Pilot Study on Improving Crash Data Accuracy in Kentucky through University Collaboration

Report Number: KTC-24-17

DOI: https://doi.org/10.13023/ktc.rr.2024.17

Kentucky Transportation Center
College of Engineering, University of Kentucky, Lexington, Kentucky

in cooperation with
Kentucky Transportation Cabinet
Commonwealth of Kentucky

The Kentucky Transportation Center is committed to a policy of providing equal opportunities for al persons in recruitment, appointment, promotion, payment, training, and other employment and education practices without regard for economic, or social status and will not discriminate on the basis of race, color, ethnic origin, national origin, creed, religion, political belief, sex, sexual orientation, marital status or age.

Kentucky Transportation Center
College of Engineering, University of Kentucky, Lexington, Kentucky

in cooperation with
Kentucky Transportation Cabinet
Commonwealth of Kentucky

Kentucky Transportation Center

Kentucky Transportation Center • University of Kentucky
176 Raymond Building • Lexington, KY 40506 • 859.257.6898 • www.ktc.uky.edu

University of Kentucky

Research Report
KTC-24-17

Pilot Study on Improving Crash Data Accuracy in Kentucky through University Collaboration

Michael A. Fields
Research Scientist

Eric Green, PhD, GISP, RSP1, PE
Program Manager

Robert Kluger, PhD, PE
University of Louisville

Xu Zhang, PhD
Kentucky Transportation Center
Research Scientist

and

Kirolos Haleem, PhD, PE, PTOE, PMP
Western Kentucky University

Kentucky Transportation Center
College of Engineering
University of Kentucky
Lexington, Kentucky

In Cooperation With
Kentucky Transportation Cabinet
Commonwealth of Kentucky

February 2024

<table>
<tr><td>1. Report No.
KTC-24-17</td><td>2. Government Accession No.</td><td colspan="2">3. Recipient’s Catalog No.</td></tr>
<tr><td colspan="2">4. Title and Subtitle
Pilot Study on Improving Crash Data Accuracy in Kentucky through University Collaboration</td><td colspan="2">5. Report Date
February 2024
6. Performing Organization Code</td></tr>
<tr><td colspan="2">7. Author(s):
Michael A. Fields, Eric Green, Robert Kluger, Xu Zhang, Kirolos Haleem</td><td colspan="2">8. Performing Organization Report No.
KTC-24-17</td></tr>
<tr><td colspan="2">9. Performing Organization Name and Address
Kentucky Transportation Center
College of Engineering
University of Kentucky
Lexington, KY 40506-0281</td><td colspan="2">10. Work Unit No. (TRAIS)

11. Contract or Grant No.
SPR 22-630</td></tr>
<tr><td colspan="2">12. Sponsoring Agency Name and Address
Kentucky Transportation Cabinet
State Office Building
Frankfort, KY 40622</td><td colspan="2">13. Type of Report and Period Covered

14. Sponsoring Agency Code</td></tr>
<tr><td colspan="4">15. Supplementary Notes
Prepared in cooperation with the Kentucky Transportation Cabinet</td></tr>
<tr><td colspan="4">16. Abstract
Without high-quality crash data and robust interpretive/analytical tools to analyze these data, transportation agencies will struggle to develop evidence-based strategies for improving road safety. Crash narratives are one element of crash reports that pose especially acute interpretive challenges. These narratives supplement coded data and give an account of incidents authored by responding law enforcement officers. Despite their value, conducting manual reviews of the 150,000+ crash reports and narratives issued in Kentucky each year is not feasible. To address this challenge, reviewers examined approximately 8,000 crash narratives from calendar year 2020 using a proprietary web-based quality control tool to identify discrepancies between narratives and coded data. The most pronounced inconsistencies between coded data and narratives were found in questions related to aggressive driving, distracted driving, intersection and secondary crashes, and travel direction. Building on this exercise, researchers developed a machine learning algorithm that automatically classifies attributes in crash records based on the interpretation of unstructured narrative text. Although this model performed well, goodness-of-fit metrics showed that a Google AI Language model (Bidirectional Encoder Representations from Transformers [BERT]) was more accurate and precise as well as having better recall. Future crash data quality control efforts that incorporate machine learning applications should use BERT, however, the latest advances in AI technology need to be integrated into new applications and models as they are developed.</td></tr>
<tr><td colspan="2">17. Key Words
crash data, crash reporting accuracy, consistency, quality control tool</td><td colspan="2">18. Distribution Statement
Unlimited with approval of the Kentucky Transportation Cabinet</td></tr>
<tr><td>19. Security Classification (report)
Unclassified</td><td>20. Security Classification (this page)
Unclassified</td><td>21. No. of Pages
28</td><td>19. Security Classification (report)</td></tr>
</table>

Table of Contents

List of Figures

List of Tables

Acknowledgements

The authors would like to acknowledge the Kentucky Transportation Cabinet (KYTC) for the grant provided to conduct this research. The opinions, findings, and conclusions in this paper are those of the authors and not necessarily those of KYTC. The authors would also like to acknowledge all students and faculty who contributed to the data collection and review process in this study.

Executive Summary

An important component of Kentucky crash reports is the crash narrative written by officers who complete crash forms. Over 150,000 crash reports are generated statewide each year. Narratives written by officers are critical for crash data quality control because they explain the nature of a crash in the reporting officer's own words, often in greater detail than coded crash data. Comparing crash narratives and coded values lets data analysts uncover inconsistencies in coded values and find new information that may not be recorded in the standard report form. Given the enormous quantity of crash reports issued each year, conducting manual reviews of each report is simply not practical. To take advantage of all information in crash reports, researchers from the University of Kentucky, the University of Louisville, and Western Kentucky University collaborated on a project to develop a web-based quality control tool (QCT) to improve the efficiency of manual crash report reviews. A review of over 8,000 crash reports underpinned the research team's development of a machine learning algorithm that classifies crash records based on unstructured narrative text.

Using the QCT, reviewers evaluated crash reports to identify patterns between narratives and crash characteristics as well as discrepancies between narratives and coded data. Omitted from review were fatal crashes and property damage only (PDO) crashes. Fatal crashes undergo in-depth investigations, while narratives for PDO crashes tend to lack detail. Reviewers found it difficult to validate some coded data based on narratives along, whereas validation of other questions was straightforward. For example, reviewers found it particularly difficult to validate instances of distracted driving, while they were able to determine manner of collision in 99% of narratives (Table 5.1 provides a full comparison). Reviewers tended to identify disagreements between narratives and coded values most often on questions related to driving behaviors (e.g., aggressive driving, distracted driving). Discrepancies in interpretation arise because of the inherent subjectivity of perception. Other questions where disagreements between narratives and coded values were common included those on intersection crashes, secondary crashes, and travel direction.

Researchers leveraged knowledge from manual crash report reviews to develop a model that automates the interpretation of crash narratives. The research team took a machine learning approach, using a logistic regression model written in the Python scripting language to assess unstructured narrative text and extract crash attributes of interest. The model is designed to process narratives and predict the classification of attributes in each record. The performance of this model was compared to a model recently developed by Google AI Language researchers called BERT (Bidirectional Encoder Representations from Transformers). Based on the Transformer architecture, BERT employs self-attention mechanisms to process text sequences. Although the proprietary logistic regression model and BERT produced broadly similar results, on all goodness-of-fit metrics, BERT outperformed the logistic regression model. Future crash data quality control that incorporate machine learning applications should use BERT, however, researchers should monitor the latest advances in AI technology to identify improvements that can be incorporated into future models.

Chapter 1 Introduction

1.1 Background

An important component of Kentucky crash reports is the crash narrative written by the officer who fills out the crash form. Statewide, narratives vary significantly in the level of detail provided but are an essential part of quality control as they explain the nature of a crash in the reporting officer's own words, often in greater detail than coded crash data. With narratives, data users can identify inconsistencies in coded values and find new information that may not be recorded in the standard report form. Full-scale review of narratives is impractical due to the time required, but researchers conducted thorough review of a sample of narratives to develop an automatic text-mining application. A web-based quality control tool (QCT) was developed to make the process more efficient.

Once researchers reviewed a representative sample of narratives, they developed a machine learning operation to automatically categorize each narrative into either the inclusion class or the exclusion class (i.e., the crash either was or was not of a certain type). These classifications can be compared to traditional inclusion/exclusion flags — based on querying structured tabular data — to quantify data quality problems that result from the miscoding of certain data elements, and to identify patterns of miscoding.

To assess errors across the crash reporting process in Kentucky, researchers from the University of Kentucky, the University of Louisville, and Western Kentucky University collaborated on this project. By developing a methodology to systematically review narratives and identify contradictions between the structured tabular data and the unstructured free text in narratives, the team can provide information to other states and traffic safety agencies to improve their crash data reporting quality. This report describes the results of this study. It:

- Presents and evaluates an efficient method for manually reviewing crash narratives using the QCT
- Presents and evaluates two machine learning algorithms developed to automatically classify crash records based on the unstructured narrative text
- Discusses findings related to crash data quality from the quality control methodology.

Chapter 2 Issues in Crash Data Quality

Crash data quality is a known problem, with two main issues recognized by researchers: completeness and accuracy *(1)*. Both impact our understanding of road safety. Completeness accounts for the occurrence of a given crash record and its full attribution in a database. Accuracy refers to the correctness of stored information. Ahmed et al.'s *(2)* global investigation of crash data quality found that among high-income countries an average of ≈ 25% of crash records had an error related to location, ≈ 39% had errors in victim information, and ≈ 15% contained errors in environmental factors. Estimates had a high variance among reporting studies, highlighting how difficult it is to quantify errors.

Inaccurate data can significantly impact safety analysis *(3)*. Sources of inaccuracy are attributable to a range of issues in the reporting process, data entry, data storage standards, and more. Location inaccuracies are becoming less common, but are still present in crash data, even with widespread use of GPS technologies *(4, 5)*. Incomplete information in specific fields, such as occupant restraints, is also problematic *(6)*. Training and requiring reporting officers to evaluate challenging issues can also lead to inaccuracies in crash data. For example, officers are asked to assess the extent of an injury, which may require working knowledge outside of their expertise and is likely to lead to differences of opinions and errors *(7, 8)*.

Crash report narratives provide information against which the quality of structured tabular data can be validated, but manual review of all crash records is time prohibitive. Automatic systems to review text-based information have been implemented to screen for specific errors in crash reports *(9)* including work-zones crashes *(10)* and secondary crashes *(11)*. Similar systems have also been applied to narratives of railroad-related crashes submitted to the Federal Railroad Administration (FRA) *(12)*.

Montella et al. *(13)* found that an electronic crash form reduced data input mistakes while saving time during the response. Researchers have also applied statistical methods to the correction, input, or estimation of features of crash data directly from a database *(14)*. Others have tried to resolve data-quality issues by establishing links to other datasets that can be used for correction or validation. For example, Hosseinzadeh et al. *(15)* investigated the quality of injury data by linking them to medical records.

In Kentucky, every crash report is submitted electronically using the Kentucky Open Portal System (KYOPS). The Traffic Records Coordinating Committee (TRCC) advises the Kentucky Transportation Cabinet (KYTC) on crash data quality. Since its inception, the electronic submission system has greatly improved crash reporting accuracy, especially for crash locations *(16, 17)*. Evaluation of crash data by safety professionals has uncovered several commonly miscoded data elements. One study found that the code for a secondary crash was used correctly in just 4% of the crashes reviewed *(18)*. More recent studies found improvements in the use of the code, likely due to enhancements to KYOPS suggested by University of Kentucky researchers *(19, 20)*. However, other coding errors were observed when trying to calculate Traffic Incident Management performance measures such as clearance time.

A recent effort by Kentucky's TRCC outlined several common errors in crash record coding. Direction of travel, for instance, can be misleading because it is not implemented in a consistent manner (e.g., sometimes the direction of travel is coded before a turning movement, and sometimes after). Misinterpretations of head-on collisions *(21)*, median crossovers *(22)*, and secondary crashes *(18)* are common as well. Table 2.1 lists recommendations TRCC has advanced for KYOPS data entry improvements, officer training, and data user training.

Table 2.1 Crash Records Recommendations from Kentucky TRCC

Improvements to Crash Data Entry
1. Develop a consistent meaning for Direction of Travel to be direction before any turning movements leading to crash 2. Railway Grade Crossing: Make this an autofill field based on spatialized Highway Information System data 3. One Way: Make this an autofill field based on spatialized Highway Information System data 4. Speed Limit: Make this an autofill field based on spatialized Highway Information System data 5. Roadway Character: Make this an autofill field based on spatialized Highway Information System data 6. Roadway Type: Make this an autofill field based on spatialized Highway Information System data 7. Intersection: Make this an autofill field based off spatialized Intersection Database 8. Ramp: Make this an autofill field based on spatialized Highway Information System data 9. Make all auto filled fields capable of being overridden by officers if necessary. Create a flag in the data for fields that have been overridden to improve QC efforts later 10. Check several other fields for conflicts if the officer indicates that it is a median crossover. Confirm with the officer that this is correct if a conflict is found. 11. Integrate updates to KYTC's HIS roadway database into KYOPS on a monthly basis (currently done annually) 12. Add a checkbox field to Vehicle Information page for "work-related vehicle" 13. Update KYOPS time-related fields with a logic to account for date and time 14. Update KYOPS Manner of Collision page such that the diagrams of crash types become selectable to indicate the manner of collision 15. Update KYOPS to remove the Land Use field. This would be replaced with a simple checkbox indicator for Private Property crashes.
Improvements to Crash Data Entry Training
1. Train officers on how new Direction of Travel definition should be properly applied 2. Train officers about the new autofill fields based off the location data 3. Train officers on the conflicts for Median Crossover crashes 4. Train officers on the standard definition of Head On collisions (cannot be a single unit) 5. Train officers on newly added "work-related vehicle" field 6. Promote and make more widely available the KYOPS/CRASH data dictionary to data creators 7. Train officers on changes to time and date field entry 8. Support a research project to systematically review a subset of crash reports to identify common discrepancies found between crash narratives and date entered into fields 9. Train officers on updates to Manner of Collision indicator that would allow clicking of the crash diagram to indicate the Manner of Collision 10. Update KYOPS Pedestrian Factors field to read "Non-Motorized Factors" and train officers on the difference 11. Train officers on updates to Land Use field 12. Train officers on addition of Private Property indicator checkbox 13. Support a research project to develop a "Narrative Wizard" that would assist officers in writing Crash Narratives by incorporating data already entered into the fields 14. Train officers on standard definition of Secondary Collision 15. Train officers on Time fields to improve accuracy

Improvements to Crash Data User Training
1. Train data users on how new Direction of Travel definition should be properly interpreted 2. Train data users on how certain fields will now be filled in automatically 3. Train data users on how HIS roadway data may change over time and why regular updates are necessary 4. Train data users on how to identify crashes mistakenly identified as Median Crossover 5. Train data users on how QC flags can be used to improve HIS 6. Train data users on how to identify crashes mistakenly identified as Head On 7. Train data users on newly added “work-related vehicle” field 8. Promote and make more widely available the KYOPS/CRASH data dictionary to data users 9. Support a research project to systematically review a subset of crash reports to identify common discrepancies found between crash narratives and date entered into fields 10. Update KYOPS Pedestrian Factors field to read “Non-Motorized Factors” and train data users on the difference 11. Train data users on removal Land Use field 12. Train data users on addition of Private Property indicator checkbox 13. Support a research project to develop a “Narrative Wizard” that would assist officers in writing Crash Narratives by incorporating data already entered into the fields. 14. Train data users on standard definition of Secondary Collision 15. Train data users on problems with Time fields and why incident clearance times are sometimes negative

Chapter 3 Manual Review of Crash Narratives

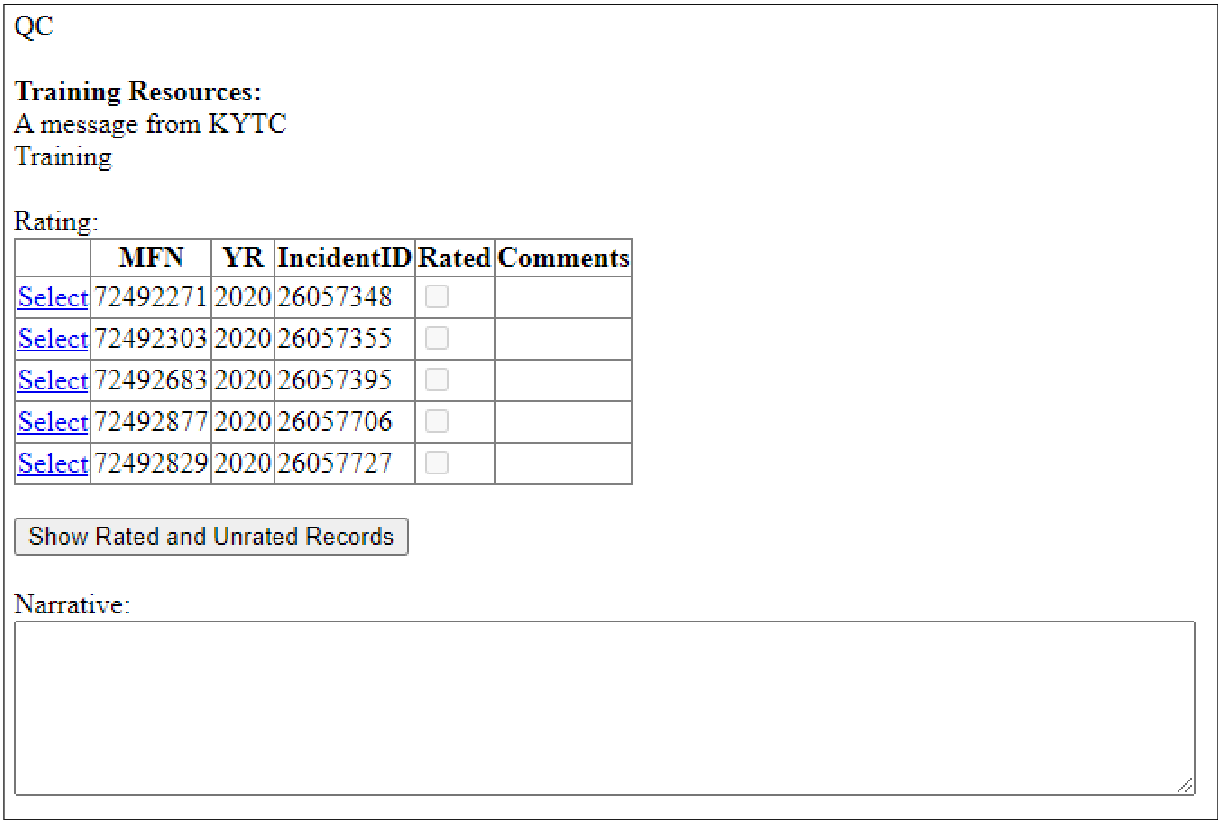
3.1 Quality Control Web-Based System and Use

Crash narratives authored by police officers are important components of Kentucky crash reports. The level of detail provided in narratives varies significantly but is an essential part of quality control. Narratives explain the nature of a crash in the reporting officer's own words and usually contain more detail than crash codes. With narratives, data users can identify inconsistencies between coded values and find new information that may not be recorded in the standard report form. While manual review is necessary to systematically use narratives, it is impractical to review every narrative due to time and labor constraints. To resolve this problem the research team developed an automatic text-mining application that can be used to conduct full-scale narrative reviews on an annual basis. In creating the app, researchers:

- Manually reviewed crash reports to identify patterns between narratives and specific crash characteristics
- Developed a machine learning algorithm to automatically classify crash records based on unstructured narrative text
- Compared machine learning outputs to structured tabular data to conduct data quality control

To efficiently review many narratives and coding errors, researchers developed a web-based quality control tool (QCT). After completing training, and with permission from KYTC, reviewers were given secure access to crash narrative data via unique login credentials. After completing each review, reviewers submitted their assessment, and the QCT stored data in a secure SQL database housed at the University of Kentucky. The QCT lets multiple users review crashes simultaneously and tracks the time spent on each record, which allowed the research team to track review efficiency.

To avoid duplication of effort, researchers assigned each crash record to a unique reviewer. For each record, the QCT displays the narrative text and asks the reviewer to assess 20 crash attributes. Most questions are answered with one of four responses: *Yes, No, Maybe,* and *Unknown*. Figure 3.1 shows the upper portion of the web interface. The QCT lets users select a crash record from a table of records assigned to them. The grid view displays which crashes have been rated and allows the reviewer to edit a rated crash. Rated crashes can be hidden, making the grid view easier to navigate. After selecting a crash to review, text populates in the *Narrative* box and presents a series of 20 questions (including comments) about the crash based on the officer's written account. Questions address the 20 crash attributes with known and suspected issues in the Kentucky Crash Database and are used to gather some auxiliary information not typically collected at the scene.



QC

Training Resources:
A message from KYTC
Training

Rating:

	MFN	YR	IncidentID	Rated	Comments
Select	72492271	2020	26057348		
Select	72492303	2020	26057355		
Select	72492683	2020	26057395		
Select	72492877	2020	26057706		
Select	72492829	2020	26057727		

Show Rated and Unrated Records

Narrative:

Figure 3.1 Upper Portion of the Quality Control Tool Web Interface

Figure 3.2 captures the first 11 questions users answer for each crash. They focus on direction of travel for the first 5 units in the crash and include yes/no questions on commonly mischaracterized crash attributes. Reviewers have four answer options for these questions: *Yes, No, Maybe,* and *Unknown*. To avoid bias in responses, reviewers did not receive coded values recorded by the officer (except manner of collision, which they were asked to confirm in Question 11). *Yes, No,* and *Maybe* indicate that the narrative explicitly provides evidence related to the question, while *Unknown* indicates the narratives does not mention evidence related to the question. All responses default to *Unknown*.

Please indicate the direction of travel for the first 5 units (be sure to indicate the unit's initial direction).

Q1 - Unit 1:
○N ○S ○E ○W ◉Unknown
Q2 - Unit 2:
○N ○S ○E ○W ◉Unknown
Q3 - Unit 3:
○N ○S ○E ○W ◉Unknown
Q4 - Unit 4:
○N ○S ○E ○W ◉Unknown
Q5 - Unit 5:
○N ○S ○E ○W ◉Unknown

Q6 - Was this a median crossover crash?
○Yes ○Maybe ○No ◉Unknown
Q7 - Was this crash on a ramp?
○Yes ○Maybe ○No ◉Unknown
Q8 - Was the crash a roadway departure?
○Yes ○Maybe ○No ◉Unknown
Q9 - Was this an intersection crash?
○Yes ○Maybe ○No ◉Unknown
Q10 - Was the crash in a parking lot?
○Yes ○Maybe ○No ◉Unknown
Q11 - Does the following statement describe the crash?
--please select a crash in the table above--
○Yes ○Maybe ○No ◉Unknown

Figure 3.2 Questions 1 – 11 of the Quality Control Tool Web Interface

The final section of the QCT interface investigates other commonly miscoded attributes, unrecorded attributes, and provides space for comments (Figure 3.3). Questions 12 – 16, disable the *No* answer option as it is unlikely a reviewer can respond accurately based on the narrative alone. For example, if the narrative does not mention if a driver is operating their vehicle as part of their work responsibilities, it is still unclear if the driver was on duty. A response of *Unknown* is preferred in these cases. Questions 17 – 19 are conditional and highlighted in green if the crash meets one of three conditions:

- Crash is coded as a secondary crash
- Time of crash notification is later than the time the crash scene was cleared, resulting in a negative clearance time
- A wrong-way driving action was coded.

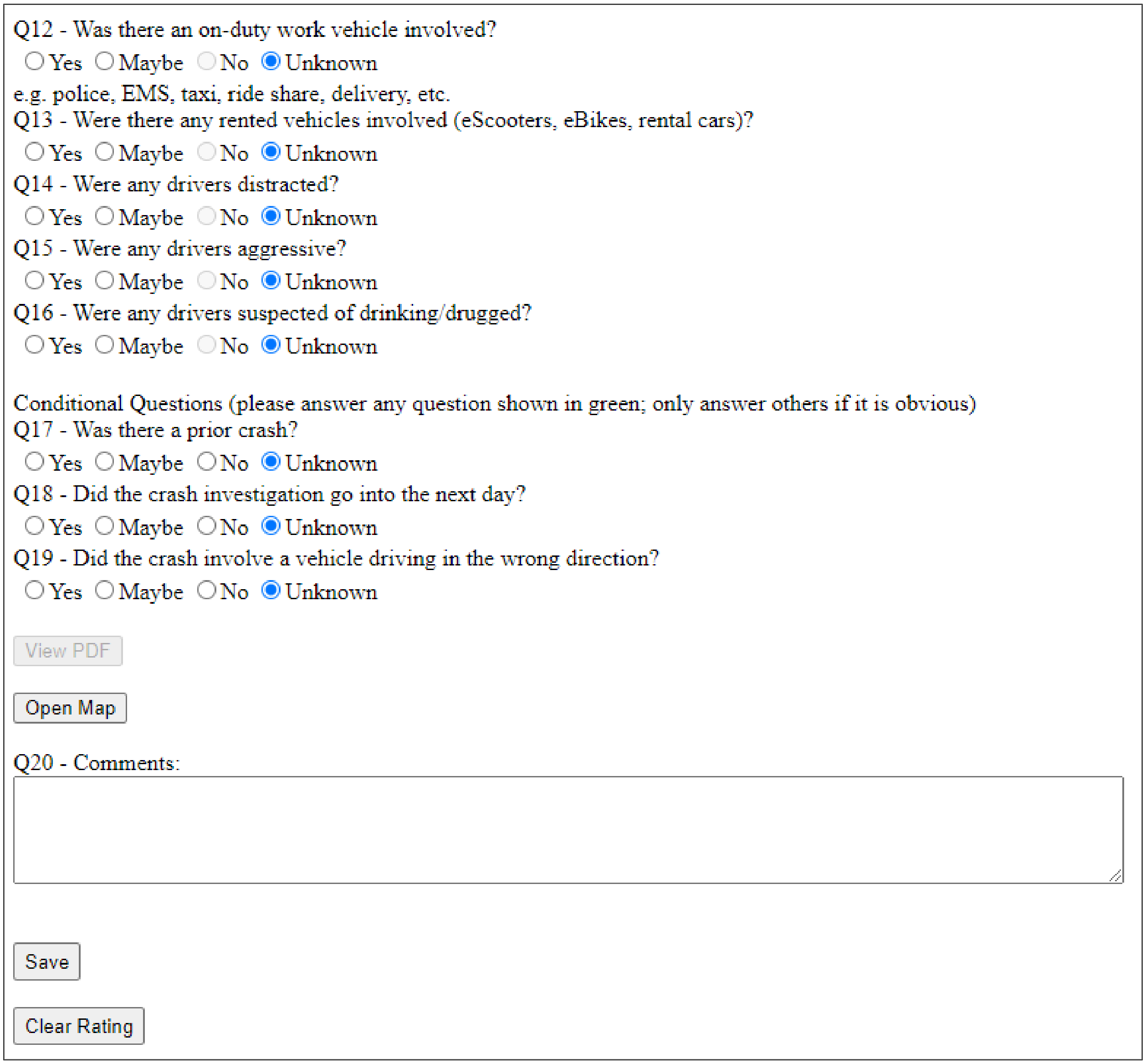


Figure 3.3 Questions 12 – 20 of the Quality Control Tool Web Interface

Reviewers do not have to answer conditional questions unless the relevant condition is met and an answer can be discerned from the crash narrative.

The QCT has a button reviewers can click on to view a map to the crash location if they want to verify information that is unclear in the narrative, such as roadway elements or travel directions. Another button opens an image of the full crash report. However, this button was disabled for data reviewers who participated in this study. Its function is for the project team to quickly access to the original report data and compare reviewer output with tabular data. Question 20 is a comment box where users could leave notes for the project team. Reviewers were instructed to add a comment any time they selected *Maybe* as a response. Users can save ratings or clear the form with the last two buttons under the *Comments* box.

Chapter 4 Data Collection

4.1 Reviewer Training

The research team trained student reviewers on crash types with known issues identified by the TRCC. To understand the process of reviewing crash narratives, student reviewers watched four videos:

1. Introduction to Crash Data and Reporting — Provides an overview of all aspects of the reporting process and introduced key fields in the crash report including direction of travel, crash types, severity, driver behaviors (e.g., impairment, aggression, distraction), and other important information relating to the questions listed in the QCT.
2. Crash Data QCT Workflow — Describes a step-by-step process for accessing the QCT and reviewing records.
3. How to Review Narratives — Shows reviewers examples of narrative reviews and responses to questions.
4. Responsibility and Ethics in Data Review — Reviews rules associated with maintaining data privacy and rules for working with data.

Each student reviewer passed Institutional Review Board (IRB) training courses and completed a Memorandum of Understanding (MOU) as part of the onboarding process.

4.2 QCT Testing

Seven students from the University of Kentucky (UK), University of Louisville (UofL), and Western Kentucky University (WKU) served as reviewers. After being trained, student reviewers, as well as a UK Department of Civil Engineering faculty member with extensive highway safety knowledge, analyzed 100 crash narratives to validate the QCT. The goal of these reviews was to determine how well reviewers understood questions, identify issues with the QCT and database processes, and ensure consistency across reviewers. The sample of 100 narratives contained narratives whose level of complexity varied significantly. As described in Chapter 5, following this exercise researchers adjusted review procedures.

4.3 QCT Implementation

Included in the sample from which narratives were drawn were all calendar year 2020 crashes except fatal *K-level* and property damage only (PDO) *O-level* crashes. Fatal crashes undergo a thorough investigation and include a more detailed narrative. PDO crashes represent a large majority of the annual crash distribution, but their narratives usually contain a low level of detail. Both were excluded to make the review process more efficient. The resulting crash database included ≈ 20,000 injury crashes.

Chapter 5 Results

5.1 QCT Validation Summary

Reviewers were initially assigned 100 of the most recent crashes from the database. After the review process began, researchers determined that the KY FARS team had reclassified 31 crashes from fatal to non-fatal based on the nature of the crash (e.g., suicide, heart attack). The research team concluded the remaining 69 crashes were sufficient to evaluate the training exercises.

Figure 5.1 illustrates the agreement percentage between pairs of reviewers.[1] The cell that intersects the names of reviewers indicates the level of agreement. For example, the level of agreement between UKY Student 1 and UofL Student 1 was 84%. Although student reviewers were broadly consistent in their levels of agreement, there tended to be less agreement between them and the faculty reviewer. Further analysis of individual answers could identify the source(s) of disagreements.

Reviewer	UKY Student 1	UofL Student 1	WKU Student 1	WKU Student 2	WKU Student 3	UofL Student 2	WKU Student 4	UKY Faculty 1
UKY Student 1	100%							
UofL Student 1	84%	100%						
WKU Student 1	85%	92%	100%					
WKU Student 2	86%	82%	85%	100%				
WKU Student 3	85%	93%	92%	83%	100%			
UofL Student 2	86%	87%	89%	86%	89%	100%		
WKU Student 4	84%	90%	90%	84%	89%	85%	100%	
UKY Faculty 1	77%	76%	77%	84%	75%	78%	77%	100%

Figure 5.1 Agreement Between Reviews

Table 5.1 shows the average time reviewers spent on each report. Excessively long review times were removed from the dataset since they were likely due to reviewers taking breaks but not logging out from the QCT. The faculty reviewer spent less time than students (mean of 1.57 minutes for each report). Mean review times varied between 1.7 and 3.51 minutes per report among student reviewers, with an overall average of 2.8 minutes.

[1] Agreement percentage was calculated as 100 × number of questions with same answers from the two reviewers / 18 questions per crash / 69 crashes.

Table 5.1 Average Report Manual Review Times

Reviewer	Review Time (min/report)
UKY Faculty 1	1.57
UKY Student 1	3.51
UofL Student 1	2.62
WKU Student 1	2.60
WKU Student 2	1.70
WKU Student 3	3.21
UofL Student 2	3.37
WKU Student 4	2.99
Student Average	**2.80**

For each question, Table 5.2 lists the number of times reviewers input a response other than *Unknown*. That is, a reviewer was able to provide an answer based on a narrative's content. What is most revealing about Table 5.2 is the numerical range. For example, on the question that asks about the Direction of Travel for Unit 1, UofL Student 1 gave an answer of *N*, *S*, *E*, or *W* 48 times, while WKU Student responded similarly 59 times. This is a range is quite dramatic considering the question is straightforward. Researchers used this information to reinforce some training procedures.

Table 5.2 Number of Times Reviewers Provided a Response Other than *Unknown* to Each Question

Question	UKY Student 1	UofL Student 1	WKU Student 1	WKU Student 2	WKU Student 3	UofL Student 2	WKU Student 4	UKY Faculty 1
Travel Direction Unit 1	53	48	51	59	51	52	51	44
Travel Direction Unit 2	34	31	32	36	33	34	31	28
Travel Direction Unit 3	6	6	6	6	7	8	6	6
Travel Direction Unit 4	1	1	1	1	1	1	1	1
Travel Direction Unit 5	0	0	0	0	0	0	1	0
Aggressive Driving	7	4	1	0	4	4	6	1
Distracted Driving	14	1	4	4	5	11	3	6
Impaired Driving	13	10	11	10	11	12	10	8

Question	UKY Student 1	UofL Student 1	WKU Student 1	WKU Student 2	WKU Student 3	UofL Student 2	WKU Student 4	UKY Faculty 1
Head On Collision	0	0	0	0	0	0	0	0
Intersection Crash	66	69	69	69	66	68	67	61
Manner of Collision	68	69	69	69	68	69	68	62
Median Crossover	65	69	68	69	67	63	68	62
Negative Clearance Time	35	6	19	37	6	17	18	66
Non-Traditional Unit	3	1	6	5	6	5	3	5
Parking Lot Crash	67	68	69	68	66	67	66	59
Ramp Crash	65	69	69	69	67	66	68	62
Roadway Departure	65	69	69	68	62	64	66	61
Secondary Crash	39	1	4	66	9	38	0	63
Work Related Crash	2	4	1	3	3	3	1	2
Wrong Way Crash	40	0	17	66	7	31	39	62

For each question, Figure 5.2 indicates how far each reviewer deviated from average response. Deviations for each question were calculated by finding the absolute value of the difference between a reviewer's submitted value and the average of all submitted values. Researchers calculated the mean deviation by averaging the deviation values of all applicable crashes. The null value was used if a reviewer determined it was not applicable to a crash.

High deviation values for some questions were due to very few answers submitted by an individual reviewer. For example, the deviation of Secondary Crash from UofL Student #1 is 1, which is based on only one crash (see Table 5.3 If we ignore those with very few answers, questions such as Roadway Departure, Distracted Driving, and Non-Traditional units have relatively high deviations, suggesting there might have been confusion among reviewers as to what should count toward those crash types. The research team held follow-up discussions with reviewers to ensure the method was applied consistently.

Question	UKY Student 1	UofL Student 1	WKU Student 1	WKU Student 2	WKU Student 3	UofL Student 2	WKU Student 4	UKY Faculty 1
Travel Direction Unit 1	0.15	0.03	0.05	0.11	0.07	0.06	0.09	0.14
Travel Direction Unit 2	0.02	0.02	0.03	0.05	0.06	0.03	0.05	0.11
Travel Direction Unit 3	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.15
Travel Direction Unit 4	0	0	0	0	0	0	0	0
Travel Direction Unit 5							0	
Aggressive Driving	0.06	0.10	0		0.10	0.15	0.10	0
Distracted Driving	0.09	0	0.15	0.13	0.22	0.33	0.28	0.06
Impaired Driving	0.09	0.09	0.05	0.11	0.05	0.09	0.04	0.18
Head On Collision								
Intersection Crash	0.24	0.14	0.16	0.22	0.13	0.22	0.22	0.27
Manner of Collision	0.17	0.20	0.11	0.12	0.12	0.12	0.20	0.16
Median Crossover	0.08	0.06	0.07	0.12	0.07	0.07	0.14	0.22
Negative Clearance Time	0.20	0.12	0.25	0.19	0.11	0.08	0.15	0.22
Non-Traditional Unit	0.17	0.00	0.42	0.10	0.25	0.30	0.08	0.05
Parking Lot Crash	0.12	0.03	0.05	0.06	0.05	0.04	0.05	0.04
Ramp Crash	0.04	0.04	0.05	0.05	0.05	0.02	0.05	0.04
Roadway Departure	0.26	0.18	0.20	0.30	0.18	0.25	0.21	0.19
Secondary Crash	0.03	1.00	0.25	0.05	0.00	0.05		0.05
Work Related Crash	0	0.25	0.50	0	0.17	0.17	0	0.25
Wrong Way Crash	0.09		0.13	0.14	0.05	0.29	0.15	0.19

Figure 5.2 Average Deviation to the Mean by Question and Reviewers

5.2 Analysis of Full Kentucky Crash Dataset

Reviewers evaluated ≈ 8,000 crash narratives. Table 5.3 summarizes responses by question. If a reviewer gave an answer of *Unknown*, researchers treated it as unanswered. The table's last column lists the percentages of responses that were *Yes*, *No*, or *Maybe*. Some issues proved difficult to validate based solely on narratives, while for other questions the narrative provided a straightforward question. For example, only about 7% of narratives indicate distracted driving, whereas the manner of collision was discernable in 99% the narratives. In 350 records, the narrative did not match the manner of collision. This would flag these crashes for further analysis.

Table 5.3 Question Response Rates

Question	Yes	Maybe	No	Unknown	Proportion Answered
Median Crossover Crash	95	17	5,815	1,920	76%
Work-Related Driver*	129	21	-	7,697	2%
Intersection Crash	2,673	53	3,601	1,520	81%
Crash on Ramp	126	25	5,793	1,903	76%
Roadway Departure Crash	3,275	50	3,357	1,165	85%
Non-Traditional Unit*	5	7	-	7,835	0%
Distracted Driver*	538	31	-	7,278	7%
Aggressive Driver*	162	14	-	7,671	2%
Parking Related Crash	165	21	5,775	1,886	76%
Crash Matched Manner of Collision	7,240	194	350	63	99%
Driver Suspected of Drinking	561	15	-	7,271	7%
Secondary Crash**	248	40	622	6,937	12%
Crash Duration Spanned Over Midnight**	119	35	434	7,259	7%
Wrong Way Driving Crash**	174	6	629	7,038	10%

* *No* was not allowed for this question as it is unreasonable to determine

** These questions were only required if the crash was flagged ahead of time based on the attributes of the crash

Table 5.4 compares reviewer interpretations to KYOPS-reported codes. *Agree* indicates that the reviewer agreed with the designation in the report; *Disagree* indicates the reviewer found evidence that conflicts with coded values. *Unclear* means the narrative did not provide sufficient information for the reviewer to ascertain information related to the question. *Null* values recorded in the database were null in the KYOPS database.

Questions with the lowest levels of agreement are those related to behaviors. The aggressive driving question was answered for 161 crashes narratives, and reviewers disagreed with the report more than half the time. Distracted driving had a similarly high disagreement rate of 35%. Discrepancies in interpretation arise because perceptions of behaviors are subjective. Providing additional definition or explanation would be valuable. Other questions with high disagreement rates include those on intersection crashes (22%), secondary crashes (26%), and travel directions (8% – 15%).

Table 5.4 Levels of Agreement Between Narratives and Coded Values

	Agree	Disagree	Unclear From Narrative	Null Record In Database	Total	Percent Agreement $\left(\frac{\text{Agree}}{\text{Agree+Disagree}}\right)$
Crash Locations						
Intersection Crash	4,896	1,373	1,572		7,841	**78.1%**
Ramp Crash	5,829	85	1,927		7,841	**98.6%**
Parking Lot Crash	5,863	72	1,906		7,841	**98.8%**
Driving Behaviors						
Distracted Driving	351	187	7,303		7,841	**65.2%**
Aggressive Driving	80	81	7,680		7,841	**49.7%**
Impaired Driving	456	105	7,280		7,841	**81.3%**
Crash Types						
Median Crossover	5,757	148	1,936		7,841	**97.5%**
Roadway Departure	5,539	1,088	1,214		7,841	**83.6%**
Secondary Crash	648	222	6,971		7,841	**74.5%**
Manner of Collision	7,234	350	257		7,841	**95.4%**
Travel Direction						
Travel Direction Unit 1	4,820	840	1,856	325	7,841	**85.2%**
Travel Direction Unit 2	3,492	406	1,024	2,919	7,841	**89.6%**
Travel Direction Unit 3	409	36	208	7,188	7,841	**91.9%**
Travel Direction Unit 4	82	6	42	7,711	7,841	**93.2%**
Travel Direction Unit 5	22	2	6	7,811	7,841	**91.7%**

5.3 Narrative Text Mining

The research team developed and tested two models that automate narrative interpretation and classify each record into one of two classes — inclusion or exclusion — based on the attribute of interest. For both models, the study sample consisted of all calendar year 2020 crash records and the results of manual narrative reviews .

First, researchers developed a targeted, machine learning approach using a logistic regression model written in the Python scripting language to assess unstructured narrative text and extract crash attributes of interest. This effort entailed a two-step process:

- Zhang et al.'s (11) model to identify secondary crashes using the Natural Language Toolkit Python library was adapted to identify other crash attributes.
- A process was designed to incorporate the logic of the training/testing model and process narratives and predict the classification of all attributes in each record.

For the crash narratives reviewed manually, each crash attribute was coded as 1 for inclusion or 0 for exclusion. For example, a narrative that stated the driver appeared intoxicated (and had been coded as *Yes* in the manual review) was coded *1* for the impaired driving attribute. If a narrative omitted language suggesting impairment (*No* in the manual review) it was coded 0. That was used as the input dataset for the machine learning model. Eighty percent of the input dataset served as the training set. The other 20% of records made up the testing set, which was used to evaluate the model's logic and quantify its level of agreement with the input dataset.

Understanding the effects of model parameters is important to maximize the model's predictive power and develop a replicable process that can be applied to subsequent years of crash data. Researchers conducted a series of model runs to test the effects of adjusting model parameters. The model uses a hyperparameter, *C,* to inform selection of internal parameters not set by the user. Higher values of *C* instruct the model to give more weight to the training data, suggesting it is a strong representation of the population, which may lead to overfitting. Lower values of *C* indicate more complexity in the data, allowing for exclusion of extreme internal parameter values and thus leading to smoothing of regression model outputs. Several *C* values between 1 and 100 were tested, with a value of 30 producing the most accurate results.

When the model assesses each narrative to predict its classification it can use individual words or contiguous multi-word phrases as the unit of analysis. Using 1-or-2-word phrases as the unit of analysis resulted in the most accurate predictions. The model's accuracy was evaluated using four goodness-of-fit indicators:

- Accuracy, based on the percentage of correctly classified records among all records (true negatives and true positives divided by total number of cases)
- Precision, based on the percentage of correct positive classifications among the model predicted positives (true positives divided by true positives and false positives)
- Recall, based on the percentage of correct positive classifications among predefined positives (true positives divided by true positives and false negatives)
- F1 score, based on a combination of recall and precision (Recall x Precision divided by their sum and multiplied by 2)

To evaluate the model's sensitivity, researchers compared its performance to a second, recently developed model by researchers at Google AI Language called BERT (Bidirectional Encoder Representations from Transformers). BERT is based on the Transformer architecture, which employs self-attention mechanisms to process input text sequences. Unlike previous language models, which processed words or phrases independently or in a single direction, BERT uses bidirectional features to consider words before and after each target word during the encoding process, allowing it to effectively capture contextual relations between words. The model was pre-trained on an extensive corpus of open-source text data from the internet to predict missing words in sentences based on contextual information provided by surrounding words. This lets the model learn underlying patterns in text, which can be used for a range of text mining tasks (e.g., text classification). Since BERT is pre-trained, it must be calibrated using a specific dataset to fine tune weights and enhance performance.

Researchers tested the BERT model using the Transformers Python library developed by the Hugging Face community. The model was calibrated on a subset of crash attributes using 90% of each dataset, with the other 10% set aside for testing. Included below are goodness-of-fit metrics for secondary crashes and impaired driving crashes (Figure 5.4 and 5.5, respectively). Across all goodness-of- fit metrics, BERT outperformed the logistic regression text narrative mining model. The research team's efforts demonstrated the effectiveness of text mining operations for reading and interpreting unstructured text narratives.

Machine learning models produce keywords showing positive and negative correlations with the narrative text. Each keyword has a coefficient indicating the strength of the correlation. Positive coefficients indicate keywords correlate positively with a narrative flagged as including the attribute of interest. Conversely, negative coefficients indicate the presence of the given keyword suggests the narrative should not be flagged for inclusion. In both cases, larger coefficients indicate a higher likelihood of the keyword's inclusion or exclusion. Figure 5.3 shows the positively correlated keywords output from the logistic regression model to identify impaired driving crashes.

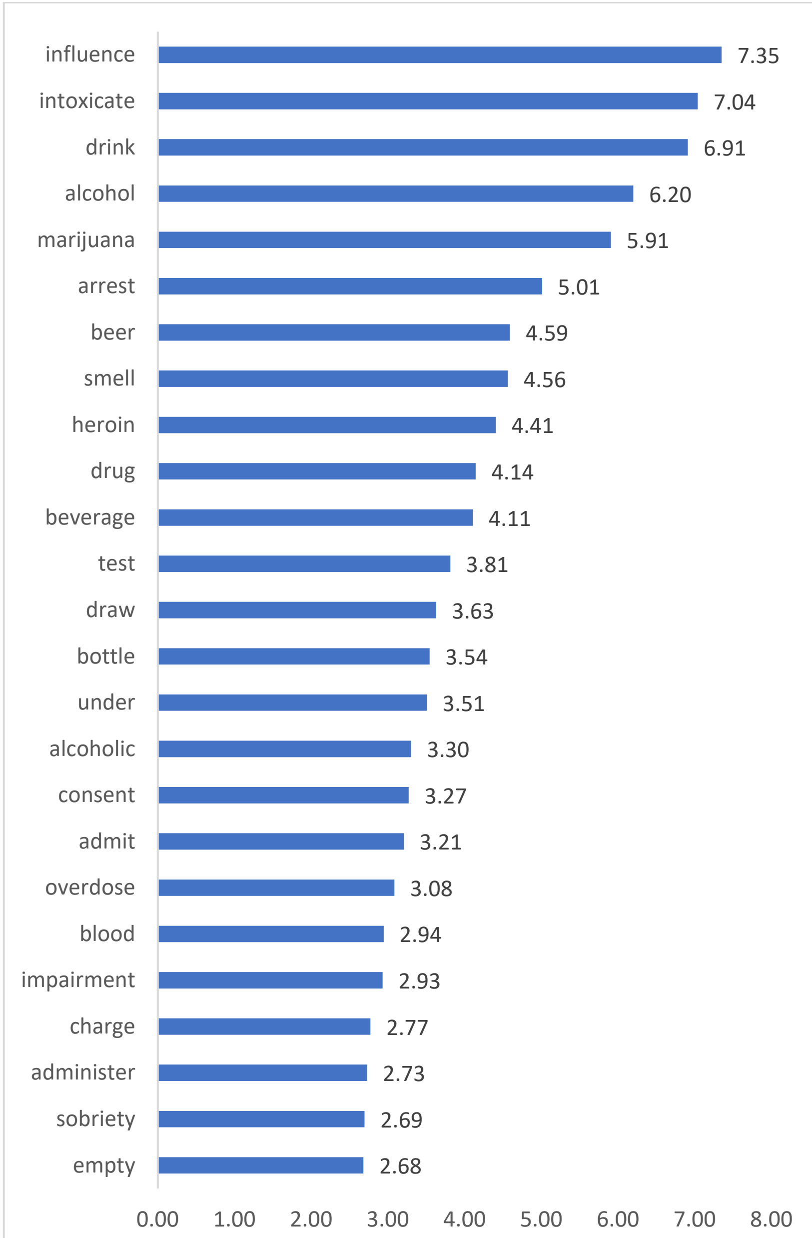


Figure 5.3 Logistic Regression Coefficients for Impaired Driving Crashes

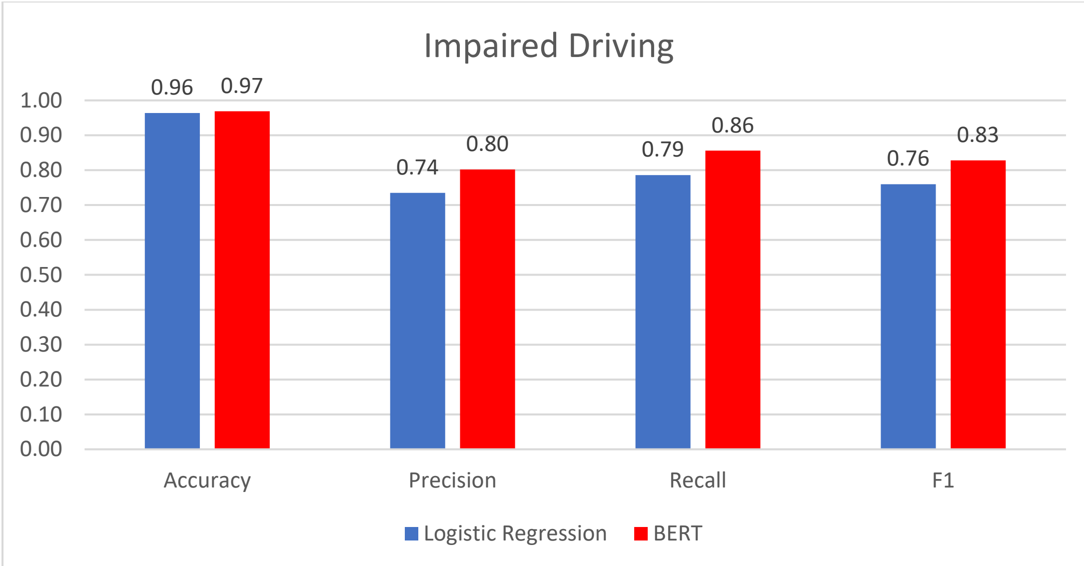


Figure 5.4 Comparison of Goodness-of-Fit Metrics Between Models – Impaired Driving

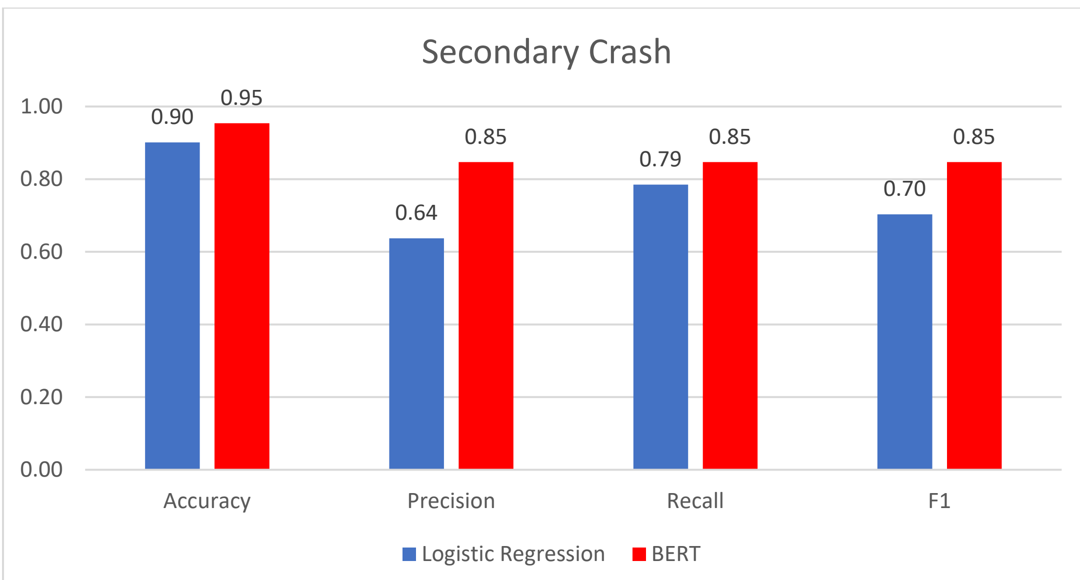


Figure 5.5 Comparison of Goodness-of-Fit Metrics Between Models – Secondary Crash

Chapter 6 Conclusions

6.1 Quality Control Tool

Manual reviews of crash narratives strengthen crash data quality control by helping reveal inconsistencies, improve knowledge of how officers articulate crash characteristics, and inform text mining approaches. For this study, researchers reviewed Kentucky crash narratives using the QCT to identify inconsistencies in 19 key questions identified as important by the TRCC. Initial QCT testing showed that reviewers interpret narratives inconsistently. Reviews uncovered possible systematic issues in the Kentucky crash database that must be investigated further to determine the underlying causes of inconsistencies, including fields related to behaviors, intersection designations, secondary crash flags, and travel direction.

Because reporting protocols, training systems, database structures, and other external factors impact crash data over time, adopting the QCT (or comparable tool) for manual reviews offers several advantages. For new fields on a report, the web-based QCT can easily be adjusted to ask different questions and provides a mechanism for evaluating the fidelity of data. It can also be used to quickly access narratives to create and validate text-based modeling methods.

While the QCT can flag potential issues, using it to correct information in crash records is not advisable due to variance in reviewer understandings of narratives and the limited presentation of information in narratives. Findings should be used to inform training protocols, adjust aspects of reporting, and support policy decisions to improve reporting consistency statewide. Another benefit of the QCT is that manual reviews can help improve machine learning techniques by providing training data that can lead to more consistent reviews.

Agencies seeking a method to improve crash data quality must understand data collection processes implemented by local law enforcement and devise questions to estimate where errors are likely to occur. Crash data users are the best resource to lean on when devising such questions. Manual labor is required to implement the QCT. In this project civil engineering students reviewed crash narratives. Students were eager to learn about the topics and received hands-on education working with crash data. This improved their knowledge of crash datasets and data literacy, increasing their awareness of common problems with data collection.

6.2 Text Mining

Crash narratives contain valuable information that is sometimes missing from structured tabular data yet is critical for making informed safety decisions. Because reading the 150,000+ crash narratives Kentucky generates each year is impractical, researchers developed an automated text mining process to interpret this vital information. Incorporating information from narratives revealed discrepancies between narrative details and structured crash reports. While potential reasons for discrepancies were not the focus of this study, some are likely due to the inflexibility of structured tabular crash data. Including the law enforcement officer's contextual details from crash narratives may improve classification of crash attributes, especially for more ambiguous attributes.

Comparison of model output keywords from the logistic regression model and BERT showed similarities across methods, but across all goodness-of-fit metrics, BERT outperformed the logistic regression model. As such, future quality control approaches that incorporate machine learning applications should use BERT. One caveat is that AI-based methodologies evolve and improve rapidly. Future efforts to adapt Kentucky's crash data quality procedures should identify and integrate appropriate advances in AI technology at the time of implementation.

References

(1) M. Imprialou and M. Quddus, "Crash data quality for road safety research: Current state and future directions," Accident Analysis & Prevention, vol. 130, pp. 84-90, 2019.

(2) A. Ahmed, A. F. M. Sadullah, and A. S. Yahya, "Errors in accident data, its types, causes and methods of rectification-analysis of the literature," Accident Analysis & Prevention, vol. 130, pp. 3-21, 2019.

(3) A. M. Iqbal et al., "Assessment of Crash Location Accuracy in Electronic Crash Reporting Systems," Transportation Research Record, vol. 2674, no. 9, pp. 311-323, Jul. 2020.

(4) W. A. Sarasua, J. H. Ogle, and K. Geoghegan, "Use of Global Positioning System to Identify Crash Locations in South Carolina," Transportation Research Record, vol. 2064, no. 1, pp. 43-50, Jan. 2008.

(5) A. J. Graettinger, T. W. Rushing, and J. McFadden, "Evaluation of Inexpensive Global Positioning System Units to Improve Crash Location Data," Transportation Research Record, vol. 1746, no. 1, pp. 94-101, Jan. 2001.

(6) L. E. Bilston and J. Brown, "Accuracy of medical and ambulance record restraint and crash data information for child occupants," Injury Prevention, vol. 14, no. 1, pp. 46 LP - 50, Feb. 2008.

(7) N. N. Ferenchak and R. B. Osofsky, "Police-reported pedestrian crash matching and injury severity misclassification by body region in New Mexico, USA," Accid. Anal. Prev., vol. 167, p. 106573, 2022.

(8) B. Burdett, Z. Li, A. R. Bill, and D. A. Noyce, "Accuracy of Injury Severity Ratings on Police Crash Reports," Transp. Res. Rec., vol. 2516, no. 1, pp. 58-67, Jan. 2015.

(9) M. S. Gerber and L. Tang, "Automatic quality control of transportation reports using statistical language processing," IEEE Transactions on Intelligent Transportation Systems, vol. 14, no. 4, pp. 1681-1689, 2013.

(10) M. A. Sayed, X. Qin, R. J. Kate, D. M. Anisuzzaman, and Z. Yu, "Identification and analysis of misclassified work-zone crashes using text mining techniques," Accident Analysis & Prevention, vol. 159, p. 106211, 2021.

(11) X. Zhang, E. Green, M. Chen, and R. R. (Reg) Souleyrette, "Identifying secondary crashes using text mining techniques," Journal of Transportation Safety & Security, vol. 12, no. 10, pp. 1338-1358, Nov. 2020.

(12) M. Heidarysafa, K. Kowsari, L. Barnes, and D. Brown, "Analysis of Railway Accidents' Narratives Using Deep Learning," in 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 2018, pp. 1446-1453.

(13) A. Montella, S. Chiaradonna, G. Criscuolo, and S. De Martino, "Development and evaluation of a web-based software for crash data collection, processing and analysis," Accident Analysis & Prevention, vol. 130, pp. 108-116, 2019.

(14) A. P. Tarko, J. Thomaz, and D. Grant, "Probabilistic Determination of Crash Locations in a Road Network with Imperfect Data," Transportation Research Record vol. 2102, no. 1, pp. 76-84, Jan. 2009.

(15) A. Hosseinzadeh, A. Karimpour, R. Kluger, and R. Orthober, "Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records," Journal of Safety Research, vol. 81, pp. 21-35, 2022.

(16) E. Green and K. Agent, "Evaluation of the Locations of Kentucky's Traffic Crash Data" (2010). Kentucky Transportation Center Research Report. 47. https://uknowledge.uky.edu/ktc_researchreports/47

(17) E. Green and K. Agent, "Evaluation of the Accuracy of GPS as a Method of Locating Traffic Collisions" (2004). Kentucky Transportation Center Research Report. 213. https://uknowledge.uky.edu/ktc_researchreports/213

(18) J. Pigman., E. Green, and J. Walton, "Identification of Secondary Crashes and Recommended Countermeasures" (2011). Kentucky Transportation Center Research Report. 25. https://uknowledge.uky.edu/ktc_researchreports/25

(19) R. Souleyrette, M. Chen, X. Zhang, E. Green, and S. Sagar, "Improving the Quality of Traffic Records for Traffic Incident Management" (2018). Kentucky Transportation Center Research Report. 1621. https://uknowledge.uky.edu/ktc_researchreports/1621

(20) X. Zhang, R. Souleyrette, A. Wang, P. Ross, E. Green, and M. Chen, "Traffic Incident Management Dashboard" (2020). Kentucky Transportation Center Research Report. 1715. https://uknowledge.uky.edu/ktc_researchreports/1715

(21) J. Smith, M. Hosseinpour, R. Mains, N. Hummel, and K. Haleem. Investigating Head-On Crash Severity Involving Commercial Motor Vehicles in Kentucky. Transportation Research Record. 2021;2675(10):133-147. doi:10.1177/03611981211010803

(22) E.R. Green, M.A. Fields. "A Methodology to Prioritize the Locations of Cable Barrier Installations in Kentucky", Roadway Safety and Simulation Conference, Orlando, FL; 2015.