Veloso, M. (2017). Visual explanations for convolutional neural networks via input resampling. arXiv preprint arXiv:1707.09641.

[4] Cha, M., Haddadi, H., Benevenuto, F., & Gummadi, P. K. (2010). Measuring user influence in twitter: The million-follower fallacy. Icwsm, 10(10-17), 30.

[5] Himelboim, I. (2017). Social Network Analysis (Social Media). The International Encyclopedia of Communication Research Methods, 1-15.

[6] Gormley, C., & Tong, Z. (2015). Elasticsearch: The Definitive Guide: A Distributed Real-Time Search and Analytics Engine. " O'Reilly Media, Inc.".