Deciphering 140 Characters:
Text Mining Tweets On #DriverDistraction

Name and affiliation information

Objective: Conduct an exploratory analysis of driver distraction tweets using text mining. **Background:** Twitter is a popular social networking site with a wealth of data that is both explanatory and predictive of current trends and events. Data from Twitter may also prove useful in understanding the attitudes and opinions surrounding distracted driving. **Method:** Tweets posted between January 29, 2012 and April 12, 2013 containing the words ‘driver distraction’ or ‘driving distraction’ were collected. Text mining was used to extract patterns from the tweets in terms of timelines, frequencies, and associations. **Results:** Tweets contained information about users’ personal experience with driver distraction as well as various news articles about driver distraction. **Conclusion:** Twitter data provide a real-time snapshot of the attitudes surrounding of distracted driving. **Application:** Information from social media can complement traditional driving data sources, such as simulator studies, naturalistic studies, and epidemiological data, to create a more holistic picture of distracted driving.

INTRODUCTION

The use of social media has exponentially increased in the last decade. Over one billion people use social networking tools such as Facebook, Google +, LinkedIn, Pinterest, and Twitter. With such a large user base also comes extensive user-created content that measures the pulse of current events and trends.

Twitter, a microblogging website where users chat, converse, share information, and report news (Java, Song, Finin, & Tseng, 2007), is of interest because their users have the ability to quickly share information using methods similar to text messaging. They do so by creating short messages of 140 characters or less, known as tweets. Tweets also contain links to pictures, videos, or other websites as well as hashtags—words that begin with # and are turned into links to make it easier to find certain terms. Once a tweet is posted, it can be retweeted, or shared by another user (https://discover.twitter.com/learn-more). Twitter’s 240 million + monthly users send 500 million tweets per day (https://about.twitter.com/company).

The wealth of information available from Twitter has prompted an extensive and diverse body of research on the prediction of box office revenue (Asur & Huberman, 2010), stock market prices (Bollen, Mao, & Zeng, 2011), earthquakes (Sakaki, Okazaki, & Matsuo, 2010), and election results (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Similarly, Twitter has been used to help understand public opinion during a political debate (Diakopoulos & Shamma, 2010), a global pandemic (Chew & Eysenbach, 2010), and natural disasters (Vieweg, Hughes, Starbird, & Palen, 2010). In general, attitudes and emotions exhibited through Twitter often reflect current social, political, cultural, and economics events (Bollen, Mao, & Pepe, 2011).

Despite its broad use in other domains, Twitter has not been used in driver safety research, particularly driver distraction. In recent years, the dangers of distracted driving have caused much public debate, especially as it relates to the use of cell phones while driving. According to the official government website for distracted driving (www.distraction.gov), 41 states ban texting while driving and 12 states ban cell phone use while driving. At the same time, it is difficult for law enforcement officials and governments to quickly determine the effectiveness of these laws and whether public opinion surrounding distracted driving is changing. In that sense, Twitter can contribute to a more holistic picture of the culture of distracted driving.

Analysis of driver distraction tweets requires methods beyond typical quantitative analysis applied to numerical data from simulator or naturalistic studies. Compared to numerical data, text is “unstructured, amorphous, and difficult to deal with” (Witten & Frank, 2005, p. 21). As such, text mining, which seeks to find patterns in text and extract useful information, is an appropriate tool for analysis (Feldman & Sanger, 2007; Witten & Frank, 2005). Therefore, the purpose of this research is to use text mining to conduct an exploratory analysis of driver distraction tweets. Given that text mining aims to discover the nature and relationship of concepts based on their co-occurrence as determined by timelines, frequencies, and associations (Feldman & Sanger, 2007), these three methodological tools are the focus of the analysis.

METHOD

Data Collection

The Twitter Application Programming Interface (API) allows researchers to access Twitter data including tweets, user information, and timelines in two formats: REST API and Streaming API. Compared to the Streaming API that allows real-time access to nearly all tweets, the REST API is not an exhaustive source of tweets. However, given its’ simplicity, ease of use, and focus on delivering quality over quantity, the REST API was the preferred choice for this research.

For 3 months (January 11, 2013 to April 12, 2013), tweets were collected using the Twitter REST API v1, which has since been upgraded to v1.1 and is no longer active as of June 2013. Using the search command, tweets written in the English language that contained the words “distraction” and “driver” or “driving” were collected. As the Twitter REST API draws from tweets within users’ most recent 3,200 tweets, tweets composed before the start of data collection could be present in the data. In addition, as the REST API can only return a maximum of 1,500 tweets per search, a script was written in Python 2.7 to conduct a search every 20 minutes to collect as many tweets as possible. From the search, the following information was stored into a JSON file: the text of the tweet, the user ID, the tweet ID, and the time stamp, in Greenwich Mean Time (GMT).

Data Reduction

Text in its raw, natural format is unsuitable for text mining. Preprocessing is required to transform the unstructured raw text into a more manageable format that is suitable for identifying patterns (Feldman & Sanger, 2007). For this research, the following preprocessing steps were taken using the 'tm' package (Feinerer & Hornik, 2014) in R 3.0.2 (R Development Core Team, 2013). First, the raw text from the tweets was turned into a collection of text documents—known as a corpus —such that each document contained one tweet. Second, the corpus was transformed to remove punctuation and to turn all text to lower case. Punctuation was removed as it adds little value to this text analysis, i.e., the pattern of commas or periods contained within 140 characters is not of interest. Given the use of '#' for hashtags and '/' in website links, punctuation removal also deleted information about hashtags and links to other websites. Conversion to lower case improves the analysis as it reduces confusion between identical words, e.g., 'Driving' versus 'driving'. Third, stop words (e.g., 'the' and 'a') were removed, as they add little value to the analysis given their frequent occurrence in the English language (Feinerer, Hornik, & Meyer, 2008). The stop words dictionary from the 'SnowballC' package (Bouchet-Valet, 2013) was used and reduced the mean number of words in a tweet from 17 to 12. Fourth, the text was stemmed—the word suffixes were removed to leave only their radicals. The stemming process reduces the complexity of the text and allows nearly identical words to be treated similarly, e.g., 'drive' is similar to 'driving (Feinerer, et al., 2008). As an example, before preprocessing, a complete tweet was:

```
I'm driving! And my favorite song is on
#distraction #[hashtag]
```

After preprocessing, the tweet became:

```
im drive favorit song distract [hashtag]
```

Using the stemmed terms, a term-document matrix was formed where the rows correspond to documents, i.e. tweets, and the columns represent terms, i.e., words. The number within each cell is the frequency that the term appears in the document (Feinerer, et al., 2008). Finally, sparse terms appearing less than 1% of the time were removed.

RESULTS

Data collection from the Twitter REST API resulted in 8,689 tweets. Preliminary exploratory inspection of the data revealed two distinct types of tweets: one based on media reports, e.g.,

```
According to the National Highway Traffic Safety
Administration, nearly 80% of crashes involved
driver distraction - [link to website],
```

and another based on personal experience, e.g.,

```
[Name] tells me that I can't listen to music
while driving cause it's a distraction. I think
he's jealous of my amazing singing voice.
```

Tweets from the media are regular and cohesive, while personal tweets are broad and random. As such, patterns within the text may be biased by similar tweets referencing the same media report.

Timeline

Tweets were posted between January 29, 2012 and April 12, 2013. Though data collection ended in the middle of April 2013, there are more tweets during that month than any other month alone. Closer inspection of individual tweets reveals that over 1,500 of the 2,556 tweets during April 2013 are retweets stating that:

```
Daydreaming is the most common cause of driving
accidents due to distraction, not people using
their cellphones.
```

As this one tweet was present in 18% of the data, thereby biasing the results, all of these tweets were removed and the proceeding results only report analyses of the remaining 7,106 tweets. Similar media report retweets were present in the data, but not nearly as frequently as the tweet presented above, and as such, they remained in the analysis.

Figure 1 shows a histogram of the number of tweets by month and year. Though any tweet within a user's most recent 3,200 tweets could be present in the data, the REST API is more likely to return recent tweets, hence there are considerably more tweets between January and April 2013.

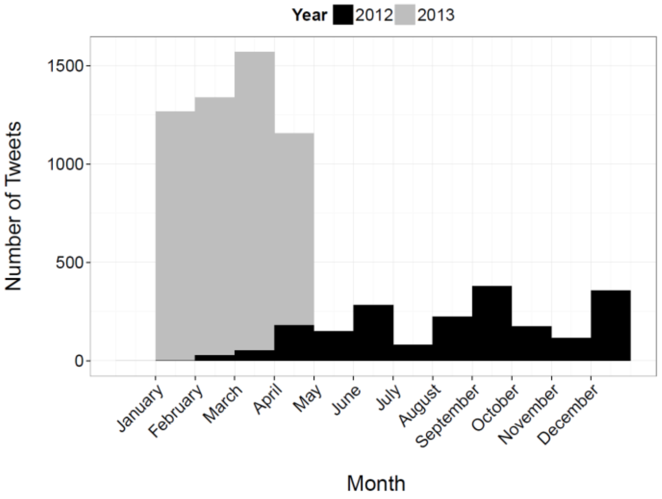

Figure 1: Timeline of tweets separated by month and year.

Term Frequencies

The word cloud in Figure 2 displays stemmed terms that appear more than 200 times. Naturally, the terms 'distract' 'drive', and 'driver' appear more frequently than other terms because only tweets containing these words were collected. As such, they are not displayed in Figure 1. The most frequent terms are 'phone', appearing 1,168 times, and 'text', appearing 869 times. Both terms reflect causes of distracted driving crashes, and are the focus of driver distraction legislation. Other terms appearing frequently require explanation. ‘Studi’ is present 456 times and is a result of tweets reporting results of research studies. ‘New’ appears 315 times as people tweet news articles about driver distraction. Finally, ‘american’ appears 286 times as there is a frequently retweeted Wall Street Journal article about where American drivers hold their cell phone (Rogers, 2012).

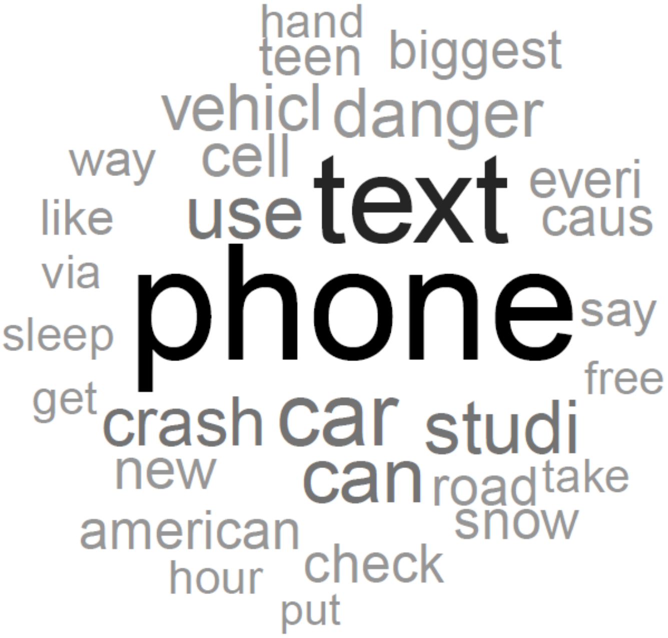

Figure 2: Word cloud displaying frequently tweeted stemmed terms; words that are darker and larger represent more frequent terms.

Associations

Correlation coefficients between ‘distract’, ‘drive’, ‘driver’, and ‘driver distract’, respectively, with other terms are shown in Figure 3. Only correlation coefficients above 0.10 are shown.

‘Month’, ‘awar’, and ‘april’ are all correlated with the term ‘distract’, leading to the conclusion that there are tweets about Distracted Driving Awareness Month, which is during the month of April. During April, the terms “Distracted Driving Awareness” and “Distracted Driver Awareness” collectively appeared in 149 tweets. Interestingly, outside of ‘text’, which already appears in Figure 2, ‘distractionfre’ and ‘pledg’ were all associated with the term ‘drive’ and collectively appeared 175 times. This suggests that many people make a safe driving pledge to be distraction free, e.g.,

```
Raise your hand and take the pledge to do your
part to end distracted driving on our roads
[link to website].
```

Making safe driving pledges has become more common with the increased publicity surrounding the dangers of distracted driving. These pledges are also a popular feature of Distracted Driving Awareness Month.

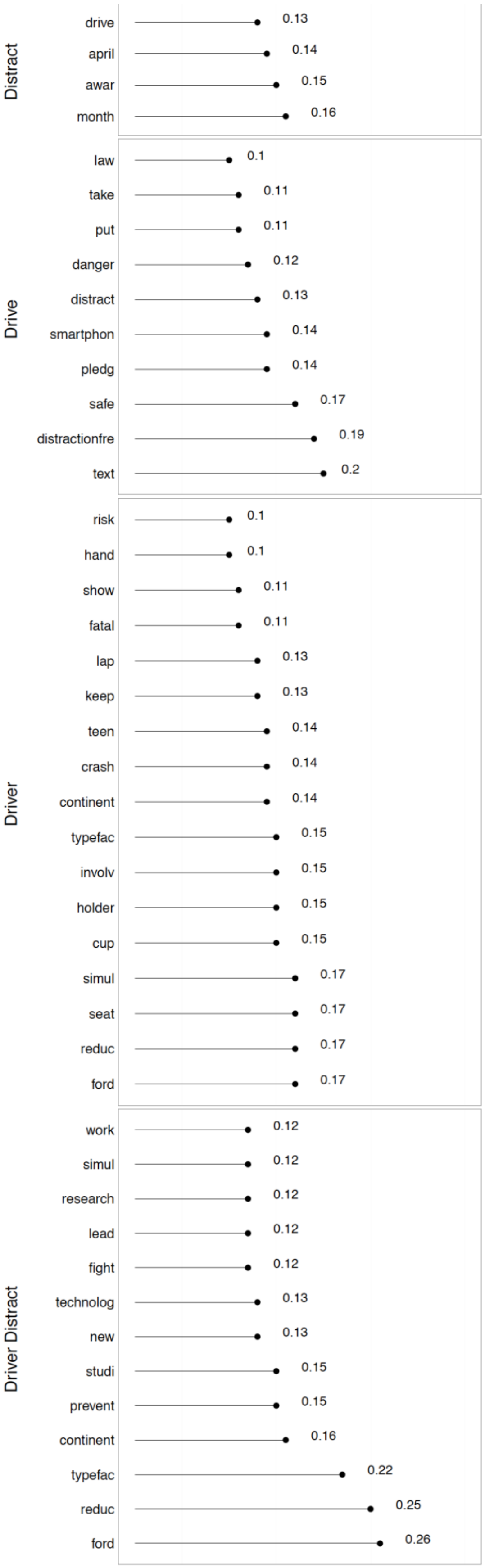

Figure 3: Associations between ‘distract’, ‘drive’, ‘driver’, and ‘driver distract’, respectively, with other terms.

The terms most associated with 'driver' and 'driver distract' represent a mixture of concepts. The collective relationship between 'driver' with 'lap', 'keep', 'holder', and 'cup' points to the frequently retweeted Wall Street Journal article about where people hold their cell phone while driving (Rogers, 2012). The association between 'driver distract' and 'fight' arises from many tweets about fighting to end driver distraction. A few terms are associated with both 'driver' and 'driver distract'. The association between these two terms and 'typefac' results from tweets that discuss and link to an article about how typefaces used on in-vehicle displays affect glance time (Reimer et al., 2012). The association with 'ford' arises from tweets about Ford's inclusion of technology in their vehicles that reduces driver distraction, e.g.,

```
Check out: Ford working on automatic “do not
disturb” function to combat driver distraction
[link to website] via @[username].
```

Similarly, the association with 'continent' points to Continental Automotive's effort to combat driver distraction:

```
Continental Automotive Unveils Concept Vehicle
to Investigate and Address Driver Distraction
[link to website].
```

The correlation with 'simul' relates to automotive manufacturers use of simulators to assess driver distraction as well as the use of simulators to teach teenagers the risk of distracted driving.

Frequencies and Associations

Analyzing distributions, frequencies, and associations independently provide information about driver distraction that goes beyond typical experimental studies. Greater benefits may be achieved by combining the methodological tools. Network analyses can combine analyses of frequencies and associations, where each node (word) is a frequent term and the edges (links) that connect nodes represent associations. The network graph in Figure 4 shows terms that appear frequently and indicates whether the terms are associated with each other, i.e., appear in the same tweet. Only terms that appear at least 71 times (1% of the total sample of 7,106 tweets) together are connected via an edge.

Many patterns in Figure 4 are worth mentioning. Terms that appear with many other words are clustered towards the center of the graph. Outside of 'distract', 'drive', and 'driver', only the term 'phone' appears frequently with many other words. Surprisingly, though the term 'text' is frequently found in tweets, it does not occur with many other frequent terms. However, 'text' is associated with 'crash', while 'phone' is not. On the opposite end of the spectrum, there are terms that appear with 'drive' and 'distract' and nothing else: 'way', 'road', 'biggest', 'get', 'say', 'via', 'take', and 'free'. 'Drive' appears with more terms than 'driver', leading to the conclusion that people tweet using the combination of 'driving' and 'distraction', rather than 'driver' and 'distraction'. This was confirmed by the fact that only 1,175 tweets had 'driver' immediately preceding 'distraction'.

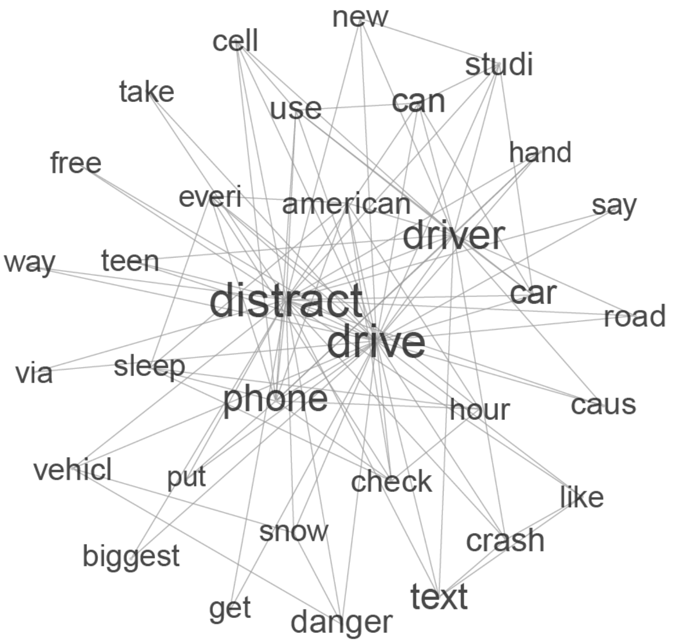

Figure 4: Network graph of frequent terms and their associations; large nodes represent frequent terms.

DISCUSSION

The purpose of this research was to use text mining tools to explore tweets about driver distraction. Text mining tools make it possible to treat tweets about driver distraction as quantitative data in terms of timelines, frequencies, and associations. Results indicated that tweets contain two types of information: (1) facts or media reports about driver distraction (2) people's personal experience with driver distraction. Many of the facts and media reports were retweeted, as was apparent given the high frequency of certain terms and their associated relationships with 'drive', 'driver', and 'driver distract'. Retweeting of this information ensures increases exposure to the same valuable information. Interestingly, two of the most frequent terms, 'phone' and 'text', relate to the use of technology and its effect on driving behavior. In addition, 'text' was directly associated to 'crash', suggesting that people associate texting with crashing. This confirms lawmakers' focus on limiting drivers' use of these devices and acknowledges the fact that introducing these laws, or at least the effect of distracted driving on car crashes, is being discussed on social media. There were also many tweets surrounding and highlighting Distracted Driving Awareness Month. Furthermore, tweets contained safe driving pledges, indicating that some people publicly show their commitment to not become distracted by cell phones while driving.

Future Work

This exploratory research uncovered many insights as to the attitudes surrounding driver distraction on social media. At the same time, there is still much to be learned. Next steps for this research include developing statistical models to describe the relationship between terms. Further analysis of these data can include examining patterns of retweets, hashtags, and tweet frequency by user. Another interesting path to pursue involves the application of clustering techniques to identify different patterns in the data. In that sense, Twitter data may complement vehicle complaint databases to understand challenges drivers face with increasingly automated vehicles (Ghazizadeh, McDonald, & Lee, 2014). In future data collection attempts, location information can be captured to determine if the topics present in tweets vary by time of day and geographic location (Xu, Bhargava, Nowak, & Zhu, 2012). Finally, data collection through the Streaming API would enhance and complement this research by providing a greater breadth of tweets about driver distraction. Data from the Streaming API could allow analysis of terms that have gained popularity over time, such as 'drivingselfie' or different combinations of terms such as 'distraction' and 'crash', that have a direct connection to driver distraction, but are not found through typical searches.

Conclusion

Analyzing tweets using text mining is a novel approach to studying driver distraction. It allows researchers to gain a quick, real-time understanding of driver opinions surrounding distraction. Twitter results can be collected and analyzed within weeks, which is a fraction of the time required for typical data collection methods such as simulator and naturalistic studies, or epidemiological data. Tweets also provide information that is typically not found through other data collection methods, like the effectiveness of campaigns such as Driver Distraction Awareness Month and how many people make safe driving pledges. As such, data from Twitter can complement and extend knowledge about driver distraction.

ACKNOWLEDGEMENTS

This work was supported by the Federal Highway Administration through a Dwight David Eisenhower Fellowship to the first author.

REFERENCES

Asur, S., & Huberman, B. A. (2010). *Predicting the Future with Social Media.* Paper presented at the Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on Web Intelligence, Toronto, Canada.

Bollen, J., Mao, H., & Pepe, A. (2011). *Modeling Public Mood and Emotion: Twitter Sentiment and Socio-Economic Phenomena.* Paper presented at the Fifth International AAAI Conference on Weblogs and Social Media.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science, 2*(1), 1-8.

Bouchet-Valet, M. (2013). SnowballC: Snowball stemmers based on the C libstemmer UTF-8 library (Version 0.5). Retrieved from http://CRAN.R-project.org/package=SnowballC

Chew, C., & Eysenbach, G. (2010). Pandemics in the Age of Twitter: Content Analysis of Tweets during the 2009 H1N1 Outbreak. *PLoS One, 5*(11), 1-13.

Diakopoulos, N. A., & Shamma, D. A. (2010). *Characterizing debate performance via aggregated twitter sentiment*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA.

Feinerer, I., & Hornik, K. (2014). tm: Text Mining Package (Version 0.5-10). Retrieved from http://CRAN.R-project.org/package=tm

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text Mining Infrastructure in R. *Journal of Statistical Software, 25*(5), 1-54.

Feldman, R., & Sanger, J. (2007). *The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data*. New York: Cambridge University Press.

Ghazizadeh, M., McDonald, A. D., & Lee, J. D. (2014). Text Mining to Decipher Free-Response Consumer Complaints: Insights From the NHTSA Vehicle Owner's Complaint Database. *Human Factors: The Journal of the Human Factors and Ergonomics Society*.

Java, A., Song, X., Finin, T., & Tseng, B. (2007). *Why we twitter: understanding microblogging usage and communities*. Paper presented at the Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, San Jose, California.

R Development Core Team. (2013). R: A Language and Environment for Statistical Computing [3.0.2]. Vienna, Austria: R Foundation for Statistical Computing.

Reimer, B., Mehler, B., Wang, Y., Mehler, A., McAnulty, H., Mckissick, E., . . . Greve, G. (2012). *An exploratory study on the impact of typeface design in a text rich user interface on off-road glance behavior*. Paper presented at the Proceedings of the 4th International Conference on Automotive User Interfaces and Interactive Vehicular Applications, Portsmouth, New Hampshire.

Rogers, C. (2012, December 5). Two Hands on the — Phone? Industry Study Looks at Driver Distraction, *Wall Street Journal*. Retrieved from http://blogs.wsj.com/drivers-seat/2012/12/05/two-hands-on-the-smartphone-industry-study-looks-at-driver-distraction/

Sakaki, T., Okazaki, M., & Matsuo, Y. (2010). *Earthquake shakes Twitter users: real-time event detection by social sensors*. Paper presented at the Proceedings of the 19th international conference on World wide web, Raleigh, North Carolina, USA.

Tumasjan, A., Sprenger, T. O., Sandner, P., & Welpe, I. W. (2010). *Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment.* Paper presented at the Fourth International AAAI Conference on Weblogs and Social Media.

Vieweg, S., Hughes, A. L., Starbird, K., & Palen, L. (2010). *Microblogging during two natural hazards events: what twitter may contribute to situational awareness*. Paper presented at the Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Atlanta, Georgia, USA.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques* (2nd ed.). San Francisco, CA: Morgan Kaufmann.

Xu, J.-M., Bhargava, A., Nowak, R., & Zhu, X. (2012). Socioscope: Spatio-temporal Signal Recovery from Social Media. In P. Flach, T. Bie & N. Cristianini (Eds.), *Machine Learning and Knowledge Discovery in Databases* (Vol. 7524, pp. 644-659): Springer Berlin Heidelberg.