


# Deep Learning and Clustering-Based Analysis of Text Narratives for Identification of Traffic Crash Severity Contributors <sup>†</sup>

Cristian Arteaga and JeeWoong Park \* 

Department of Civil and Environmental Engineering and Construction, University of Nevada, Las Vegas, NV 89154, USA; arteagas@unlv.nevada.edu

\* Correspondence: jee.park@unlv.edu

<sup>†</sup> Presented at the Second International Conference on Maintenance and Rehabilitation of Constructed Infrastructure Facilities, Honolulu, HI, USA, 16–19 August 2023.

**Abstract:** Crash narratives provide valuable information to understand traffic crashes and develop roadway safety countermeasures. However, manually reading long text narratives is time-consuming and error-prone. This study presents a deep-learning and clustering-based approach to identifying contributors to traffic crash severity in text narratives. We evaluate the approach using a dataset of narratives from Massachusetts and compare different deep-learning models for semantic similarity. The approach clusters semantically similar phrases in the narratives and provides an overview of frequent topics related to severe crashes, offering a valuable tool for roadway safety analysis and countermeasure development.

**Keywords:** crash narratives; clustering; deep learning; semantic similarity; severity contributors



**Citation:** Arteaga, C.; Park, J. Deep Learning and Clustering-Based Analysis of Text Narratives for Identification of Traffic Crash Severity Contributors. *Eng. Proc.* **2023**, *36*, 31. <https://doi.org/10.3390/engproc2023036031>

Academic Editor: Hosin (David) Lee

Published: 10 July 2023



**Copyright:** © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Understanding the factors that worsen the severity of traffic crashes is of paramount importance in prioritizing and implementing traffic safety countermeasures. Crash narratives, which describe in detail the context and circumstances of the crashes, are a valuable source of information to identify injury severity contributors. However, extracting insights from this unstructured text data is challenging, particularly when manual reading of thousands of narratives is required.

Past studies [1–3] have provided important advances in exploiting information from narratives for decision-making, but these studies have two main limitations. First, they focus on analyzing at the word level, which could suffer from ambiguous or incomplete insights. Second, their methods offer limited modeling of language semantics at the word level, which negatively impacts the quality of the revealed insights.

In a recent study, Arteaga et al. [4] proposed a method that addresses the limitations of past studies by identifying meaningful phrases (instead of individual words) from the narratives that describe potential crash severity contributors. However, for large databases of narratives, the method simply yields a large number of phrases, making the interpretation of the results challenging. To address this limitation, this study proposes a method that synthesizes numerous phrases that describe severity contributing factors in narratives, which facilitates analysis and decision-making for traffic safety.

## 2. Materials and Methods

To synthesize topics recurrently found in the narratives as correlated with severe crashes, this study integrates deep-learning techniques for semantic similarity and a clustering technique. The deep-learning techniques are Transformer-based models that provide enhanced semantic modeling capabilities by capturing interrelationships between words

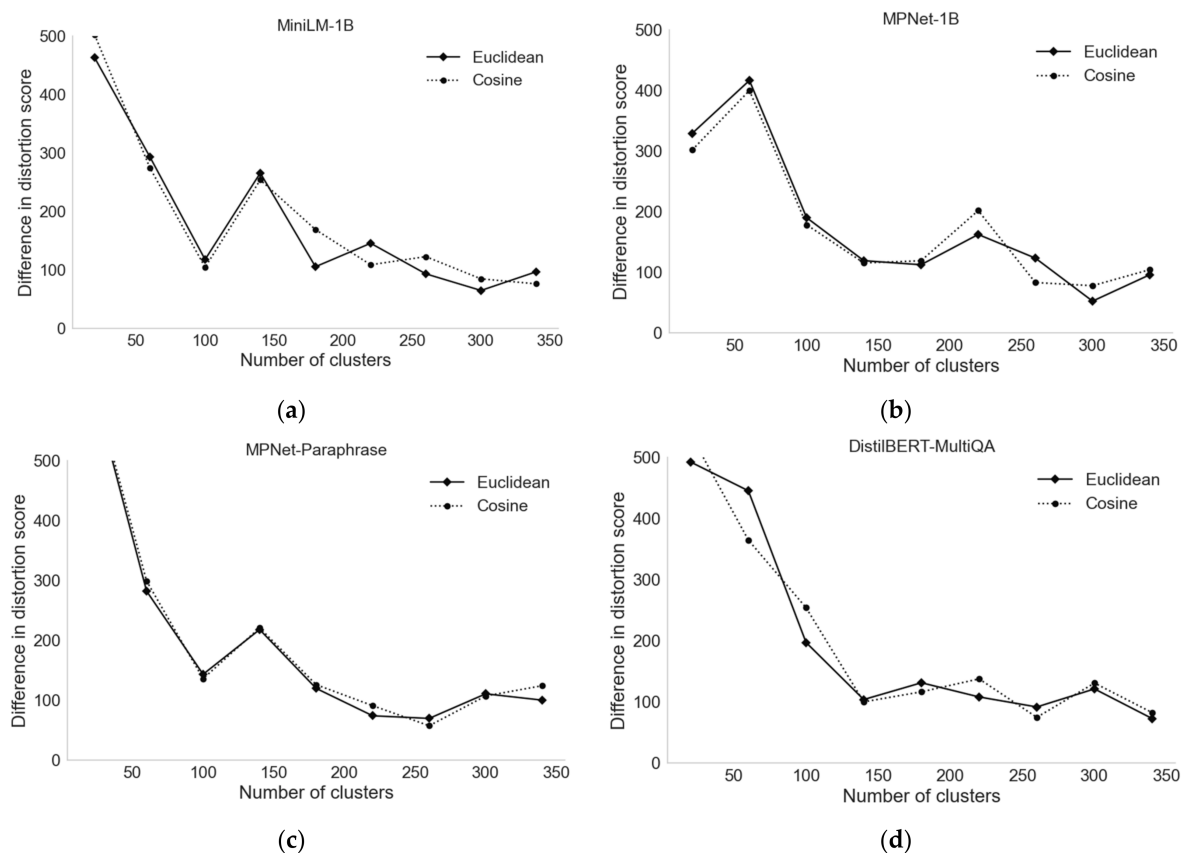
using an attention-based mechanism. Researchers [5] have pre-trained several types of Transformer-based models (e.g., MiniLM, MPNet, and DistilBERT) to excel in sentence similarity tasks by using large datasets of sentence pairs, such as the one billion sentence pairs dataset (1B), as well as datasets with paraphrasing information and question-answering sentence pairs (MultiQA). These Transformer-based models take a set of phrases as input and return a vector representation of the phrases. The vector representations enable the comparison of phrases based on their semantics (ideas expressed) instead of their individual words.

For the clustering task, this study uses the Agglomerative Clustering technique, which has shown promising results when applied in conjunction with Transformer-based models [5]. The proposed approach uses the vector representations provided by the Transformer-based models and clusters them based on their semantic similarity.

For evaluation, this study leverages a dataset of 1131 narratives provided by the Massachusetts Department of Transportation. By following the approach outlined in Arteaga et al. [4], we identified 5783 phrases correlated with severe crashes and synthesized them using the proposed approach. We evaluated four Transformer-based models for semantic similarity, two distance metrics for agglomerative clustering, and different numbers of clusters to evaluate their effect on the cluster distortion scores (the average of the squared distances from the cluster centers to each element in the clusters).

### 3. Results

Figure 1 shows the distortion scores for different numbers of clusters, similarity metrics, and Transformer-based models for semantic similarity.



**Figure 1.** Distortion scores for different numbers of clusters and deep learning models for semantic similarity: (a) MiniLM-1B; (b) MPNet-1B; (c) MPNet-Paraphrase; and (d) DistilBERT-MultiQA.

These results show that all the deep-learning models exhibit similar patterns in the decrease of distortion scores as the number of clusters increases. The plots provide an indi-

cation of the points at which increasing the number of clusters yields diminishing returns in terms of reduction of distortion scores, which is equivalent to an analysis using the elbow method. Therefore, given that the goal is to have a low distortion score while keeping the number of clusters low, the value of 100 was selected as the number of clusters, and the MiniLM-1B model was used for subsequent analysis, as this model was the fastest to reach a low distortion score without excessively increasing the number of clusters. In terms of the distance metric, the Cosine and Euclidean distances yielded similar results for distortion scores, which indicates that both metrics are equally suitable for the clustering task.

Table 1 shows examples of the clusters returned by the developed approach. The table includes information about the most common words within a cluster to provide an indication of the type of contents that a cluster captures. For instance, the words “alcohol”, “intoxicated”, “bottles”, and “marijuana” are top words in a cluster, and all the phrases are intrinsically related to the topic of driving under the influence of alcohol or drugs.

**Table 1.** Examples of clusters and phrases within clusters provided by the developed approach.

| Top Words in a Cluster                       | Number of Phrases | Examples of Phrases in a Cluster   |
|--|-------------------|--|
| EMS, hospital, transported, and ambulance    | 2191              | <ul style="list-style-type: none"> <li>“... EMS was requested, and he was later transported to the hospital for his injuries.”</li> <li>“... requested at least an ambulance.”</li> <li>“... needed medical attention.”</li> </ul>   |
| Vehicle, lane, travel, and towing            | 1392              | <ul style="list-style-type: none"> <li>“... partially in the travel lane.”</li> <li>“Wheels towing arrived on ... ”</li> <li>“... traveling west on Route 195 ... ”</li> <li>“... traveling in the middle travel lane ... ”</li> <li>“... blue hills towing ... ”</li> </ul>           |
| Scene, arrived, officer, and trooper         | 324               | <ul style="list-style-type: none"> <li>“The trooper arrived on scene ... ”</li> <li>“The scene was photographed by ... ”</li> <li>“... masters arrived on scene and ... ”</li> <li>“... arrived on scene ... ”</li> <li>“Also, on scene was ... ”</li> </ul>                           |
| Guardrail, crosswalk, pedestrian, and struck | 217               | <ul style="list-style-type: none"> <li>“A pedestrian was struck after exiting ... ”</li> <li>“... struck the guardrail, stopped and rolled ... ”</li> <li>“... communications reported a pedestrian who was hit by ... ”</li> </ul>  |
| Head on, vehicle, pole, and crash            | 101               | <ul style="list-style-type: none"> <li>“... strike the motorcycle head on and ... ”</li> <li>“... the utility pole was severed near the base.”</li> <li>“... reported head-on crash ... ”</li> <li>“... a head-on crash ... ”</li> <li>“... pole with heavy front end ... ”</li> </ul> |
| Speed, rate, high, and mph                   | 91                | <ul style="list-style-type: none"> <li>“... at a high rate of speed, presumably in ... ”</li> <li>“... very high rate of speed.”</li> <li>“... vehicle 1 began at a high rate of speed ... ”</li> <li>“... at a speed greater than 90 mph.”</li> </ul>                                 |
| Alcohol, intoxicated, bottles, and marijuana | 33                | <ul style="list-style-type: none"> <li>“... round face and she appeared intoxicated.”</li> <li>“... bottles of Smirnoff ... ”</li> <li>“... appeared to be heavily intoxicated.”</li> <li>“... was heavily intoxicated. ”</li> </ul>   |

Table 1. *Cont.*

| Top Words in a Cluster             | Number of Phrases | Examples of Phrases in a Cluster   |
|------------------------------------|-------------------|--|
| Control, swerve, lose, and lost    | 16                | <ul style="list-style-type: none"> <li>• “... unknown male lost control of ... ”</li> <li>• “... she did not want to swerve ... ”</li> <li>• “... lose control of basic ... ”</li> <li>• “... then lose control and cross ... ”</li> <li>• “... saw swerve abruptly ... ”</li> </ul> |
| Ice, weather, raining, and vehicle | 7                 | <ul style="list-style-type: none"> <li>• “... ice on the road.”</li> <li>• “... with snow-covered/icy road conditions.”</li> <li>• “... raining with low temperatures.”</li> </ul>   |

#### 4. Discussion

The results in Table 1 indicate that the proposed approach effectively clusters phrases based on their semantic contents. Some clusters contain phrases that recurrently appear in narratives for severe crashes, although they are not necessarily severity contributing factors (e.g., EMS transporting people to hospitals and vehicles being towed from the scene). However, most of the clusters identified by the developed approach provide important insights about severity contributing factors, such as the involvement of pedestrians, speeding, the influence of intoxicating liquor, head-on crashes with utility poles, suspected marijuana use, and adverse roadway conditions. These phrases provide important insights to traffic safety analysts about the factors that require urgent attention for the implementation of countermeasures. Thus, the developed approach provides traffic engineers with a valuable tool to easily exploit the information in crash narratives for data-driven decision-making.

#### 5. Conclusions

This paper presented an approach to synthesizing the results of an analysis of severity contributors in crash narratives. The developed approach addresses the limitations of past studies by integrating deep-learning-based semantic similarity and a clustering approach to provide an overview of frequent topics in the narratives associated with crashes of different severity levels (e.g., fatality, severe injury, minor injury, and property damage only). The insights returned by the approach can significantly help crash analyses as it enables the use of narratives as a valuable information source to compare and complement the insights derived from conventional statistical analyses of quantitative crash data, thereby facilitating a comprehensive diagnosis of traffic crashes. The identification of contributing factors based on the analysis of both text narratives and quantitative data can enhance analysts' confidence in the significance of such factors. Thus, by facilitating the extraction of insights from narratives, the proposed approach offers considerable value for the identification and prioritization of crash factors that need prompt attention.

**Author Contributions:** Both authors participated in the conceptualization, investigation, and writing of the paper. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by the National Cooperative Highway Research Program (NCHRP) (IDEA Project #231).

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** Data is unavailable due to privacy restrictions.

**Acknowledgments:** The authors express gratitude to MassDOT for supplying the data utilized in this investigation. The authors bear complete responsibility for the viewpoints, discoveries, and inferences presented in this work, which may not align with the positions of NCHRP or MassDOT.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Gao, L.; Wu, H. Verb-Based Text Mining of Road Crash Report. In Proceedings of the Transportation Research Board, 92nd Annual Meeting, Washington, DC, USA, 13–17 January 2013; pp. 5–16.
2. Arteaga, C.; Paz, A.; Park, J. Injury Severity on Traffic Crashes: A Text Mining with an Interpretable Machine-Learning Approach. *Saf. Sci.* **2020**, *132*, 104988. [[CrossRef](#)]
3. Rakotonirainy, A.; Chen, S.; Scott-Parker, B.; Loke, S.W.; Krishnaswamy, S. A Novel Approach to Assessing Road-Curve Crash Severity. *J. Transp. Saf. Secur.* **2015**, *7*, 358–375. [[CrossRef](#)]
4. Arteaga, C.; Park, J.; Paz, A. Enhanced Identification of Crash Severity Contributors from Text Narratives Using Natural Language Processing. In Proceedings of the 102nd Annual Meeting of the Transportation Research Board, Washintong, DC, USA, 8–12 January 2023.
5. Reimers, N.; Gurevych, I. Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *arXiv* **2019**, arXiv:1908.10084.

**Disclaimer/Publisher’s Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.