

From Police Narratives to Crash Insights Using AI: Language Models for Systematic Crash Data Analysis

Mitta Hage (mitta.hage@gmail.com), Paula Granlund (paulagranlund@hotmail.com),
Carlos M. Lima Azevedo (climaz@dtu.dk), and Mette Møller (mette@dtu.dk)

The Technical University of Denmark (DTU),

Department of Technology, Management and Economics, Intelligent Transportation Systems

Every police-reported crash in Denmark includes a written description of what happened. Compared with structured crash variables, these police narratives often provide richer descriptions of crash sequences, intended manoeuvres, interactions between road users, reporting circumstances, and other contextual details. Such information can be important when interpreting crashes, but its unstructured nature makes it difficult to analyse systematically across large crash databases. As a result, the narratives are primarily used through manual and case-specific procedures, while large-scale crash analyses rely mainly on structured crash data. This creates a challenge: a potentially valuable source of information exists within the crash database, but it remains difficult to utilise systematically across hundreds of thousands of crash reports. Consequently, important information contained in the narratives may remain inaccessible for large-scale crash data analysis.

This study investigates how language models and semantic topic modelling can support more systematic use of information contained in Danish police crash narratives. The analysis uses police-reported crash data provided by Vejdirektoratet, covering nearly 900,000 reported traffic crashes in Denmark since 1985. Two complementary approaches are applied. First, a Danish BERT model is fine-tuned to classify narratives into the official `main_crash_situation` categories. This serves as a proof of concept, examining whether the narratives contain sufficient information to recover existing structured crash classifications. Second, BERTopic is used to identify recurring patterns in the narratives that may not be fully represented by predefined crash categories.

The results show that police crash narratives contain systematic and useful information that can complement existing structured crash data. The classification model is able to recover official crash categories with high accuracy, demonstrating that the narratives contain sufficient information to reproduce much of the information already represented in structured variables. This suggests that the narratives are not merely supplementary notes, but a consistent source of information about crash circumstances that can support validation of existing registrations.

The topic modelling analysis further reveals recurring patterns and subgroups within the crash data that are difficult

to identify through structured variables alone. Examples include queue-related rear-end crashes, yield-line conflicts, hit-and-run reports, and other recurring narrative themes that cut across traditional crash categories. Several topics also provide a more detailed view of existing categories by separating crashes that share the same official classification but differ in circumstances, road-user interactions, or reporting characteristics. Together, these findings demonstrate that police narratives contain both confirmatory information that aligns with existing registrations and complementary information that can support a richer understanding of crash patterns.

The societal value of the project lies in improving the usability of existing crash data. The results demonstrate three potential stakeholder-oriented applications. First, narrative-based models can support validation by identifying possible inconsistencies between structured registrations and narrative descriptions. Second, they can support enrichment by identifying meaningful subgroups within existing crash categories. Third, they can support pattern discovery by revealing recurring themes that do not have direct equivalents in the structured crash database. Together, these applications can help analysts search, interpret, and explore large collections of crash reports more efficiently.

The project also addresses practical and ethical considerations associated with sensitive police data. All models are open-source and can be operated locally within secure research environments, avoiding the need to transfer sensitive narratives to external services. This supports responsible use of language-based methods within public-sector crash data analysis.

Overall, the study demonstrates how Danish language models and topic modelling can transform police crash narratives from an underutilised text source into a systematic complement to structured crash data. The findings suggest that police narratives should not only be viewed as supporting documentation for individual crash reports, but also as a valuable data source that can be analysed systematically alongside traditional structured crash variables. Rather than replacing expert judgement, the methods can support validation, enrichment, and pattern discovery within crash data analysis.