A Data-driven Approach to Study the Influence of Social Media on Human Behaviours in Transportation

Denis Asenov
d.asenov@student.utwente.nl
University of Twente
Enschede, The Netherlands

ABSTRACT

Estimating traffic conditions and accounting for accidents when deciding on a daily commute route is challenging. Modern technologies like GPS and navigation systems such as Google Maps[1] are commonly used, and with the rise of social media, many people also share traffic-related information online. This research focuses on X (formerly Twitter), one of the biggest social media platforms. Previous studies suggest that social media posts can accurately detect traffic events like incidents and congestion. This raises the question: do these influence users' routing decisions? This research analyses how traffic-related information from X can be used to understand its impact on travel decisions by comparing real-time traffic conditions with and without the presence of relevant social media posts to investigate a potential correlation, providing insights into whether this influence should be embraced or approached with caution.

KEYWORDS

Traffic analysis, navigation, social media, commuting patterns, social influence, driving behaviours, route planning, traffic estimation, real-time data

1 INTRODUCTION

Over the years, adaptation and integration of technology in everyday life have positively impacted many aspects of people's lives. Managing road trips is no exception, whether it's the daily commute to work or a longer journey. In the past, before the widespread availability of digital navigation tools, people used physical maps, which required extensive planning for the route and did not provide any information about the current traffic. The steady increase of cars on the road [5] means that the shortest path to the destination is often not the fastest one. Factors like the busyness of the road, speed limits and the amount of traffic lights on the path have a big influence on trip duration. Luckily, alongside the traffic networks, our tools for navigation have also improved massively. Modern systems like Google Maps [6] use GPS [17] to provide real-time updates, but the service also monitors the aforementioned traffic factors and considers them when asked for directions.

At the same time, the rise of social media has changed how many people share information. Platforms like Facebook and X have become popular outlets for individuals to share their thoughts, views and experiences on any topic, including traffic conditions. Due to the nature of social media, this information is real-time, and it is no surprise that many people may use it to stay up to date with current traffic, helping them make decisions for the route they want to take.

1.1 Paper Outline

The rest of the introduction discusses the motivation behind the research, the problem statement and what the contributions of the paper are. Afterwards, the paper covers the state of the art in the field of traffic detection through social media platforms, as well as which social media platforms are relevant to research.

Then, Research Questions (RQs) are defined - each one with its section, covering the methodology, results and relevant discussion of the RQ. Finally, a section compiling all of the results of the research, the limitations encountered along the way, and suggestions for future research. The paper ends with a conclusion section, summarising the contents.

1.2 Contribution

The goal of the paper is to contribute to the field of transportation and traffic management by exploring the influence of social media on those subjects.

1.2.1 New Datasets. Two novel datasets are constructed for the sake of this research, collecting social media posts from X and traffic flow & incident information from HERE Technologies. These datasets are quite large and can be used for future research in the field. More details about the datasets can be found in Section 7

1.2.2 Novel Analysis. Many papers have been written on being able to detect traffic accidents and other such events through social media, however, almost no research deals with the influence of social media on people's traffic behaviours. This paper hopes to bring a new perspective into the overlapping networking fields of social media and traffic.

1.2.3 Insight into Available APIs. Since the research requires data from both social media platforms and traffic info providers, all the relevant findings about publicly available APIs will be shared.

1.2.4 Future Research Directions. The paper will outline how the work can be continued and expanded by including the exact methodologies used, the limitations in the process and how they could be circumvented, should this research be continued.

[1] https://www.google.com/maps

1.3 Research Motivation

The motivation for this research lies in the undeniable impact of social media on many aspects of daily life. The proposed initial idea involved inspecting the influence of social media on more microscopical traffic behaviours - overtaking, speeding, lane changing, proper use of turn signals, keeping distance from other vehicles, etc., however obtaining relevant data would have been beyond the scope of the research. Thus, the scope was broadened to a more macroscopic one - route decisions.

Despite the altered scope, the motivation remains the same - finding a statistical significance in the influence of social media could mean avenues for integrating social media into existing navigation tools, or standalone software that aggregates only relevant data for user convenience, with tangible research for its application and usefulness.

1.4 Problem Statement

The main problem that this research addresses is the influence and integration of social media data on route selection and traffic management. Modern navigation systems like Google Maps provide real-time updates and optimal route suggestions based on GPS and other sensor-based sources. These systems could potentially be improved by incorporating real-time user-generated information from social media, if this data is reliable.

The difficulty of the problem is determining whether social media can enhance the accuracy and efficiency of existing navigation systems, and to what extent. This involves analysing the impact of social media on route choices. By addressing this problem, this research aims to find new possibilities for improving navigation tools, which could lead to the development of more advanced systems that provide higher quality traffic info.

2 RESEARCH QUESTIONS

The following two RQs were defined to help drive the research:

- **RQ1: How can data from social media platforms, such as X, and real-time traffic APIs be used to detect alterations in driving routes?**

This RQ focuses on the different ways of using social media APIs and real-time traffic APIs together to identify changes in driving routes. Many factors go into individuals' choice of route, so it is important to find a methodology that isolates social media as the only variable. A side focus of this RQ goes into the general use of such APIs for research.

- **RQ2: What is the impact of social media influence on altered routes in the traffic patterns of a studied city?**

By following a methodology from RQ1, this question answers how big of an impact social media has, and what the implications are.

3 STATE OF THE ART

This section will not only discuss other work relevant to the paper, but also past work in the field of traffic and social media. It will also cover the current technologies for navigation and social media platforms with their usability interfaces.

3.1 Relevant Work

A lot of research addresses the influence of social media on people's daily lives and decision-making, however their scope rarely covers traffic and route-related choices. Searching for papers related to "traffic" and "social media" leads to many false positive matches, most of which address the "web traffic" on those platforms - how many people visit them. The closest paper [10] explores the idea that increased media coverage on road safety results in a decrease in traffic accidents. The study quantifies the presence of news coverage in major newspapers over nine years, and the analysis reveals a strong negative correlation between media coverage and traffic accident rates. This paper outlines that media coverage clearly influences traffic decisions of people, albeit it uses traditional media outlets - newspapers, as the media of investigation. Translating the work [10] to adapt it for social media is a good first step to conducting the research of this paper.

3.2 Topic-adjacent work

Despite the lack of research done on the specific topic of this paper, many papers have been written about the detection and identification of traffic accidents through social media. This section will summarise the findings of papers that contributed to the broader field of traffic monitoring. One such research [12] achieved between 91% and 94% accuracy in detecting real-time jams and accidents through a text-mining approach combined with logistic regression and a support vector machine (SVM). The same research outlined one of the biggest challenges with the accurate classification of social media posts - the variability in post quality and reliability. Language ambiguity is still a blockade that remains to this day, albeit some partial solutions have been developed and used by papers like [14], which will be further discussed in a later section.

Similar methodologies can be observed in many other papers [4, 7, 8, 19] - either employing text-mining techniques or different machine learning models to analyse traffic-related posts, or in some cases both. Furthermore, the methodologies are not the only consistent thing throughout the research - but the results too. All of the mentioned papers have a very high accuracy in detecting traffic congestion and accidents.

Building upon the foundation provided by the previous research, this paper aims to extend the application of the methodologies to explore a novel aspect of traffic studies — how the visibility and sharing of traffic-related information on social media might influence individual driving behaviours.

3.3 Current Platforms

To conduct this research, two sources of data are required: a social media platform and a traffic conditions or navigation service. The following is not an exhaustive list of platforms

that can be used, but relevant findings about the biggest services that were considered during this research.

3.3.1 *Traffic Information providers.*

Google Maps[1] - The largest navigation service nowadays provides excellent real time information, however, the specificity of their API makes it hard to get the exact data that is needed for this research. Traffic index information, such as average speed, congestion index, amount of cars, etc. is not directly available through the Directions API. Furthermore, the target area is specified by two points, source and destination, making it hard to construct a dataset for an entire city. Google Maps also does not provide any historical data, only real-time.

TomTom[2] - TomTom's database of traffic information is probably the biggest and most extensive, and their various APIs provide access to all kinds of useful information, dating back to 2008. Their MOVE service allows users to demo some of the best features, including historical traffic index data in an area [1], which is exactly what is needed for this research. Unfortunately, almost all of their services are behind a paywall, and they do not offer an educational package for research purposes.

HERE Technologies[3] - HERE's platform offers many different services, one of which is their Traffic API. Similar to Google Maps[1], the Traffic API offers a real-time traffic flow overview, as well as an endpoint for incidents, however unlike Google, it allows for an area to be specified. Additionally, the traffic flow information contains traffic indices for average speed, jam factor, traversability and others. This makes it very convenient to collect data for this research.

OpenStreetMaps[4] - While a great free service, OpenStreetMaps only provides geographical data. This includes streets, buildings, terrain and other useful map details, which can be very useful for other research.

3.3.2 *Social Media Platforms.*

X (formerly Twitter) - X's real-time nature allows for immediate sharing of traffic updates, which can be vital for detecting changes in traffic patterns. Combined with their public, albeit locked behind a paywall, API makes it extremely useful for any research involving social media data, including this one. While it does not have the largest active user base, the availability of data through their API makes this platform the best one for research.

TikTok - While incredibly popular, this platform focuses on short video content and their API only allows developers to embed content from TikTok into other platforms. Additionally, the age of the platform makes it nearly impossible to look at historical data for analysis, thus making it unusable for this research.

Facebook - Facebook's user groups are perfect for the aims of this research - dedicated user spaces for sharing traffic updates. That, in addition to having a more than 3 times bigger user base than X [15], makes this platform a great candidate for the target of this paper. However, Meta's policies on data access are a lot more restrictive and do not provide an option to search through posts' content. This also means other social media platforms owned by Meta - WhatsApp, Instagram, and Threads are also not usable for the sake of this research.

[2]https://www.tomtom.com/
[3]https://www.here.com/
[4]https://www.openstreetmap.org/

4 RQ1: DETECTING ROUTE ALTERATIONS VIA SOCIAL MEDIA AND TRAFFIC STATISTICS APIS

4.1 Methodology

4.1.1 Data Availability Analysis. The first step to answering RQ1 lies in the decision of what data is relevant to the research. Section 3.3 alludes to the necessity of traffic metrics in the traffic dataset, but it does not explore the different available metrics and their varying usefulness. Conversely, a social media dataset only requires the content of the post and its timestamp, however the content must also be relevant. Thus, for the sake of this research, various databases with existing datasets will be investigated, as well as different available APIs for constructing a new dataset.

4.1.2 Literature Review. As the previous section mentions, the content of the social media posts must be relevant to traffic conditions. Many papers exist on the detection of traffic accidents through social media, and their methodologies will shine a light on the best approach when it comes to parsing through a large collection of unsorted posts. Reviewing them and collecting the best ideas will be key to being able to construct a good traffic-related social media dataset.

4.2 Process and Results

A lot of the research done in this section is a continuation and expansion of section 3.2 in the State of the Art. To answer RQ1, sub-questions are defined:

4.2.1 Sub-question 1: **What traffic metrics are relevant to quantifying road conditions?**

To be able to compare two states of the same road, or different roads, there must be a formalised set of metrics that are considered for the comparison. "Traffic Flow Fundamentals" [11] provides methodologies for measuring traffic volume, speed and density. The book's goal is to offer a better understanding of how traffic concepts and methodologies can be applied to solve real-world transportation problems. Another similar book is the "Highway Capacity Manual" [2], which focuses on methodologies for evaluating traffic flow, congestion and road capacity. It introduces concepts such as Level of Service, which is a qualitative measure describing conditions within a traffic stream. Most papers seem to reference these two as the baseline for traffic and congestion metrics, however, some offer a novel approach like [13], which is a case study of six areas through "weighted congestion" and "normalised congestion". The following metrics were compiled

as the best for the sake of this research, as the data processing requirement with them is fairly low, and, as general metrics, they are more likely to be available through traffic information platforms:[5]

- **Traffic Volume** - The number of cars on the road.
- **Traffic Speed** - Average speed of cars on the road.
- **Congestion Level** - The degree of saturation in traffic flow, quantifying how "full" a road is and how close to freely flowing the vehicles are.

4.2.2 *Sub-question 2:* How can social media posts be connected to traffic conditions?

While there is not a complex collection of different metrics when it comes to social media, tying posts to traffic conditions can still be a challenge due to the nature of user-generated content with little-to-no moderation. Social media users are free to post anything, which leads to language ambiguity, which [3] identified as a major limitation. Examples of such ambiguities are words with multiple meanings - "jam" could mean traffic jam, but is also a term used between musicians and even a food item. Furthermore, sarcasm and hyperbole can drastically change the meaning of text compared to its face value.

The same paper [3] also identifies location ambiguity as the second biggest limitation. Most people use shortened versions of street names, skip out on the type of road or use abbreviated forms - "fwy" instead of "freeway" or "blvd" for "boulevard". One potential solution to this problem is considering the geolocation of the post - platforms like X provide the coordinates of the post, so it is safe to assume that the closest matching road is the one in question. However, this means the research would no longer be from the perspective of an average user browsing the platform, and is, as such, not a valid solution.

A clever way to address both language and location ambiguity is employed by [14], who identify X accounts owned by traffic organisations or the government, with the purpose of traffic updates. These accounts follow a standardised approach in their posts, making it significantly easier to process and increase the reliability of the data. The downside of this approach is the more limited availability of the data, as well as reducing the social aspect of social media. Following such accounts for traffic updates is no different from listening to the radio or watching the news on TV, which is not of interest for the sake of this paper.

Similar to the traffic conditions information, data availability must be considered. For an existing dataset, language processing through a large language model to identify relevant posts is the best approach. This requires a large enough dataset that is not already classified or filtered, as well as a training set for the LLM, time and computational power to train the model, all of which are not available in this research. A simpler alternative is constructing a list of explicit traffic keywords and an extensive list of street names in the target area, using regular expressions to match both abbreviated and full forms of road types, and using these lists to query matching posts.

[5]The traffic metrics are considered in a given period of time

4.2.3 *RQ1:* How can data from social media platforms, such as X, and real-time traffic APIs be used to detect alterations in driving routes?

After answering the two sub-questions, the research is ready to address the potential relationship between the two datasets and how it can be investigated. Many factors go into the decision-making when it comes to routes that people take. Navigation systems like Google Maps already consider traffic conditions and travel time when suggesting a route. Staying up to date with news allows drivers to know that roads are closed and whether they need to take an alternative route. General intuition for factors like rush hour is in the back of the minds of commuters. Isolating social media's influence among all these factors requires two groups of data - one with social media coverage and one without. Averaging out the traffic conditions in both sets should provide a pair which only differs in the amount of social media posts about the state of the road, assuming the datasets are large enough, about the same area and during similar traffic-altering windows. To specify the problem and allow for these requirements to be met, it is beneficial to only consider traffic flow around accidents, in an area with high social media coverage. Then, all of the aforementioned available traffic metrics in section 4.2.2 can be compiled into a single traffic index, the change of which can be observed around the time of the accident. Finally, investigate whether the number of social media posts about the accident has an impact on the changes of the traffic index.

4.3 Discussion

4.3.1 Conclusions: To conclude this RQ, the best approach for this paper is to investigate whether there is a correlation between the amount of social media posts about a traffic accident and the traffic index of the road around the time of the accident. The traffic index is constructed from relevant road metrics like traffic volume and density, and congestion levels, which are representative of how many cars are on the road. A lower traffic index than normal means fewer cars on the road, meaning more people altered their routes.

4.3.2 Limitations: This methodology comes with some limitations. It is entirely dependent on data availability and quality, which can severely threaten the validity of the conclusion. Furthermore, this research can potentially identify correlation, but cannot establish causation. The amount of external factors makes it impossible to be conclusive, even if the methodology aims to exclude them from the target of investigation.

4.3.3 Future Work: Future research on this topic can be improved in multiple ways. Integration of broader data sources would provide a more comprehensive view of the traffic dynamics. A better model for collecting usable social media posts through language models for sentiment and traffic relevancy would be extremely beneficial for data validity. Traffic

behaviour model simulations can do a much better job of portraying potential correlations.

5 RQ2: IMPACT OF SOCIAL MEDIA ON DRIVING ROUTE DECISIONS

5.1 Methodology

To investigate the impact of social media, the methodology from RQ1 is followed. For the sake of this research, the area of interest is Los Angeles (LA), a city in southern California, US. LA is known for its extensive and complex road, and is especially notorious for its low "walkability" [9]. This, alongside a relatively high X usage, means it is an excellent target for the research.

The high X usage of LA, plus other beneficial parts of it mentioned in section 3.3.2, means X is the best social media platform to be used. Due to most of their API no longer being free [18], the research requires X's Developer Basic plan to construct a new dataset. The plan provides access to the "search_recent_tweets" endpoint of their API, which allows for a queried search of all tweets in the last 7 days. Existing datasets were considered, but very few contained any tweets that are related to LA traffic.

For traffic statistics, TomTom was the initial best candidate, with their extensive coverage and easy to use traffic metrics. However, after getting in contact with their customer support and sales team, the company made it apparent that they have no interest in providing data for educational or research purposes. Consequently, the research uses HERE Technologies, as most appropriate accessible option, as discussed in previous sections. Requests to their two endpoints for real-time traffic flow and real-time incident tracking were made every 30 minutes to construct a dataset over 2 weeks. The target area for these requests is a circle at latitude 34.052235 and longitude -118.243683 - the centre of downtown LA, and a radius of 50 kilometers.

Since the data was requested every 30 minutes, at 19 and 49 minutes past each hour, traffic flow data from exactly before an incident is not necessarily available. Thus, "before incident" refers to the most recent traffic flow snapshot that was acquired prior to the accident, and "after incident" is the following snapshot, 30 minutes later. To continue monitoring changes, the next three snapshots are also considered - labelled "after incident 30", "after incident 60", and "after incident 90" respectively. Figures 3 and 4 follow this convention.

The traffic flow information provides several metrics: speed, speedUncapped, freeFlow, jamFactor, confidence, traversability, the full description of which can be seen in their documentation [16]. Unfortunately, the data was filled with inconsistencies in the speeds and freeflow, rendering those metrics unusable. This leaves "jamFactor" as the only usable traffic metric, and is what is as the traffic index.

The incident data is also used to construct a list of streets of interest. Since every accident provides the streets it happened on, it is convenient to compile them into a list that can be used to query tweets, alongside relevant traffic keywords. After requesting data from X with this query parameter, the posts are further filtered to be within the time of an incident and 2 hours after. Finally, for each accident, the number of tweets referring to the incident in the specified time frame is counted.

The result of all this data collection and processing is a table containing 20000 incidents, with the time of the incident, streets of the incident, amount of tweets for the incident, and the traffic index at the affected streets in the time window of 2 hours. To further isolate the problem and reduce the variables, incidents during rush hour are filtered out. This brings down the total amount of incidents by about half, while only reducing the incidents with tweets related to them by about 20%. The distribution of traffic indices can be seen in Figure 1. It also shows that no road had an index above 80 before the incident, and perhaps an indication that accidents are less likely to occur on congested roads.

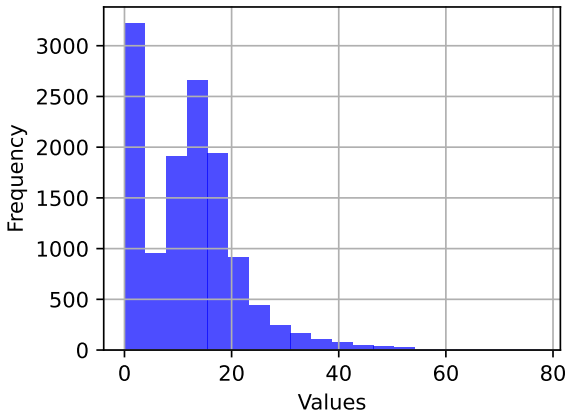

Figure 1: Distribution of traffic indices at the last measurement point before an accident occurred

To determine whether there is a relationship between the number of tweets, a formal statistical approach is employed:
1. **Variables**: Let X represent the *tweet_amount* and Y represent the *after_incident* measurements.
2. **Correlation Coefficient**: Define ρ (rho) as the population correlation coefficient between X and Y.
3. **Null Hypothesis (H_0)**: $H_0 : \rho = 0$ - there is no correlation between the number of tweets and the traffic index changes (i.e., the correlation coefficient ρ is zero).
4. **Alternative Hypothesis (H_a)**: $H_a : \rho \neq 0$ - there is a correlation between the number of tweets and the traffic index changes (i.e., the correlation coefficient ρ is not zero).
5. **Statistical Test**: Construct a correlation matrix and an Independent Two-Sample t-Test. To conduct the tests, an approximately normal distribution is required. In Figure 1 it can be observed that the distribution appears to be normal if we exclude samples with traffic index < 2. Shapiro-Wilk's test on normality is used to investigate this.

6. **Decision Rule**: - If the p-value is less than the significance level $\alpha = 0.05$, reject the null hypothesis H_0, concluding there is statistically significant evidence of a correlation. If the p-value is greater than the significance level, do not reject the null hypothesis, concluding there is not enough evidence to suggest a correlation.

5.2 Results

Figures 3 and 4 indicate that the presence of tweets does have some impact on the traffic index, albeit not the expected one - across the board, all traffic indices are higher when there are posts about the accident. One potential explanation is that smaller scale incidents with better traffic indices are less likely to be tweeted about, and conversely, big accidents are more likely to be covered by user posts.

The correlation matrix in Figure 2 shows an extremely weak correlation factor between the tweet amount and the traffic index changes, suggesting that there is not enough evidence to reject H_0.

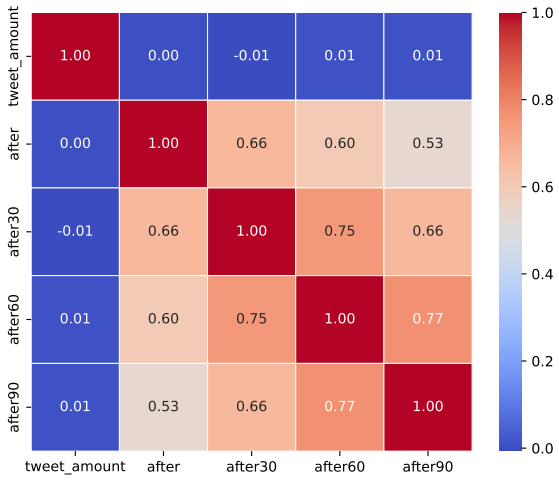

Figure 2: Correlation Matrix between Tweet Amount and Traffic Index changes

To verify these results, an Independent Two sample t-Test is conducted, the results of which can be seen in Table 1. While the test can be used on non-parametric data, its accuracy is better on a normal distribution. To confirm normality, Shapiro-Wilk's test is used, the statistic of which suggests that the distribution is indeed approximately normal. The table 1 has the calculated P-values from the t-Test for each of the 4 traffic flow snapshots, and they align with the correlation matrix - since neither of them is less than the specified significance level $\alpha = 0.05$, there is not enough evidence of a strong correlation between the amount of tweets and the traffic index. While there might be some correlation, as pointed out in Figures 3 and 4, it is not a statistically significant one.

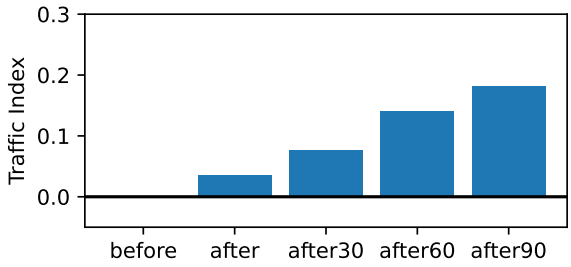

Figure 3: Traffic indices for incidents without related Tweets, showing the relative change immediately after the incident, and every 30 minutes afterwards.

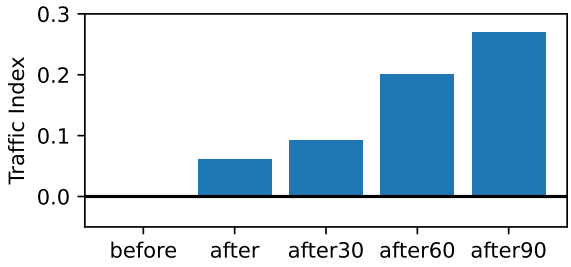

Figure 4: Traffic indices for incidents that have related Tweets, showing the relative change immediately after the incident, and every 30 minutes afterwards.

5.3 Discussion

The investigation of the relationship between social media activity and traffic condition changes has provided some interesting insights. Although the results indicate a generally weak correlation, this also has relevant implications. It means that while social media can be an informal traffic news network through real-time updates about traffic, its efficacy in impacting drivers' choice of route is limited.

The overall worse traffic conditions in the case of incidents that were covered on social media suggest a bias in how likely a driver is to tweet about an accident. Bigger incidents, with larger consequences on traffic, are more commonly brought up in social media, which is another variable that influences the statistics. To address this in future research, comparing traffic events of equal severity can be beneficial.

Time Interval	P-value
After the incident	0.28297
30 minutes later	0.55132
60 minutes later	0.36045
90 minutes later	0.21659

Table 1: Results of the t-Test

The research also underlines the importance of robust data for reliable traffic metrics, as inconsistencies in the data either render it useless, or require further processing, which is costly and prone to errors.

6 RESULTS

This section compiles all the findings from the research and analyses done in the previous sections. Utilising data from social media platforms and real-time traffic information providers, namely X and HERE Technologies, this study investigated the potential correlation between social media activity and traffic patterns.

6.1 Traffic Metrics and Social Media Influence

This research noted down which traffic metrics are relevant in quantifying traffic conditions - traffic volume, density, average speed and congestion levels. Then analysed the metrics' links to the volume of social media posts. The analysis indicated a moderate correlation, suggesting that increased social media activity does have some influence on traffic patterns. In the case of incidents with higher social media coverage, the traffic indices were also noticeably higher.

6.2 Statistical Analysis of Social Media Impact

This correlation between the number of social media posts and the traffic index changes was further examined through a statistical analysis. Results from the correlation matrix (Figure 2) showed an extremely weak correlation factor, suggesting that while there is an observable change in traffic indices, it may not be statistically significant.

The results from the Independent Two-Sample t-Test further supported this observation:

- At the first snapshot after the incident, the p-value was 0.28297, indicating no significant correlation.
- 30 minutes after the incident, the p-value increased to 0.55132, further diminishing any assumed correlation.
- Similar trends were observed 60 and 90 minutes after the incident, with p-values of 0.36045 and 0.21659, respectively.

These results suggest that while social media posts may affect traffic conditions, the impact is not one of statistical significance.

6.3 Implications for Traffic Management

These findings showcase the potential utility of social media as a real-time information source for traffic management. While the correlations are weak, the research shows that useful information can be drawn out of social media posts - e.g. if an accident is covered on social media, it is more likely the traffic conditions are worse than if there are no posts about it.

7 CONTRIBUTIONS

New social media dataset - A collection of X posts between June 1^{st} and June 14^{th}, 2024, referring to a traffic accident and a street in Los Angeles. New X datasets are scarce due to the platform no longer supporting an open API, and their usefulness expands to many fields of research.

Traffic flow and incidents dataset - Many platforms offer real-time information, however not many have historical data. The snapshot of traffic metrics data, plus a list of incidents was collected between June 1^{st} and June 14^{th}, 2024, matching the time frame of the X dataset.

Novel Analysis of social media relationships - This research pioneers a novel analysis of the impact of social media on traffic behaviour, specifically how information from X influences route choices around traffic incidents. Unlike previous studies, which mainly focus on detecting traffic events through social media platforms, this paper links the usage of such platforms to changes in real-world traffic patterns.

List of limitations and recommendations for future work - As a topic with low coverage, it is important to ensure future work investigating this subject is familiar with the limitations of previous research. As such, this paper provides an extensive list of limitations, how some of them can be avoided and how the work can be improved in the future.

Meta analysis of semi-relevant papers - With a lot of work covering the broader topic of detecting traffic accidents through social media, it is important to make sure results from different papers are consistent with each other.

Insight and criticism of available APIs - As most research deals with data analysis, or is reviewing other research which uses data analysis, the availability of data is indisputably one of the most important things in research. This paper outlines the current state of social media platforms and traffic information providers, with insights into their API policies, and a criticism of their limitations that might threaten research as a whole.

8 LIMITATIONS

This research faced many limitations during the process. One such limitation, which is a hindrance for most, if not all, research, is the availability of free open APIs. As already discussed, X is the only social media that makes it available to query content from the platform. Due to this, the platform can also take advantage of this monopoly by employing harsh limitations on the availability of the API for free users. A similar issue can be observed in traffic information providers as well. While a free API is unreasonable to ask for, an educational or research plan would be extremely beneficial for scientists around the world, that ultimately conduct research for the good of everyone. Without one, the ability to do robust research through relevant and extensive data will only continue to be limited.

Another limitation was already alluded to in previous sections of the paper. Machine learning models were discussed as a way to classify relevant social media posts. This method was, however, not applicable for this research, due to the X API limitations. The "search_recent_tweets" endpoint allowed for a query of a maximum of 512 characters, a maximum of 100 results and a maximum of 60 requests per 15

minutes. The total amount of allowed requests is 1500. With a list of almost a thousand street names, a big limitation was how many traffic-related keywords were included. The exact query was as follows:

```
((traffic OR road) AND
(accident OR congestion OR jam))
AND
(<street name> (OR <street name>)*)
```

This query means that if someone wrote "road X is congested" instead of "there is congestion on road X", it would not get picked up by the query. Naturally, this leads to much fewer relevant tweets being collected.

Furthermore, as RQ1 addresses, many different factors influence traffic. In section 4, only factors with data available for them are mentioned, since the research can only account for them. Other factors include weather conditions, public events and holidays, infrastructure issues or changes - all of these can have an impact on traffic, introducing variation in the data and threatening the validity of the results.

All of these limitations stress the difficulty in establishing causality from correlational data. While the research identifies some weak correlation, the actual influence of social media on driving behaviours cannot be conclusively determined without prolonged study with vigorous and extensive data.

8.1 Recommendations for Future Research

To expand on the framework provided by this paper, future research should take all the aforementioned limitations into account. As concrete points for improvement:

- **Expanding data sources:** Incorporating more traffic data sources with more information about the traffic conditions, as well as expanding the keyword parameters for the queried word search through X's API. More extensive data, and a generally larger dataset, will increase the validity and certainty of any results.
- **Alternative data sources:** As there are endpoint limitations with X's developer tools, an alternative approach can include using existing unlabelled datasets of tweets, out of which relevant ones need to be extracted. Natural language processing techniques can be used to better filter and interpret social media content, resulting in a dataset containing more relevant posts compared to the regular word search through the "search_all_tweets" endpoint of X.
- **Prolonged studies:** Due to the nature of social media, and the continuous evolution of society's internet usage, the influence of social media might be drastically different throughout the years. Investigating this could further improve the understanding of this influence.
- **Improved statistical techniques:** Given the traffic indices being ordinal, a Mann-Whitney U Test, also known as the Wilcoxon rank-sum Test, could be performed. By comparing ranks of the traffic indices instead of the values, this test can be more effective than a t-Test when dealing with non-normal distributions. It can help identify outliers more accurately - so it can be used to determine whether incidents with a higher associated volume of tweets are outliers.
- **Cross-disciplinary approach:** As the original idea of the research proposed, an investigation on all traffic behaviours could be conducted with the assistance of sociology or psychology experts. Traffic navigation systems already suggest route alterations, however, no technology exists for microscopic behaviours, so research that focuses on them can provide beneficial insights for drivers.

9 CONCLUSION

This study has explored the significant role of social media, specifically X, in influencing route selection during and after traffic incidents. The findings demonstrate that real-time traffic updates do have some correlational influences on the decision, however, a statistical analysis proved that this correlation is not statistically significant.

9.1 Limitations

The research faced several limitations, most notably data availability and processing. As the research dealt with user-generated content, it had to consider language and location ambiguity as part of the social media post texts. Posts often contain colloquialisms, which make it hard to request all relevant data through a simple word search. Furthermore, a methodological limitation was isolating the impact of social media from all the other factors that influence traffic patterns.

9.2 Future Directions

Future studies should focus on expanding the data, such that it can be separated into different control and testing groups of sufficient size. Establishing a ground truth for traffic conditions with and without incidents can help isolate how much of an influence social media is, which can improve the validity of the results. Furthermore, different traffic metrics like the relative volume of cars on the road can be incorporated into the general traffic index, such that it is more indicative of the traffic situation.

Expanding the target traffic behaviours beyond just route alteration, as well as the geographic and demographic range of the data, can help showcase differences in driving cultures and social media influence in different groups, which can be valuable insight for all drivers.

9.3 Implications for Traffic Management

The findings of this research outline areas where social media could be incorporated into traffic management platforms, however due to the weak correlations and the fact that existing navigation platforms like Google Maps already notify their users of accidents on their route, and suggest the best alternative route.

10 USAGE OF AI

This section is an overview of the usage of AI tools in accordance with the Univesity's regulations. In the process of writing this paper, ChatGPT[6] and Grammarly[7] were used in order to catch any grammar mistakes or inconsistencies in the writing, ensuring the writing remains formal and clear. ChatGPT was also utilised for its language model capabilities in attempt to classify tweets into relevant and not-relevant to traffic, however the result was never used for the research.

REFERENCES

[1] D. Asenov. 2024. TomTom MOVE example. https://ts.tomtom.com/reports/share/details/CPC/4421392?t=dc1afa8b-3311-4d09-b0ec-eba71e8a4dcb

[2] Transportation Research Board. 2016. *Highway Capacity Manual* (6th ed.). Transportation Research Board, Washington, D.C.

[3] Po-Ta Chen, Feng Chen, and Zhen Qian. 2014. Road Traffic Congestion Monitoring in Social Media with Hinge-Loss Markov Random Fields. (2014), 80–89. https://doi.org/10.1109/ICDM.2014.139

[4] Eleonora D'Andrea, Pietro Ducange, Beatrice Lazzerini, and Francesco Marcelloni. 2015. Real-Time Detection of Traffic From Twitter Stream Analysis. *IEEE Transactions on Intelligent Transportation Systems* 16, 4 (2015), 2269–2283. https://doi.org/10.1109/TITS.2015.2404431

[5] Stacy C Davis and Robert Gary Boundy. 2021. *Transportation energy data book: Edition 39*. Technical Report. Oak Ridge National Lab.(ORNL), Oak Ridge, TN (United States). https://doi.org/10.2172/1767864 This, and all previous editions of the book are relevant to the paper..

[6] Google LLC. 2024. Google Maps. https://www.google.com/maps Accessed: 2024-06-21.

[7] Yiming Gu, Zhen (Sean) Qian, and Feng Chen. 2016. From Twitter to detector: Real-time traffic incident detection using social media data. *Transportation Research Part C: Emerging Technologies* 67 (2016), 321–342. https://doi.org/10.1016/j.trc.2016.02.011

[8] Guntur Budi Herwanto and Deny Prasetya Dewantara. 2018. Traffic Condition Information Extraction From Twitter Data. In *2018 International Conference on Electrical Engineering and Informatics (ICELTICs)*. 95–100. https://doi.org/10.1109/ICELTICS.2018.8548921

[9] Donghwan Ki and Zhenhua Chen. 2023. Walkability inequity in Los Angeles: Uncovering the overlooked role of micro-level features. *Transportation Research Part D: Transport and Environment* 122 (2023), 103888. https://doi.org/10.1016/j.trd.2023.103888

[10] Antonio Javier Lucas, Francisco Alonso, Mireia Faus, and Arash Javadinejad. 2024. The Role of News Media in Reducing Traffic Accidents. *Societies* 14, 5 (2024). https://doi.org/10.3390/soc14050056

[11] Adolf D. May. 1990. *Traffic Flow Fundamentals*. Prentice Hall, Englewood Cliffs, NJ.

[12] Prabu Kresna Putra, Rahmad Mahendra, and Indra Budi. 2022. Traffic and road conditions monitoring system using extracted information from Twitter. *Journal of Big Data* 9 (2022), 65. Issue 1. https://doi.org/10.1186/s40537-022-00621-3

[13] Jeong Seong, Yunsik Kim, Hyewon Goh, Hyunmin Kim, and Ana Stanescu. 2023. Measuring Traffic Congestion with Novel Metrics: A Case Study of Six U.S. Metropolitan Areas. *ISPRS International Journal of Geo-Information* 12, 3 (2023). https://doi.org/10.3390/ijgi12030130

[14] Dayong Shen, Longfei Zhang, Jianping Cao, and Senzhang Wang. 2018. Forecasting Citywide Traffic Congestion Based on Social Media. *Wireless Personal Communications* 103, 1 (11 2018), 1037–1057. https://doi.org/10.1007/s11277-018-5495-x

[15] Jack Shepherd. 2024. Essential Facebook Statistics In 2024. https://thesocialshepherd.com/blog/facebook-statistics

[16] HERE Technologies. 2024. HERE Traffic API Documentation. https://registry.services.api.platform.here.com/v1/services/hrn%3Ahere%3Aservice%3A%3Aolp-here%3Atraffic-api-7/openApi

[17] U.S. Department of Defense. 2024. Global Positioning System. Online. https://www.gps.gov/ Accessed: 2024-06-21.

[18] XDevelopers. 2023. XDevelopers announces free access to Twitter API will no longer be supported. https://x.com/XDevelopers/status/1621026986784337922?lang=en

[19] Zhenhua Zhang, Qing He, Jing Gao, and Ming Ni. 2018. A deep learning approach for detecting traffic accidents from social media data. *Transportation Research Part C: Emerging Technologies* 86 (2018), 580–596. https://doi.org/10.1016/j.trc.2017.11.027

[6] https://chat.openai.com/

[7] https://grammarly.com/