LITERATURE REVIEW OF TRAFFIC SAFETY JOURNAL PAPERS USING TEXT MINING

Edward Clay, College of Engineering, California State Polytechnic University, Pomona, CA 91768, 909-749-7358, erclay@cpp.edu
Wen Cheng, College of Engineering, California State Polytechnic University, Pomona, CA 91768 909-869-2957, wcheng@cpp.edu

ABSTRACT

Text Mining is an analytical process that analyzes both term frequencies and overall topics from different texts. Despite its broad utilization in various social sciences and other similar fields, Text Mining's use has been rather constrained in traffic safety. By utilizing a number of abstracts gathered from the Traffic Research Record between 1996 and 2018, it is possible to investigate the capacities of Text Mining in terms of thorough literary analysis in traffic safety. The outcome of which will stipulate the effectiveness that Text Mining has for literary analysis in the transportation field.

Keywords: Text Mining, Literature Review, Word Correlation, Latent Dirichlet Allocation, Traffic Safety.

INTRODUCTION

Traffic safety is a major concern due to the consequential externalities related to human, social and financial loss as a result of traffic accidents. In the previous decades, there has been a substantial increase in medical costs and human fatalities on both national and international scales. The World Health Organization released a report that states in 2018, there were 1.25 million deaths and 50 million injuries from motor vehicle crashes annually [2]. This claim is solidified in a report provided by the National Highway Traffic Safety Administration in that "roadway fatalities are the leading cause of high medical expenses, financial burdens, productivity losses, and property damage" [3]. Likewise, a reported 4.6 million medicinally consulted injuries were accounted for in motor vehicle accidents whose costs totaled to $433.8 billion [4]. Due to the severe effects of traffic accidents, a tremendous number of safety studies have been directed in previous decades.

The literary analysis uncovers that the studies performed have investigated the various parts of traffic accidents such as injury severity, crash type, and crash frequency through the exploration of the various traffic safety effects including driver behavior, roadway geometry, and overall traffic characteristics. In an attempt to accurately predict the outcomes of traffic accidents, a variety of studies that rely on various statistical models including both regression models and machine learning models. Despite the fact that these research endeavors convey valuable data which helps in the advancements of traffic safety, the issue evolves into a colossal challenge when determining which statistical model is effective; therefore, for a literary analysis to be effective, it is imperative that it is performed with data that is consistently expanding in number.

Text Mining exhibits its superiority in its ability to extricate significant patterns from unstructured text documents. It is for this reason that Text Mining has been depended on to handle the categorization and

clustering of text. Since Text Mining is a relatively new process, its applications have been rather limited. What's more, none of the previous investigations applied Text Mining to writing surveys, which is necessary given the quickly expanding number of published articles. To address these confinements, the authors take advantage of applying Text Mining to an assortment of traffic safety paper abstracts gathered from the Traffic Research Record (TRR). Text Mining has shown to have different capabilities exhibited through the performance of pertinent tasks such as pairwise correlation calculations and the topic analysis. It is foreseen that through the exploration of Text Mining, some light will be shed on future literary analysis for the review of research articles.

METHODOLOGY

The goal of this experiment is to determine the effectiveness of Text Mining using a number of abstracts from TRR's articles published between 1996 and 2018. The "tidy" method is a unique way to perform Text Mining in R, being genuinely new as the package utilized for it was released in 2016. This technique will analyze term frequencies as well as topic frequencies. However, before being able to utilize the advantages of having a "tidy" dataset, the text should initially be composed into a single data frame. The data frame then needs to be singularized which allows the algorithm to differentiate between plural and singular versions of words so that multiple tenses of the same word don't offset the analysis. After which, the data must be sifted of all stop words, which are words that either grammatical errors or words that occur so commonly that do not provide content in the article such as: "to", "a", "the" etc. [6]. From here each word can be isolated for follow up examination of interest such as term counting, topic clustering, term correlations, etc. Finally, the data is then cut up into sections containing 150 words each for easier modeling and analysis.

The "tidy" method utilizes a well-known algorithm, Latent Dirichlet Allocation (LDA), for topic modeling [6]. LDA follows two fundamental standards: that each text is a blend of topics and each topic is a blend of terms. In LDA, the distribution of topics in texts pursues a dirichlet distribution, while terms of a specific topic pursue a multinomial distribution. The analysis of LDA was performed by analyzing the per-topic-per-term probabilities called β ("beta"). The contrast between different topics (state subjects 1 and 2) can be estimated by using the following log-ratio:

$$\text{Topic Difference} = \log_2\left(\frac{\beta_2}{\beta_1}\right) \tag{1}$$

Through the utilization of this log-ratio, it is discovered that the difference between topics 1 and 2 will be symmetrical. In addition, it is possible to filter out more common terms, which can be represented in any term that has a beta value greater than 0.001 in at least one topic [6].

DATA DESCRIPTION

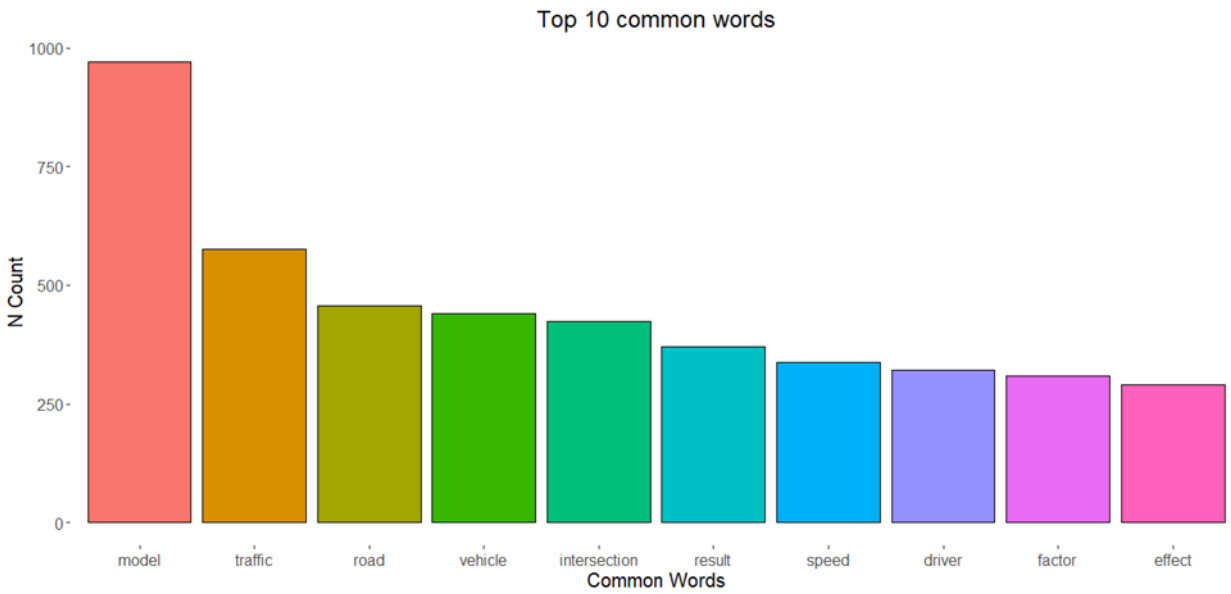
The data used in this research was gathered from TRR: The Journal of the Transportation Research Board. Readers worldwide have access to the wide scope of TRR papers that offer broad and more profound subtleties in different transportation-related characteristics, such as safety, policy, planning, etc. In this research, the abstracts of the published papers were extricated from TRR, containing 446 abstracts from published articles between 1996 to 2018. There are a number of reasons behind the utilization of Text Mining on only the papers' abstracts rather than entire papers. To begin with, abstracts are often readily available since full paper access usually requires some form of payment while the abstract itself is free to view. Additionally, abstracts act as a point of passage for any academic paper, an independent ground-

breaking articulation conveying a concise outline of an article, and along these lines, the words used are typically significant and to the point. Finally, in relation to other text sources, such as a basic paper title of the main text of papers, the assortment of the papers' abstracts act as an appropriate database size, and fits to be an ideal testbed for the generally new Text Mining procedures, before they can be applied to full papers for further analysis.

RESULTS

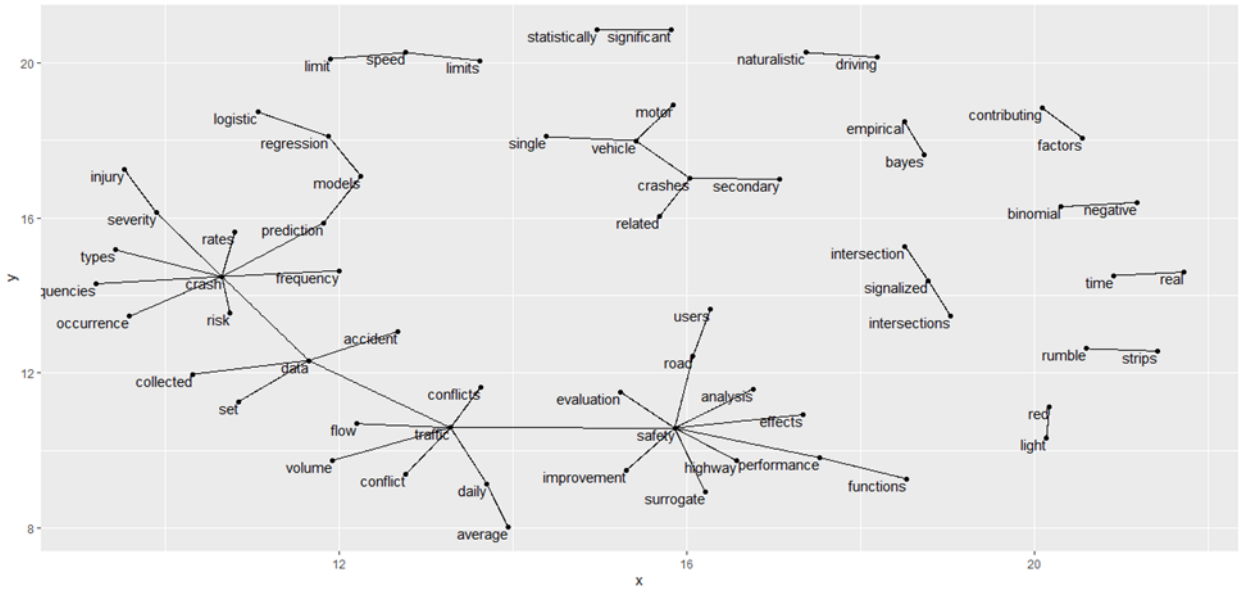
Among the most common tasks in Text, Mining is to determine term frequencies within a specific dataset. **Figure 1** shows the top 10 most commonly used terms in TRR's Abstracts from the years previously mentioned. More commonly used terms such as: "crash", "data", and "datum" were excluded to identify more insightful terms.

FIGURE 1 TOP 10 COMMON WORDS USED IN TRR ABSTRACTS



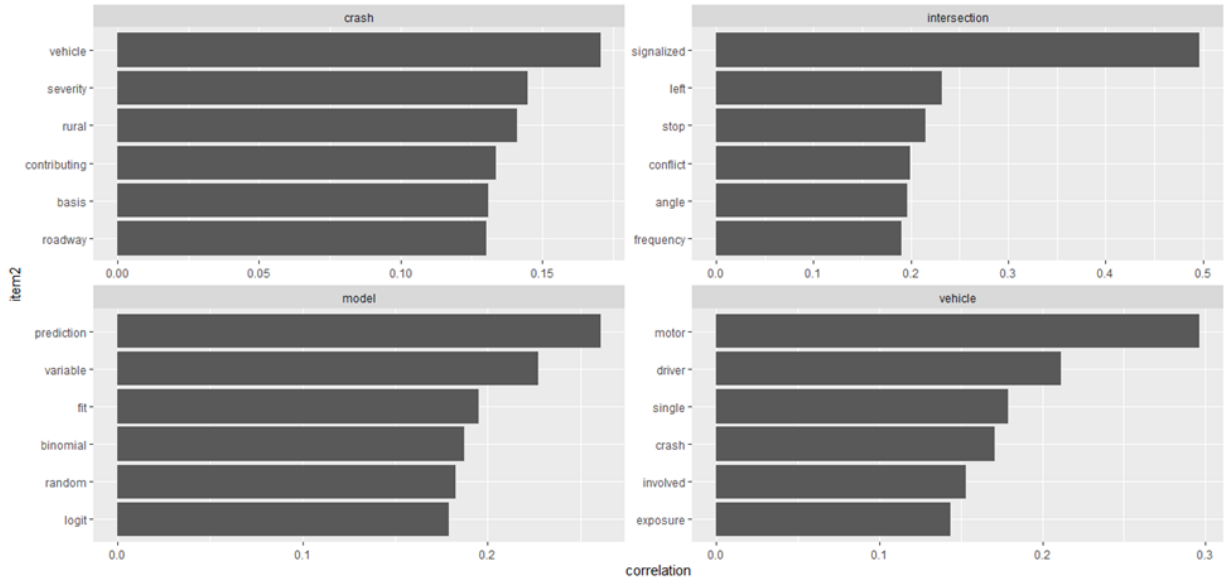
As seen in **Figure 1**, the "model" is the most utilized term in the data, with nearly 400 more instances than the second most utilized word, "traffic". This is within the author's expectations as all abstracts were gathered from papers managing safety data, investigation, and models. In the field of traffic safety, there are normally a few words that are strung together which indicate a major topic that is being researched/discussed. It is for this reason that bi-gram analysis was also performed in this study, and the most common term-pairs appear in **Figure 2**.

FIGURE 2 COMMON BIGRAMS IN SAFETY ABSTRACTS, SHOWING THOSE THAT OCCURRED MORE THAN 20 TIMES AND WHERE NEITHER WORD WAS A STOP WORD



Upon inspection of **Figure 2**, it is clear that the terms: “crash”, “data”, “traffic”, and “safety” are strung together and are the main focuses of “hubs”, which are seceded by more specific topics. There are a few term combinations that are commonly used in model analysis, such as: "negative binomial" and “empirical bayes”. It is also necessary to mention some of the other interesting discoveries that came about due to **Figure 2**. For instance, "signalized intersection" seems to draw more attention than their unsignalized counterparts; “Red lights” have also appeared to be a common topic to explore as well; “Real-time” data is clearly more utilized than most other data sources for simulated data and finally; among the different roadway highlights, “rumble strips” show up regularly in the safety research. To get a better understanding of how these terms related to one another, the authors selected four interesting terms: "crash", "intersection", "model", and "vehicle". The correlation analysis was then done based on term frequency and the results are displayed in **Figure 3**.

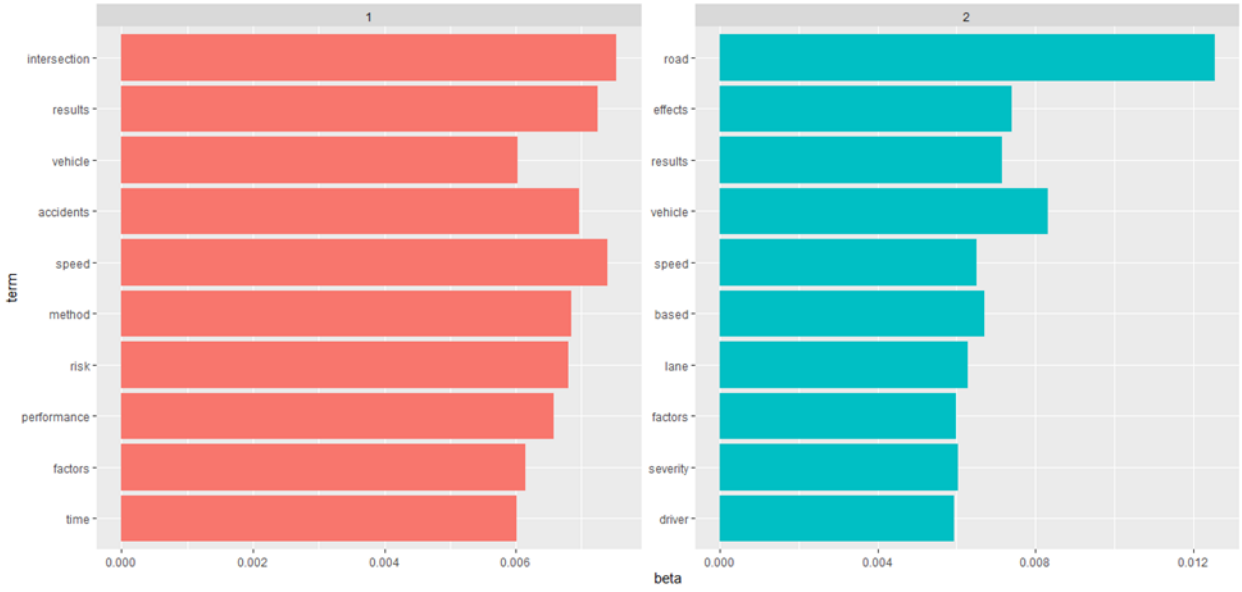
FIGURE 3 WORDS FROM SAFETY ABSTRACTS THAT WERE MOST CORRELATED WITH “CRASH”, “INTERSECTION”, “MODEL”, AND “VEHICLE”



Exploring **Figure 3** more in-depth uncovers some intriguing patterns concerning the four key-terms previously mentioned. For "intersection", one term that is commonly associated is "signalized", which lines up with the discoveries from **Figure 2**. Another term worth mentioning is "frequency", which may indicate crash frequency analysis at the intersection level [5], which is more frequently utilized than the crash rate. For "crash", the most commonly associated term is "vehicle" which as a crash component, draws significantly more attention than other terms such as "roadway" or "pedestrians". For the "model", more researchers are interested in model predictions, followed model variables and model fitness. For "vehicle", motor vehicles have been regularly explored in previous research, followed by vehicle drivers then single-vehicles. Such discoveries not only outline hot research topics in the previous two decades but also allude to which subjects may require more research efforts going onward; which perhaps incorporates unsignalized intersections, non-motorized vehicles, or multiple vehicles involved crashes. For a more comprehensible analysis of topics, a LDA point model, one of the most popular natural language processing strategies, was utilized which can not only recognize topics but assign terms to the various topics identified as well [1]. This research utilizes 2 topics to show the capacity of LDA in safety research review: "intersection" and "road".

Figure 4 showcases terms that are most common within the two topics previously mentioned. The topics "factor", "vehicle", "speed", and "results" appear in both parent topics which solidifies the advantage of topic clustering as it provides overlapping with respect to term-based probabilities. For "intersection", **Figure 4** shows that some of the most correlated topics are "factors" "vehicle", "speed", and "method", all of which most likely pertains to intersection crash analysis and relative influential factors. In the second topic, "road", the most common correlated topics include: "factor", "vehicle", "severity", and "drivers", which indicate that this topic is also highly related with the previous topic, "intersection" in terms of crash severities and other related variables. In general, it is proposed that the division of safety research regarding locations such as intersections or roadways is a fitting method for clustering research topics.

FIGURE 4 THE WORDS THAT ARE MOST COMMON WITHIN EACH TOPIC



CONCLUSIONS

In the wake of the performance of Text Mining on the abstract data used, the significant revelations including the term frequencies and correlations and word assignment in different topics are summarized as follows:

1. The bi-gram correlation relationship uncovers some basic term combinations such as Negative Binomial, Empirical Bayesian, rumble strips, and signalized intersections. Such discoveries suggest potential hot research subjects in the previous two decades.
2. LDA modeling demonstrated to be a legitimate strategy in the safety field to group topics, which exhibits itself as an extra instrument for Literature Review.

Despite the fact that this paper introduces a promising application of a new Text Mining method in traffic safety, further research is required to either affirm or debunk these conclusions. First off, the data used came solely from TRR; if additional investigations were to yield valid results, it would be useful to incorporate full papers from other well-known journals as well. Second, it would be useful to also see the least used term combinations to see which topics have been inadequately explored. Third, the "tidy" method contains multiple "versions" of LDA, future research should incorporate these different versions for further analysis. Fourth, the "tidy" method also includes a form of sentimental analysis, it would prove to be beneficial to incorporate this analysis as well to determine the author's opinion of their own analysis. Finally, the only form of topic analysis was in the form of a bi-gram analysis. In other words, other N-gram analysis might prompt distinct insightful information.

ACKNOWLEDGMENTS

We are grateful to Mr. Curtis Lee for his insightful comments on our model development.

REFERENCES

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

[2] Global status report on road safety 2018: summary. Geneva: World Health Organization; 2018 (WHO/NMH/NVI/18.20). License: CC BY-NC-SA 3.0 IGO).

[3] NHTSA, Traffic Safety Facts Annual Report Tables, https://cdan.nhtsa.gov/tsftables/tsfar.htm, accessed on June 1st, 2019.

[4] NSC, Injury Facts, https://injuryfacts.nsc.org/motor-vehicle/overview/introduction/. Accessed on May, 3, 2019.

[5] Poch, M., & Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. Journal of transportation engineering, 122(2), 105-113.

[6] Silge, J., & Robinson, D. (2017). Text mining with R: A tidy approach. " O'Reilly Media, Inc.".