Special Issue: Text Mining and Natural Language Processing Applied to Risk Analysis, Maintenance and Safety

Institution of MECHANICAL ENGINEERS

Journal of RISK AND RELIABILITY

Proc IMechE Part O: J Risk and Reliability
1–15
© IMechE 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1748006X221140196
journals.sagepub.com/home/pio
SAGE

A framework based on Natural Language Processing and Machine Learning for the classification of the severity of road accidents from reports

Dario Valcamonico[1], **Piero Baraldi**[1], **Francesco Amigoni**[2] **and Enrico Zio**[1,3]

Abstract

Road safety analysis is typically performed by domain experts on the basis of the information contained in accident reports. The main challenges are the difficulty of considering a large number of reports in textual form and the subjectivity of the expert judgments contained in reports. This work develops a framework based on the combination of Natural Language Processing (NLP) and Machine Learning (ML) for the automatic classification of accidents with the final aim of assisting experts in performing road safety analyses. Two different models for the representation of the textual reports (Hierarchical Dirichlet Processes (HDPs) and Doc2vec) and three ML-based classifiers (Artificial Neural Networks (ANNs), Decision Trees (DTs) and Random Forests (RFs)) are compared. The framework is applied to a repository of road accident reports provided by the US National Highway Traffic Safety Administration. The best trade-off between accuracy of the classification and explainability of the obtained results is achieved by combining HDP topic modeling and RF classification.

Keywords

Road safety, road accident reports, Natural Language Processing, Hierarchical Dirichlet Process, Doc2Vec, Artificial Neural Networks, Decision Tree, Random Forest

Date received: 13 October 2021; accepted: 2 November 2022

Introduction

Although road safety has improved with a reduction of the number of yearly fatalities, road accidents remain the eighth leading cause of death worldwide, with the number of deaths peaking at 1.35 million people in 2018.[1] The primary objective of road safety analysis is the identification of the factors influencing accident severity and frequency, and it is typically performed by experts using large repositories of reports of road accidents, which contain textual descriptions of the accidents and the results of post-event investigations. The main challenges encountered in the analysis are: (1) the poor quality and inhomogeneity of the textual content of the reports,[2] (2) the subjectivity of the police officers who wrote the reports,[3] and (3) the large number of accident reports to be considered.[4]

Natural Language Processing (NLP) offers the possibility of transforming the textual reports into a set of numerical features, which can be used as input of machine learning methods for the extraction of knowledge. This work explores the possibility of supporting the identification of the factors influencing accident severity and frequency by developing a classifier to automatically assign the accidents to the correct class of severity.[5] Despite many methods based on NLP techniques have been proposed for the classification of accident reports (Section 1.1), they tend to lack interpretability and/or they heavily rely on external

[1] Energy Department, Politecnico di Milano, Milan, Italy
[2] Department of Electronics, Information and Bioengineering (DEIB), Politecnico di Milano, Milan, Italy
[3] MINES Paris-PSL, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France

Corresponding author:
Piero Baraldi, Energy Department, Politecnico di Milano, Via Ponzio 34/3, Milan, 20127, Italy.
Email: piero.baraldi@polimi.it

supervision for the extraction of the factors influencing the classification of the accident severity.

In view of the above, the objective of the present work is to develop a NLP framework for the classification of accident reports, which provides an interpretable representation of the text, and, therefore, can be used in support to road safety analysis. To this aim, the problem of classifying the severity of the accident reports has been decomposed into the two sequential problems of: (1) obtaining a transparent representation of the accident reports in the form of numerical vectors, and (2) developing a Machine Learning (ML)-based classifier of the vectors.

With respect to (1), we consider two possible NLP approaches: (a) a topic modeling technique based on Hierarchical Dirichlet Processes (HDPs),[6] and (b) Doc2Vec.[7] Topic modeling extracts topics, that is, distributions of words, which can be thought of as themes or concepts, from a repository of textual documents, and, then, transforms each document into a vector whose elements are a measure of how well the document is represented by the corresponding topics.[8] Given the semantic meaning associated to the topics, topic modeling techniques are expected to meet the requirements of providing an interpretable representation of the accident reports. Specifically, HDP has been considered since, differently from traditional topic modeling techniques, such as of Latent Semantic Analysis (LSA)[9] and Latent Dirichlet Allocation (LDA),[10] it does not assume independence among the topic distributions and it does not require to a-priori specify the number of topics. Doc2Vec has been considered as alternative to topic modeling since it represents the accident reports as vectors of fixed dimension in a space in which semantically similar documents are mapped in dense clusters.[7] Also, it has been shown able to provide satisfactory classification performances when combined with an empirical classifier.[11] Other NLP techniques, such as Bidirectional Encoder Representation from Transformers (BERT)[12] have not been considered in this work, despite they have shown superior accuracy in the classification of textual documents in many applications.[13] This is motivated by the difficulty of obtaining a semantic interpretation of the features that they extract.

With respect to (2), once numerical representations of reports have been obtained, they can be classified in classes homogeneous with respect to the severity of the accident consequences using empirical models trained on repositories of prelabelled reports. In this work, we consider Artificial Neural Networks (ANNs), Decision Tree (DT), and Random Forest (RF) classifiers. ANNs are selected due to their robustness and good performance in classification problems,[14–16] DTs due to the interpretability of the model that they develop and the associated small computational efforts, and RFs due to the satisfactory trade-off that they can offer between computational efforts, interpretability and performance.[17,18]

This work is an extension of,[19] where a method combining HDP and ANNs has been proposed for the classification of accident reports and validated on a repository of synthetic reports. The proposed framework has been applied to a repository of real accident reports provided by the US National Highway Traffic Safety Administration (NHTSA),[20] containing the narrative of crash accidents recorded by police officers and the corresponding assessment of the severity of the consequences.

The main novel contributions of this work are:

(a) The development and comparison of different combinations of text representation and ML-based classification models for the classification of accident reports for safety analysis, considering also the interpretability of the report representation;
(b) the analysis of the interpretability of the developed models with respect to the identification of the factors influencing accident severity.

The remaining of the work is organized as follows. In Section 1.1 a literature review on NLP applications in classification problems is presented. In Section 2, the problem of classifying the reports is stated. In Section 3 the developed framework is described. In Section 4, the case study is introduced and the obtained results are presented. Finally, in Section 5, conclusions and future works are discussed.

NLP for accident classification

Text classification is a largely investigated field of study in NLP with applications ranging from sentiment analysis to named entity recognition.[21] Yang et al.,[22] an approach combining LSA and Convolutional Neural Networks (CNN) is adopted for the classification of textual maintenance records, with the final objective of developing a stochastic multi-stage model of the degradation of excavator components used in the mining industry. Guimarães et al.,[23] Principal Component Analysis (PCA) is combined with k-means clustering for grouping occupational accident reports. Bezerra et al.,[24] a deep learning approach based on Bidirectional Encoder Representations from Transformers (BERT) is applied to occupational accidents reports of a hydropower company to model whether a given type of injury is expected to cause a leave of the employee. Zhang et al.,[25] Term Frequency Inverse Document Frequency (TFIDF) is combined with five different ML-classifiers (Support Vector Machine (SVM), Linear Regression (LR), K-Nearest neighbor (KNN), Decision Tree (DT) and Naïve Bayes (NB)) and an ensemble model based on Sequential Quadratic Programming (SQP) is developed to identify the cause of accidents in construction accident reports. Heidarysafa et al.,[26] different combination of Word2Vec embedding and three different deep learning classifiers are compared for the classification of causes of railways accidents. Zhang,[27] Word2Vec is combined with a hybrid structured neural network based on CNNs and Bidirectional Long Short Term Memory neural networks (BDLSTM) to classify construction site accidents. Martinčić-Ipšić et al.[11] different combinations of Bag of Words (BoWs), Word2Vec and Doc2Vec embeddings and Random Forest (RF) classifiers are compared for the classification of benchmark NLP datasets. Rane and Kumar [28] Doc2Vec is combined with seven different ML-classifiers (DT, RF, SVM, KNN, LR, Gaussian NB and AdaBoost) for the analysis of customer feedbacks provided by the US Airline Service. Despite the satisfactory performances of these classification approaches, they do not provide an interpretable representation of the semantic context of the text, which is fundamental in road safety analysis to identify the factors influencing the accidents severity.

Some recent works have applied NLP techniques to accident reports with objectives different from the classification of accident severity. Sarkar et al.,[29] a NLP technique based on the analysis of the frequency of words used in occupational accident reports is developed for the identification of events initiators of accidents in steel plants. Williams and Betak,[30] LSA and LDA are applied to railroad accident reports to identify common trends in the failure of train equipment. Kwayu et al.,[31] Structural Topic Modeling (STM) is combined with network topology analysis to discover accident factors in road crash narratives, and for the classification of road accident reports. Limsettho et al.,[32] the performances of LDA and HDP in the classification of software bug reports are compared considering a wrapper classifiers based on DT, LR and NB. Tahvili et al.,[33] Doc2Vec is combined with two clustering algorithms (Hierarchical Density Based Spatial Clustering of Applications with Noise (HDBSCAN) and Fuzzy C-Mean (FCM)) to automatically discover functional dependencies in on-board train control systems. Ansaldi et al.[34] an ontology has been defined considering safety documents and applied to the analysis of equipment aging in a liquid fuel depot of an industrial establishment. Macêdo et al.,[35] BERT is combined with information coming from risk assessment documentation and pre-hazard analysis spreadsheets to identify risk features and potential hazards in O&G refineries. Bin et al.,[36] a NLP technique based on text chains is developed to extract fault features from accident reports of high-speed trains, with the objective of maintenance improvement. These works show that the integration of a representation model of the text with information coming from the system can facilitate the identification of the factors influencing accident severity and, therefore, is of great interest for safety analysis in various sectors.

Problem statement

We consider a repository (corpus in the NLP technical jargon[37]) of D road accident reports, $R=\{d_i, i=1,\ldots,D\}$. A label, $l_i \in \{0,\ldots,L\}$, is associated by a domain expert to each report, $d_i, i=1,\ldots,D$, according to the severity of its consequences where 0 and L correspond to the least and most impactful consequences, respectively.

The objective of this work is to develop a framework to automatically assign the class label l to a new test accident report d. Since the final aim is to support road safety analysis for the identification of the critical factors with respect to accident severity, the classification method is required to provide interpretable results, that is, to explain the reasons for which a report is assigned to a given class. For this reason, we develop a framework based on a combination of: (1) NLP methods for the representation of the reports $\{d_i, i=1,\ldots,D\}$ in numerical vectors $\{\boldsymbol{\gamma}_i, i=1,\ldots,D\}$ with an associated semantic interpretation of the textual information and (2) ML-based classification models for assigning the vectors $\{\boldsymbol{\gamma}_i, i=1,\ldots,D\}$ to the class of severity of the accident $\{l_i, i=1,\ldots,D\}$. Different modeling options will be evaluated considering the level of interpretability of the developed methods and the classification performances, which are measured using the metric of accuracy:

$$A = \frac{Number\ of\ correctly\ classified\ reports}{Number\ of\ reports} \tag{1}$$

In the case of interest of this work, characterized by two classes of severity ($L=1$) with very different number of reports since events of high severity are typically rarer than events of moderate severity, the $F_{measure}$ is also considered as performance metric[38]:

$$F_{measure} = \frac{2}{\frac{TP+FN}{TP} + \frac{TP+FP}{TP}} \tag{2}$$

where TP, TN, FP, and FN are the numbers of true positive, true negative, false positive and false negative classifications, respectively.

Framework

The framework proposed for the classification of reports combines (Figure 1):

(1) text preprocessing, which extracts the dictionary of the repository, $\Delta = \{t_j, j=1,\ldots,T\}$, made by T tokens, and converts a report d in the list of the tokens it is formed of;
(2) text representation, which transforms the preprocessed textual report, $\tilde{d}$, into a numerical vector, $\boldsymbol{\gamma}$, representing the semantic content of the reports;
(3) vector classification, which receives in input the numerical vector $\boldsymbol{\gamma}$ and provides in output the class of severity of the accident.

Sections 3.1, 3.2, and 3.3 describe the framework.

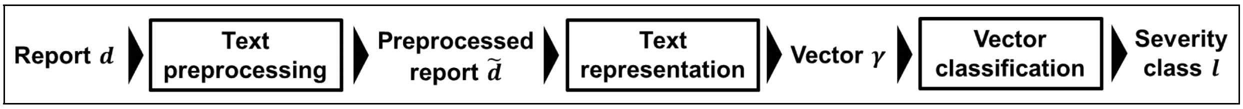

Figure 1. Developed framework.

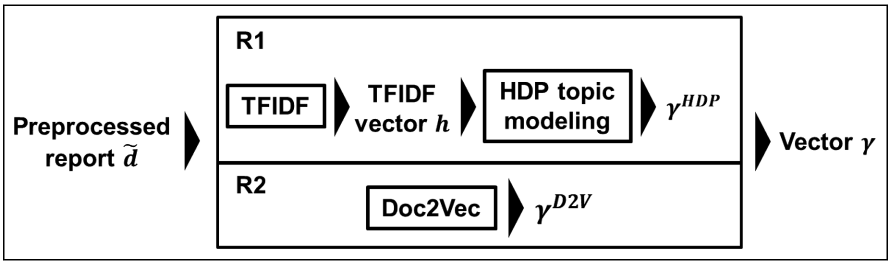

Figure 2. Text representation.

Text preprocessing

Text preprocessing is performed by:

(a) identifying the list of unique words forming all reports;
(b) cleaning the text from stop words, such as articles and prepositions, road and street proper names, car brands, vehicle serial numbers, date of the accident and generic words such as "*road*" or "*vehicle*," which do not provide useful semantic information;
(c) reducing the words to their base forms by applying a lemmatization algorithm. In this work, the Python library Gensim[39] is used;
(d) substituting each word preceded by the negation words "*no*" and "*not*" with the tokens "*no_word*" and "*not_word*," respectively; this step allows accounting for negative sentences and improve the semantic interpretation of the results. Notice that words formed by the combination of negative prefixes, that is, "*un*" or "*im*," and the root words are automatically considered in (a) as unique words, different from their corresponding root words;
(e) identifying the set, B, of frequent pairs of contiguous words, called bigrams in the NLP technical jargon. A generic bigram u_{pq}, is formed by the union of two single words, called unigrams, u_p and u_q, with $p \neq q$. The identification of B is performed by applying a procedure which associates to u_{pq} the score[40]:

$$s_{pq} = \frac{c(p,q) - c_{min}}{c(p)c(q)} \quad (3)$$

where $c(p)$ and $c(q)$ are the counts of the number of reports in which the unigrams u_p and u_q occur, respectively, $c(p,q)$ is the count of the number of reports in which unigrams u_m and u_n occur contiguously, and c_{min} is a parameter used to establish the minimum number of times a pair of contiguous words should appear in the reports to constitute a bigram. The list of frequent bigrams B is found by considering only the bigrams with s_{pq} larger than a preset threshold s_{thresh}, which allows controlling the total number of bigrams in the dictionary. In this work, the analysis of n-grams is limited to bigrams since n-grams of higher order (e.g. trigrams) tend to be repeated less often in the reports, and their identification and inclusion in the dictionary would require large computational cost.[40]

The dictionary Δ of the corpus is defined as the list of tokens $\{t_j : t_j \in U \cup B, j = 1, \dots, T\}$ obtained by the union of the list U of the preprocessed unique words and the list B of the selected bigrams At the end of this text processing stage a generic report d_i is converted into the list $\tilde{d}_i$ of the tokens it is formed of.

Text representation

Two different alternatives are considered for the representation of the preprocessed report $\tilde{d}$ as a numerical vector $\boldsymbol{\gamma}$: $R1$) the combination of Term Frequency Inverse Document Frequency (TFIDF) and HDP topic modeling, and $R2$) Doc2Vec (Figure 2).

TFIDF and HDP. The preprocessed reports $\{\tilde{d}_i, i = 1, \dots, D\}$ of the repository are firstly transformed into the vectors $\{\boldsymbol{h}_i, i = 1, \dots, D\}$, whose generic element h_{ij} measures the semantic importance of the token t_j in the preprocessed document $\tilde{d}_i$, by applying the TFIDF weighting procedure[41]:

$$h_{ij} = tf_{ij} \ \ log\left(\frac{D}{df_j}\right) \quad (4)$$

where tf_{ij} is the number of times the token t_j occurs in $\tilde{d}_i$ and df_j is the number of preprocessed documents in which the token t_j occurs. The idea of the measure h_{ij} is that the importance of token t_j in document d_i is directly proportional to tf_{ij} (a token repeated several times is expected to be relevant for the document) and inversely proportional to df_j (a token present in a large number of documents is expected to provide less specific semantic information about the document, given its scarce specificity, than a token present in few documents).[37]

Once the reports have been converted into numerical vectors based on the frequencies of tokens, a topic modeling algorithm is applied to represent their semantic content. Topic modeling algorithms are statistical methods that infer distributions of tokens, called topics, which represent themes or concepts, considering the co-occurrence of tokens in a corpus of reports.[8,42] Several topic modeling algorithms have been proposed in the NLP literature.[43] LSA is based on singular value decomposition of the token-document matrix, that is, the matrix containing the number of occurrences of each token (rows) in each document (columns). By reducing the matrix dimension, LSA aims at inferring the relations among words with large expectation of appearing together across documents.[9] The main drawback of LSA is the fact that it does not produce semantically interpretable word embeddings. LDA is a generative probabilistic model which represents reports as mixtures of topics and topics distributions over tokens.[10] The main drawback of LDA is the assumption of independence among topic distributions, which can prevent the model to capture the correlations between topics.[44] In this work, we employ a topic modeling technique based on HDP, which considers the correlations between the topics inferred from a corpus of documents by modeling the topic distributions as mixture distributions.[6] Following the approach proposed in[45] the prior distribution $P_i(t_j)$ of the tokens t_j, $j = 1, \ldots, T$, of the vocabulary in the generic i^{th} report is set equal to the normalized TFIDF measure of the tokens:

$$P_i(t_j) = \frac{h_{ij}}{\sum_{j=1}^{T} h_{ij}} \tag{5}$$

The choice of using TFIDF is motivated by the needs of decreasing the weights associated to general tokens, which are often used in the reports but carry negligible semantic meaning, and of increasing the weights associated to specific tokens, which appear less often in the reports but are expected to influence the severity of the accidents. The posterior inference of HDP is solved via an approximation algorithm based on the application of Variational Bayes,[46] which estimated the posterior distribution of the topics in the reports. HDP receives in input the vectors $\{\boldsymbol{h}_i, i = 1, \ldots, D\}$ and provides as outcomes:

(a) the set of topic distributions $\{\phi_k, k = 1, \ldots, K\}$, where a generic topic distribution ϕ_k is represented by a set of weights, $w_{jk} \in [0, 1]$ associated to the token t_j, $j = 1, \ldots, T$, of the dictionary Δ, with $\sum_{j=1}^{T} w_{jk} = 1$;

(b) the set of vectors $\{\boldsymbol{\gamma}_i^{HDP}, i = 1, \ldots, D\}$, where each vector $\boldsymbol{\gamma}_i^{HDP}$ is associated to the preprocessed report $\tilde{d}_i$ and whose generic element, $\boldsymbol{\gamma}_{ik}^{HDP} \in [0, 1]$ with $\sum_{k=1}^{K} \boldsymbol{\gamma}_{ik}^{HDP} = 1$, is a measure of the contribution of topic ϕ_k to the description of the preprocessed report $\tilde{d}_i$.

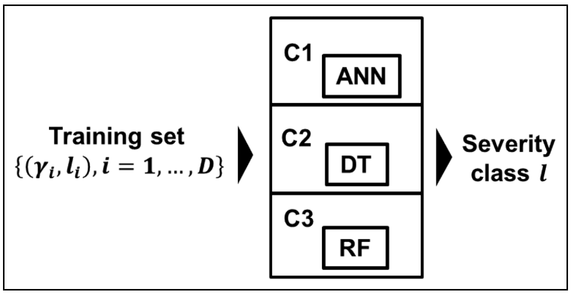
Training set
$\{(\boldsymbol{\gamma}_i, \boldsymbol{l}_i), \boldsymbol{i} = \mathbf{1}, \ldots, \boldsymbol{D}\}$
C1
ANN
C2
DT
C3
RF
Severity class l

Figure 3. Training of the classification models.

Doc2Vec. Word embedding methods are based on the distributional hypothesis, that is, words occurring in a similar context tend to have similar meanings.[47] They have yielded satisfactory results in capturing the semantic information of documents in many application fields.[48] Embedding based on neural networks were first introduced in[49] in the form of a feed-forward neural network language model. More recently, simplified models based on NNs with a single hidden layer, called Word2Vec and Doc2Vec, have been introduced in.[7] They are trained for the task of predicting a word given its context, for example, the following word in the sentence or the words in a window around the selected word. When Word2Vec is fed by a test word, it produces a vector of fixed dimension encoding its semantic information. The idea behind Word2Vec is that words with similar meaning will be converted in vectors in the same direction. Doc2Vec is an extension of Word2Vec capable of encoding entire documents in vectors, similarly to what Word2Vec does for words. In this work, Doc2Vec is applied to transform each report d_i, $i = 1, \ldots, D$, into a vector $\boldsymbol{\gamma}_i^{D2V}$, $i = 1, \ldots, D$, of real numbers, where each element, $\boldsymbol{\gamma}_{in}^{D2V}$, represents the projection of the vector $\boldsymbol{\gamma}_i^{D2V}$ on the axis n, $n = 1, \ldots, N$, of the N-dimensional semantic space found by applying Doc2Vec to the corpus of D reports.

Vector classification

Three different classification methods are considered for the classification of the vectors, $\boldsymbol{\gamma}^{HDP}$ or $\boldsymbol{\gamma}^{D2V}$, into the class of severity, l, of the road accident (Figure 3): ($C1$) a DT, ($C2$) a RF composed of an ensemble of decision trees built using bootstrap samples of the training set and validated using the out-of-bag scores, and $C3$) a fully connected feedforward ANN, whose architecture is composed by an input layer with a number of neurons equal to the dimension of the input vectors, a single hidden layer and the output layer with L neurons characterized by a softmax activation function. The output of the ANN is interpreted as the degree of confidence in the classification of the report to the corresponding class $l \in \{0, \ldots, L\}$.[50] The hyperparameters of the RF and ANN classifiers have been set by trial-and-error using only the training data within a 10-fold cross validation approach.

Considering all the possible combinations between the two text representation methods ($R1$ and $R2$) and the three classification methods ($C1$, $C2$ and $C3$), six classifiers have been trained. Each one is developed using as training set $\{(\boldsymbol{\gamma}_i, l_i),\ i = 1, \ldots, D\}$ where $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_i^{HDP}$ for the combinations ($R1$, $C1$), ($R1$, $C2$), and ($R1$, $C3$), and $\boldsymbol{\gamma}_i = \boldsymbol{\gamma}_i^{D2V}$ for the combinations ($R2$, $C1$), ($R2$, $C2$), and ($R2$, $C3$). Vectors $\boldsymbol{\gamma}_i^{D2V}$, $i = 1, \ldots, D$, have been normalized using the L2 norm:

$$\tilde{\boldsymbol{\gamma}}_{in}^{D2V} = \frac{\boldsymbol{\gamma}_{in}^{D2V}}{\sqrt{\sum_{n=1}^{N} (\boldsymbol{\gamma}_{in}^{D2V})^2}} \tag{6}$$

The normalization is motivated by the fact that, from one hand, ANNs tend to provide better classification results on L2-normalized inputs,[51] and, from the other hand, Eq. (6) preserves the direction of the Doc2Vec vectors in the semantic space, thus maintaining the representation capability of Doc2Vec.

Case study

The developed framework is applied to a repository of accident reports provided by the US National Highway Traffic Safety Administration.[20] This repository has been already exploited for statistical analyses of the crash data with the objective of estimating the contributions of different factors on the severity of the injuries sustained by the drivers in the accidents. Specifically, logistic regression models between the degradation of the vehicle and the injury severity of the drivers have been developed in Conroy et al.,[52] Augenstein et al.[53] for supporting the emergency medical services in deploying adequate protocols. Shannon et al.,[54] injuries suffered in road traffic accidents are related to expected trauma compensation pay-outs with the objective of deriving a quantitative cost function. Differently, the present work aims at developing a framework to classify the textual reports in support to road safety analysis.

The considered repository is made of 1159 reports, with average length of 356 words and containing a total of 5902 unique words. Each report is formed by:

(1) a free text reporting the narrative of the accident, wrote by a police officer after the event. It contains information that can be organized into the following five categories:

(1.a) involved vehicles and driver;
(1.b) weather conditions;
(1.c) infrastructure conditions, such as the state of the pavement and of the illumination;
(1.d) crash reconstruction, including the position and speed of the vehicles before the crash, their position after the accident, the way the vehicles collided with each other and/or with the road infrastructure barriers;
(1.e) consequences to the people and vehicles.

Among all the above categories, this work considers information only of categories *1.b* and *1.c*, which are not directly connected to the drivers behavior and its effect on the accident, and, therefore, are useful for the identification of critical factors related to road safety, which are under the control of the road network owner (type *1.c*) and/or can be subject to safety improvements by preventive or mitigative interventions (type *1.b*). The identification of the sentences of the reports related to categories *1.b* and *1.c* is performed by defining a taxonomy of the vocabulary used in the reports, which contains the two categories “weather conditions” and “infrastructure conditions.” In practice, a list of words related to the two categories is defined and a sentence of the reports is selected only if it contains at least one word belonging to one of the two categories. Table 2 reports an example of the procedure for the identification of the sentences related to categories *1.b* and *1.c*. Notice that alternative methods based on state-of-the-art NLP techniques of text segmentation[55] could also be used for the identification of the type of information in the sentences of the reports: these methods will be object of future work for fully automatizing the process.

(2) A set of integer numbers between 0 and 5 is associated to the class of injury severity sustained by the people involved in the accident, where 0 indicates little to no injury and 5 indicates very serious injury. These labels are assigned by a team of medical investigators of the police;
(3) the number of convalescence days of each person involved in the accident. This information is retrieved by the police officers from hospital medical records.

Table 1 reports the whole narratives of two accidents, the extracted weather and infrastructure conditions, the classes of injury severity and the number of convalescence days of the persons involved.

A preliminary analysis of the repository has shown that there are significant inconsistencies among the text descriptions of the accident consequences reported in (1.e), the classes of severity in (2) and the number of convalescence days in (3). For example, in several cases the class of injury severity five has been associated to

Table 1. Two examples of accident descriptions taken from the repository before and after processing, where the words of the taxonomy belonging to categories "Weather" and "Infrastructure conditions" are report in bold."

Narrative of the accident	Extracted weather and infrastructure sentences	Class of injury severity	Convalescence days
Case Focus: The focus of this case is on a 22-year old, male driver and a 25-year old, right front male passenger, both of a 2010 Chevrolet Camaro, which was involved in a left side impact. Collision Sequence Pre-Crash: This single-vehicle collision occurred during the afternoon hours (**daylight**), of a late **winter** weekday, on a two-lane, east/west **bituminous** roadway. At the crash area, the roadway has a long left turning curve which transitions into a right turning curve ("S-type"), for the eastbound travel direction. The speed limit is 35 mph (56 km/h). The overall **environment** is **rural**, with **trees/woods** bordering either side of the roadway. At the time of the crash, the **weather** was **clear** and the roadway **surface** was **dry**. Vehicle 1, the 2010 Chevrolet Camaro, was being operated by the 22-year old male driver (case occupant), in the eastbound travel lane. He was negotiating the left-turning curved section of the roadway and intended to continue traveling east. Occupying the right front seating position was a 25-year old male passenger (case occupant). Both the driver and passenger were not belted, but the vehicle was noted to be equipped with advanced frontal impact air bags, front outboard side impact hip/torso air bags and roof side rail curtains. Crash: For unknown reasons, while negotiating the left-turning curve, Vehicle 1 traveled across the westbound travel lane and its left side wheels departed the north edge of the roadway. Vehicle 1 continued along the north roadside edge (negatively graded embankment present) straddling the same. At this time, the driver of Vehicle 1 may have tried to steer right to regain the roadway, but was unsuccessful as the embankment and right steering maneuver likely contributed to Vehicle 1 entering a slight left side leading yaw. Vehicle 1 subsequently struck a large tree with its left side plane and came to rest, at the point of impact, facing a southeasterly direction. As a result of the impact, Vehicle 1′s frontal air bags deployed, as did the driver's side hip/torso air bag and the side rail curtains. Post-Crash: Both occupants of Vehicle 1 were removed from the vehicle by responding emergency medical personnel (EMS) and transported, by air unit, to a local trauma center and hospitalized with minor to severe injuries. Vehicle 1 was towed from the scene due to damage sustained in the crash.	This single-vehicle collision occurred during the afternoon hours (daylight), of a late winter weekday, on a two-lane, east/west bituminous roadway. At the time of the crash, the weather was clear and the roadway surface was dry. The overall environment is rural, with trees/woods bordering either side of the roadway.	[3,3]	[13, 22]
The focus of this case is on an 18-year old, male driver (case occupant), of a 2009 Honda Civic (V1), involved in two off road side impact crashes with trees, one to each side of V1. This single vehicle collision occurred during the **early evening** hours (**light**), of an **autumn** weekend day (Sunday), on a **winding** and hilly stretch of east/west roadway. The east/west roadway was a non **divided** two lane **bituminous**. The **weather** was **clear** and the roadway **surface** was **dry**. The speed limit for the roadway is 48 km/h (30 mph). V1 was reported on the PAR to be traveling 96 km/h (60 mph) just before the crash. The 18-year old male driver (case occupant) was operating V1 in the eastbound travel lane. He was wearing his 3-point lap/shoulder belt and had the benefit of an advanced frontal impact air bag. Two other non-case occupants were in the vehicle in the front right and second row right seating positions (positions 13 and 23). V1 crested the hill at a high rate of speed (60 mph), lost control and departed the roadway continuing to travel east as the roadway turned to the northeast. The right side plane of V1 impacted a tree causing V1 to rotate clockwise. The left side plane of V1 then contacted a second tree and came to rest against the tree. The top of the second tree broke off approximately 2 m above ground level as a result of the impact. The driver of Vehicle 1 (case occupant) was removed from the vehicle and transported, by air rescue, to a trauma center and hospitalized with serious injuries. All other occupants were transported to a local medical facility. The extent of their injuries and/or treatments is unknown. V1 was towed due to damage sustained in the crash.	This single vehicle collision occurred during the early evening hours (light), of an autumn weekend day (Sunday), on a winding and hilly stretch of east/west roadway. The east/west roadway was a non divided two lane bituminous. The weather was clear and the roadway surface was dry	[3,5,5]	[2,0,0]

accidents characterized by minor injuries of the driver according to the description in 1.e), whereas in other cases a large number of convalescence days has been associated to accidents assigned to class of injury severity 0. Clearly the presence in the repository of these contradictory patterns, with similar inputs associated to different outputs, can be harmful for the training of the empirical classifiers. For this reason, the reports have been relabeled into the two macro classes of severity: "Minor to Serious" and "Severe," following the procedure described in.[56] Specifically, the rules used for the assignment of the new classes to a generic report d_i are:

(1) if more than one person is involved in the accident, the largest class of injury severity and the largest number of convalescence days are considered;
(2) if the class assigned by the police officer is lower than or equal to three and the number of convalescence days is lower than 14, then the new class is "Minor to Serious";
(3) in all the remaining cases, the new class is "Severe";

The main motivations behind the definition of the new severity classes have been to:

(a) reduce the subjectivity of the assignments made by the police officers and the number of contradictory patterns in the repository. This can be obtained by decreasing the number of classes from six to two and by using the least subjective information available on the number of convalescence days;
(b) be able to identify the factors influencing the accidents associated to the mildest and most severe consequences, since the most critical factors are expected to cause major modifications of the accident consequences, that is, from minor to severe;
(c) be conservative by choosing the worst condition between the police officer assignment and the number of convalescence days among the victims of the accident.

In the following, the class "Minor to Serious" will be represented by the label 0 and the class "Severe" by the label 1. As expected, the distribution of the reports into the two classes of severity is largely unbalanced, with 1052 reports of class 0 and 107 reports of class 1.

Framework development

The D = 1159 reports are split into a training set and a test set, made of D_{train} = 1043 and D_{test} = 116 reports, respectively. The partition is performed keeping in both sets the same (imbalance) ratio between reports of classes 0 and 1 that there is in the whole repository. To generate the vocabulary, the preprocessing procedure described in Section 3.1 is applied only to the training set. Bigrams have been identified by setting the parameters c_{min} and s_{thresh} (Section 3.1) equal to 1 and 0.01, respectively, in accordance to Mikolov et al.[40] Then, a generic document of the test set d_{test} is preprocessed by:

(1) Applying steps from (a) to (d) of the procedure described in Section 3.1. In this way the reports d^{test} is converted into the list of single tokens in their base form $\tilde{d}_{test}$;
(2) An n-gram formed by n consecutive tokens is added to $\widetilde{d}_{test}$ if and only if it is present in the vocabulary of the repository obtained from the training set.

This procedure guarantees that the test set is not used for the training of the models and the setting of their hyperparameters.

The preprocessed corpus of reports of the training set, $\left\{\left(\tilde{d}_i, l_i\right), i = 1, \ldots, D_{train}\right\}$, where a generic pair $\left(\tilde{d}_i, l_i\right)$ is composed by the preprocessed report $\tilde{d}_i$ and its associated class $l_i \in \{0,1\}$, is converted in the vectors $\{(\boldsymbol{\gamma}_i^{HDP}, l_i), i = 1, \ldots, D_{train}\}$ and $\{(\boldsymbol{\gamma}_i^{D2V}, l_i), i = 1, \ldots, D_{train}\}$ following the procedures described in Section 3.2. The number of topics, K, searched by the HDP has been set equal to 16 by adopting a trial-and-error procedure. It has been verified that smaller values of K tend to provide topics which assign large weights to tokens with very different semantic meaning, whereas larger values of K tend to spread the tokens with similar semantic meaning in multiple topics.[57] Regarding the hyperparameters of Doc2Vec, the learning rate, the number of epochs and the dimension N of the vectors have been set equal to 0.025, 100 and 150, respectively, according to Li et al.[55] HDP and Doc2Vec models trained on the training set are also adopted, separately, to infer the vectors $\{(\boldsymbol{\rho}_i^{HDP}, l_i), i = 1, \ldots, D_{test}\}$ and $\{(\boldsymbol{\rho}_i^{D2V}, l_i), i = 1, \ldots, D_{test}\}$ from the test set.

The model hyperparameters used to train the DT, RF and ANN classifiers have been set by adopting a trial-and-error procedure within a 10-fold cross validation approach applied only to the training data and based on the following steps:

i. a set of possible tentative values of the hyperparameters is fixed;
ii. the training set is randomly divided into 10-folds with the same (imbalance) ratio between classes 0 and 1 reports. Then, 9 out of 10-folds are used to train the classifier, whereas the remaining fold is used as validation set;
iii. Steps *i* and *ii* are repeated 10 times, using each time a different fold as validation set. The combination of hyperparameters corresponding to the largest average $F_{measure}$ on the 10-folds (equation (2)) is selected.

The optimization of the hyperparameters is done independently for the classifiers based on the HDP and Doc2Vec extracted features. Specifically, a grid search has been performed considering the number of neurons

Table 2. Sets of values considered for the grid search, and resulting optimal hyperparameters of the RF and ANN classifiers.

Text model	Classifier	Hyperparameter	Searched values	Optimal value selected
HDP	ANN	Number of neurons in the hidden layer	$\begin{bmatrix} 5, 10, 15, 20, 25, \\ 50, 100, 150, 200 \end{bmatrix}$	15
		Learning rate	$[10^{-2}, 10^{-3}, 10^{-4}]$	10^{-3}
HDP	RF	Number of trees	[50, 100, 150]	100
		Fraction of bootstrap samples	[0.5, 0.67, 0.8]	0.67
Doc2Vec	ANN	Number of neurons in the hidden layer	$\begin{bmatrix} 5, 10, 15, 20, 25, \\ 50, 100, 150, 200 \end{bmatrix}$	100
		Learning rate	$[10^{-2}, 10^{-3}, 10^{-4}]$	10^{-3}
Doc2Vec	RF	Number of trees	[50, 100, 150]	100
		Fraction of bootstrap samples	[0.5, 0.67, 0.8]	0.67

Table 3. Classification performances achieved by the different combinations of methods. The classification accuracy (top) and $F_{measure}$ (bottom) are reported as average ± 1 standard deviation over a 10-fold cross validation.

Combination of methods	Text model	Classifier	Classification performance metric	Classification performance		
				Training	Validation	Test
(R1, C1)	HDP	ANN	Accuracy$F_{measure}$	0.767±0.0180.768±0.022	0.735±0.0410.734±0.050	0.6870.747
(R1, C2)	HDP	DT	Accuracy$F_{measure}$	0.891±0.0010.891±0.002	0.764±0.0210.763±0.038	0.6860.747
(R1, C3)	HDP	RF	Accuracy$F_{measure}$	0.906±0.0030.905±0.006	0.788±0.0320.788±0.023	0.7550.789
(R2, C1)	Doc2Vec	ANN	Accuracy$F_{measure}$	0.967±0.0060.968±0.021	0.925±0.0220.925±0.023	0.8710.849
(R2, C2)	Doc2Vec	DT	Accuracy$F_{measure}$	1.000±0.0001.000±0.000	0.866±0.0180.869±0.017	0.7530.779
(R2, C3)	Doc2Vec	RF	Accuracy$F_{measure}$	0.959±0.0020.958±0.001	0.917±0.0200.908±0.015	0.7620.791

in the hidden layer and the learning rate for the ANN classifiers, and the number of trees and the fraction of bootstrap samples for the RF classifiers. Table 2 reports the sets of values considered for the grid search and the resulting optimal hyperparameters of the classifiers.

The input layer of the ANN is characterized by 16 neurons when fed by the vectors $\{(\boldsymbol{\gamma}_i^{HDP}, l_i), i = 1, \ldots, D_{train}\}$ and 150 neurons when fed by the vectors $\{(\boldsymbol{\gamma}_i^{D2V}, l_i), i = 1, \ldots, D_{train}\}$. This difference accounts for the different lengths of the input vectors, which are 16 for HDP and 150 for Doc2Vec. The ANN training is stopped when either the number of iterations reaches the maximum number of epochs, which has been set equal to 2000 to avoid underfitting, or when the loss function computed on the validation set did not decrease for 50 iterations, to avoid overfitting. During the training phase, a report of class $l_i = 0$ is associated to the two-dimensional output $o = [o_0, o_1] = [1,0]$, whereas a report of class $l_i = 1$ is associated to the output $o = [o_0, o_1] = [0,1]$. Thus, when the ANN is used for the classification of a test report, d, the output value o_0 (o_1) is interpreted as the degree of confidence in the classification of the report to the class "Minor to Moderate" ("Severe").[50] Finally, the test report is assigned to the class with the associated largest degree of confidence.

Results

Due to the large imbalance ratio between the two classes of accidents in the repository, data augmentation is used to improve the classification accuracy.[57] According to Wei and Zou,[58] a new report is artificially generated from a report of the repository by: (1) randomly swapping the position of a generic token with the position of another one with probability 0.3, (2) randomly deleting a token from the report with probability 0.3. To obtain a training set made by a similar number of reports of the two classes, the augmentation procedure is applied nine times to each report of class 1. In this way, a balanced training set formed by $D_{aug} = 1819$ reports (946 of class 0 and 873 of class 1) is obtained. Notice that the application of the developed data augmentation procedure does not significantly change the distribution of the number of tokens in a class, since replicating the reports has the effect of replicating the tokens contained in them, and, therefore, the proportions among the tokens in a class are not significantly affected.

Table 3 reports the classification performances achieved by the different combinations of methods.

Notice that:

(a) As expected, the RF classifiers tend to outperform the DT classifiers, since they use and ensemble of decision trees[17];
(b) The best performances are obtained by the ANN classifier trained on the Doc2Vec vectors, which is consist with the findings in the NLP literature.[59]

Figure 4 shows the confusion matrixes of the RF fed by vectors $\boldsymbol{\rho}^{HDP}$ and of the ANN classifier fed by the vectors $\boldsymbol{\rho}^{D2V}$, respectively.

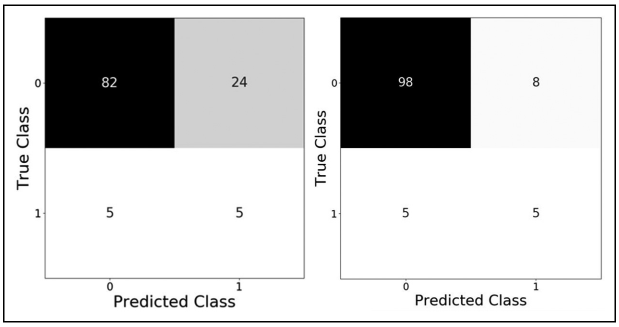

Figure 4. Confusion matrix of the developed RF tested on the vectors $\boldsymbol{\rho}^{HDP}$ (left) and of the developed ANN tested on the vectors $\boldsymbol{\rho}^{D2V}$ (right).

Notice that the classifiers tend to assign the class of severity "Minor to Serious" to some reports whose true class is "Severe." This is due to the small number of available reports of class 1 and to the presence of similar words in the reports of the two classes. Specifically, out the 390 unique tokens used in reports of class 1, 351 are also used for reports of class 0. Table 4 gives the five reports in the test set whose true class is "Severe" but have been erroneously assigned to class "Minor to Serious." It has been verified that the same reports are misclassified by both models, and that most of them are assigned to the original class of injury severity "3," which is at the border between the new classes "Minor to Moderate" containing the original classes "0," "1," and "2," and "Severe" containing the original classes and "3," "4," and "5."

Investigation of the effect of performing data augmentation

To investigate the effect of data augmentation, the combinations (R1, C3) and (R2, C1) have been directly applied to the imbalanced repository in a 10-fold cross-validation procedure. Table 5 reports their performances, which remarkably decrease.[60]

Investigation of the robustness of the classifier

To further evaluate the robustness of the classification results obtained by the combinations (R1, C3) and (R2, C1), the repository is divided in 10 batches characterized by similar (imbalance) ratio between class 0 and class 1 reports. Then, a 10-folds cross-validation procedure is applied by developing 10 classifiers, each one trained with a different combination of 9 batches and tested on the remaining batch. The augmentation technique defined in Section 4.2 is applied to the training sets of each classifier. Notice that this cross-validation procedure should not be confused with the cross-validation procedure used to set the hyperparameters of the models, which is applied only to the training set.

Table 4. Reports of the test set whose true severity class is "Severe" but are erroneously assigned to the class "Minor to Serious" by RF and ANN classifiers.

Report	Class	Class of injury severity in the original scale [0,5]	Convalescence days
The crash occurred during night hours in dry and clear weather.	Severe	3	8
It was daylight, cloudy, and the bituminous road was dry. The driver of V1 was traversing a left curve when the right side tires departed the paved surface.	Severe	3	22
It was day time, the weather was clear and the bituminous (asphalt) road was dry and level.	Severe	4	5
The weather was clear and the bituminous roadway surfaces dry during the afternoon, weekend crash.	Severe	3	4
It was daylight, the weather was cloudy and the asphalt road surface was dry.	Severe	3	7

Table 5. Classification accuracy (top) and $F_{measure}$ (bottom) of combinations (R1, C3) and (R2, C1) in a 10-fold cross-validation procedure with and without performing data augmentation.

Text model	Classifier	Augmentation	Classification performance metric	Classification performance		
				Training	Validation	Test
HDP	RF	No	Accuracy$F_{measure}$	0.500±0.0010.503±0.001	0.506±0.0570.401±0.101	0.5190.337
		Yes	Accuracy$F_{measure}$	0.906±0.0030.905±0.006	0.788±0.0320.788±0.023	0.7550.789
Doc2Vec	ANN	No	Accuracy$F_{measure}$	0.502±0.0050.324±0.019	0.493±0.0220.330±0.019	0.4910.316
		Yes	Accuracy$F_{measure}$	0.967±0.0060.968±0.021	0.925±0.0220.925±0.023	0.8710.849

Table 6. Classification accuracy (top) and $F_{measure}$ (bottom) of (R1, C3) and (R2, C1) in 10-fold cross-validation procedure.

Combination of methods	Text model	Classifier	Classification performance metric	Classification performance on the test set
(R1, C3)	HDP	RF	Accuracy$F_{measure}$	0.768±0.0260.795±0.033
(R2, C1)	Doc2Vec	ANN	Accuracy$F_{measure}$	0.854±0.0320.855±0.024

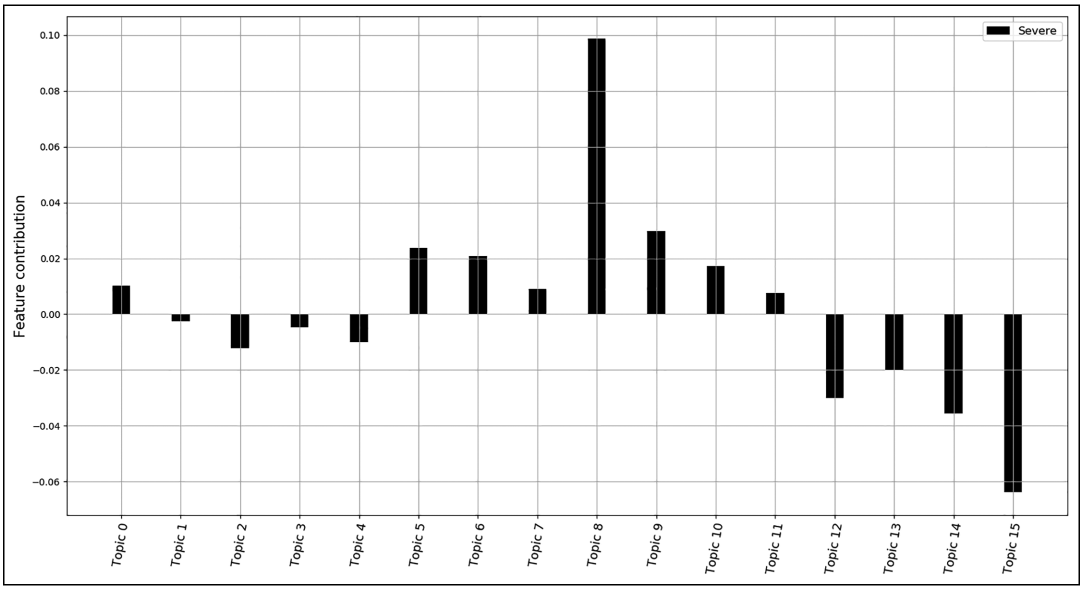

Figure 5. Contribution of each topic to the classification of the test report d_{test}.

Table 6 reports the obtained classification performances on the test set. The results confirm that the model is robust with respect to variations of the input training set and that the results obtained in the previous Sections were not biased by a particular combination of training and test reports.

Interpretability of the results

RF allows retrieving the contribution of each feature with respect to the classification of a report to the different classes.[61] Specifically, considering a RF model made by N_{tree} trees, the contribution of the f^{th} feature to the classification of a generic test vector $\boldsymbol{\rho}_{test}$, extracted from the test report d_{test}, is[62]:

$$c_f^{RF}(\boldsymbol{\rho}_{test}) = \frac{\sum_{tree=1}^{N_{tree}} c_f^{tree}(\boldsymbol{\rho}_{test})}{N_{tree}} \qquad (7)$$

where $c_f^{tree}(\boldsymbol{\rho}_{test})$ is the sum of the differences between the values of the parent nodes and the children nodes associated to the f^{th} feature along the decision path.[62] A positive contribution to class 0 (1) means that the f^{th} feature has contributed in most trees to assign the vector $\boldsymbol{\rho}_{test}$ to class 0 (1), and vice versa to the class 1 (0) in case of negative contribution.

Therefore, considering the combination (R1, C3) and the correspondence between the features of the HDP vectors and the topics, the index $c_f^{RF}(\boldsymbol{\rho}_{test})$ can be interpreted as the criticality of the f^{th} topic in the classification of the report. For example, Figure 5 shows the contribution of the different topics to the classification of the report d_{test} = "*The crash occurred during daylight hours at the hilltop of a four lane trafficway. It was raining at the time of the crash and the straight asphalt roadway was wet*" to class "Severe" (1), which has been correctly classified by the combination (R1, C3).

It can be observed that the largest index $c_f^{RF}(\boldsymbol{\rho}_{test}^{HDP})$ is associated to topic 8, which is a list of tokens whose associated weights w_j, $j = 1, \ldots, T$, are shown in Figure 6. Therefore, the tokens of the text that are most responsible for the classification of the report d_{test} to the class of severity "Severe" can be identified by combining the topic contributions (Figure 5) and the token weights in the topic (Figure 6). Since the token weights are found by HDP considering all reports, only the tokens appearing in the specific report d_{test} should be considered in the analysis. For example, the token "*wet*," which is among those with largest weights for topic eight, can be considered as a factor of influence of the severity of the accident.

The example shows that the vector $\boldsymbol{\rho}^{HDP}$ allows extracting useful semantic relationships between tokens and accident severity. Contrarily, the elements of the vectors $\boldsymbol{\rho}^{D2V}$, do not offer a direct way to retrieve the semantic information since they are not directly mapped to distributions of words. Therefore, it is

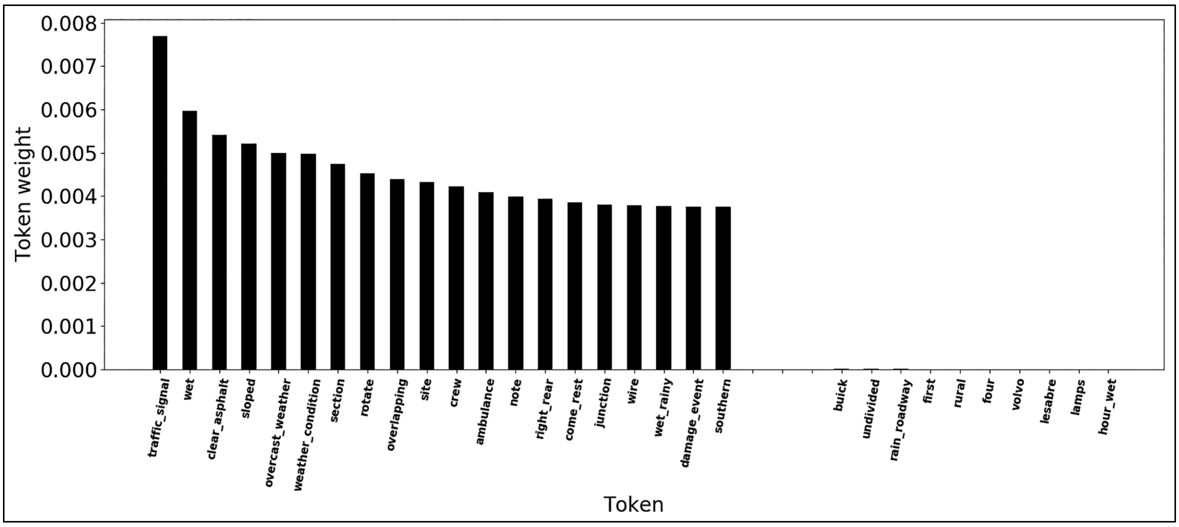

Figure 6. Distribution of the word weights associated to topic 8.

possible to conclude that the representation of the text based on HDP topic modeling allows a clearer interpretation of the results than the representation provided by Doc2Vec, despite the slightly lower classification performance.

A possible solution to extract knowledge from $\boldsymbol{\rho}^{D2V}$ is to define a distance metric to measure the similarity among the reports and to use it to cluster the repository.[63] Then, a semantic interpretation of the clusters can be searched by domain experts considering the most frequent tokens in the clusters. Since the accuracy achieved by the classifiers built on the two different representations are not remarkably different and the interpretation of the clusters by domain experts is subjective, methods to interpret the feature space extracted by Doc2Vec have not been developed in this work.

Conclusions

In the context of road safety analysis, it is fundamental to assist domain experts in the identification of the critical factors influencing frequency and severity of road accidents. This work has developed a framework based on NLP and ML for the automatic classification of road accident reports in classes homogeneous with respect to the severity of the accident consequences. The main novelty is that the classification problem is addressed by performing an intermediate step of text representation, which transforms each textual report into a numerical vector whose elements are correlated with its semantic content, with the objective of identifying the factors influencing accident severity. To identify the best approach with respect to the objectives of classification performance and interpretability of the results, different combinations of text representation and ML-based classification models have been considered. In particular, HDP and Doc2Vec have been considered as methods for text representation and ANN, DT and RF as classifiers. A repository of road accident reports provided by the US National Highway Traffic Safety Administration has been used to assess the performance of the combinations of text representation and classification methods. The obtained results have shown that: (1) the RF classifier outperforms the other alternatives when combined with HDP topic modeling, (2) the ANN classifier outperforms the other alternatives when combined with Doc2Vec and achieving the best overall performance, (3) HDP transforms reports into vectors from which useful information about the correlation between tokens and accident severity can be extracted, whereas Doc2Vec transforms reports into vectors from which knowledge is more difficult to retrieve, (4) the combination of RF and HDP allows for identifying the tokens that mostly influenced the assessment of the class of severity to the accident report. The main contribution of this work to road safety analysis is the development of a framework able to accurately classify road accident reports and from which knowledge on the factors that are critical with respect to severity of the accident consequences can be retrieved.

Future work will be devoted to: (a) introducing a more systematic processing of the data by investigating the possibility of applying text segmentation techniques to automatically divide text in sections without resorting to predefined list of tokens[55]; (b) exploring methods for the automatic identification of contradictory patterns, such as in Baraldi et al.[64]; (c) investigating the possibility of applying advanced text representation methods, such as transformer-based models,[12] in combination with explainability methods for obtaining interpretations of the extracted features. With respect to this latter issue, the authors will consider the exploitation of the attention mechanism to identify the tokens that have mostly influenced the classification[65]; (d) verifying the classification framework on repositories containing a larger number of reports and more than two classes of severity; (e) integrating the classification framework into a methodology to automatically extract the factors that influence the accident severity. The final aim is to develop a decision-making framework that considers the information content of repositories of car accident reports and identifies the intervention strategies needed to improve safety, in assistance to the safety operators.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Piero Baraldi and Enrico Zio disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by "Smart maintenance of industrial plants and civil structures by 4.0 monitoring technologies and prognostic approaches – mac4pro," sponsored by the call BRIC-2018 of the National Institute for Insurance against Accidents at Work – INAIL (CUP J56C18002030001).

ORCID iDs

Dario Valcamonico https://orcid.org/0000-0002-7621-6831
Piero Baraldi https://orcid.org/0000-0003-4232-4161
Enrico Zio https://orcid.org/0000-0002-7108-637X

References

1. World Health Organization. *Global status report on road safety*. Geneva: World Health Organization, 2018.
2. Imprialou M and Quddus M. Crash data quality for road safety research: Current state and future directions.. *Accid Anal Prev* 2019; 130: 84–90.
3. Rolison JJ. Identifying the causes of road traffic collisions: Using police officers' expertise to improve the reporting of contributory factors data. *Accid Anal Prev* 2020; 135: 105390. DOI: 10.1016/j.aap.2019.105390
4. Krause S and Busch F. New insights into road accident analysis through the use of text mining methods. In: *6th international conference on models and technologies for intelligent transportation systems (MT-ITS)*, 2019. New York: *IEEE*.
5. Kwon OH, Rhee W and Yoon Y. Application of classification algorithms for analysis of road safety risk factor dependencies. *Accid Anal Prev* 2015; 75: 1–15.
6. Teh YW, Jordan MI, Beal MJ, et al. Hierarchical Dirichlet processes. *J Am Stat Assoc* 2006; 101: 1566–1581.
7. Le Q and Mikolov T. Distributed Representations of Sentences and Documents. In: *Proceedings of the 31 st international conference on machine learning*, 2014.
8. Blei D, Carin L and Dunson D. Probabilistic topic models. In: *IEEE signal processing magazine*, 2010, Vol. 27, pp.55–65. New York: *IEEE*.
9. Landauer TK, Foltz PW and Laham D. An introduction to latent semantic analysis. *Discourse Process* 1998; 25: 259–284.
10. Blei DM, Ng AY and Jordan MT. Latent dirichlet allocation. *Adv Neural Inf Process Syst* 2002; 3: 993–1022.
11. Martinčić-Ipšić S, Miličić T and Todorovski A. The influence of feature representation of text on the performance of Document Classification. *Appl Sci* 2019; 9: 743.
12. Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need. In: *Proceeding of the 31st conference on neural information processing systems (NIPS 2017)*, 2017.
13. Fang W, Luo H, Xu S, et al. Automated text classification of near-misses from safety reports: an improved deep learning approach. *Adv Eng Inform* 2020; 44: 44.
14. Mauni HZ, Hossain T and Rab R. Classification of underrepresented text data in an imbalanced dataset using deep neural network. In: *IEEE region 10 symposium (TENSYMP)*, 2020, pp.997–1000. New York: *IEEE.*
15. Mishu SZ and Rafiuddin S. Performance analysis of supervised machine learning algorithms for text classification. In: *19th international conference on computer and information technology (ICCIT)*, 2016, pp.409–413. New York: *IEEE*.
16. Zaghloul W, Lee SM and Trimi S. Text classification: neural networks vs support vector machines. *Ind Manag Data Syst* 2009; 109: 708–717.
17. Sun Y, Li Y, Zeng Q, et al. Application research of text classification based on random forest algorithm. In: *3rd International conference on advanced electronic materials, computers and software engineering (AEMCSE)*, 2020.
18. Vries VD Classification of aviation safety reports using machine learning. In: *International conference on artificial intelligence and data analytics for air transportation (AIDA-AT)*, 2020, pp.1–6. New York: IEEE.
19. Valcamonico D, Baraldi P, Amigoni F, et al. Text mining for the automatic classification of road accident reports. In: *Proceedings of the 30th European safety and reliability conference and the 15th probabilistic safety assessment and management conference*, 2020.
20. NHTSA. Crash injury research (CIREN), https://www.nhtsa.gov/research-data/crash-injury-research. Access ed 20 October 2019.
21. Gasparetto A, Marcuzzo M, Zangari A, et al. A survey on text classification algorithms: from text to predictions. *Information* 2022; 13: 83–39.
22. Yang Z, Baraldi P and Zio E. A novel method for maintenance record clustering and its application to a case study of maintenance optimization. *Reliab Eng Syst Saf* 2020; 203: 107103.
23. Guimarães MS, Gomes de Araújo HH, Lucas TC, et al. An NLP and text mining – based approach to categorize occupational accidents. In: *Proceedings of the 30th European safety and reliability conference and the 15th probabilistic safety assessment and management conference*, Venice, Italy, 2020.
24. Bezerra C, de Santana JMM, Moura M das C, et al. Automated classification of injury leave based on accident description and natural language processing. In: *Proceedings of the 30th European safety and reliability conference and the 15th probabilistic safety assessment and management conference*, Venice, Italy, 2020. DOI: 10.3850/981-973-0000-00-0.

25. Zhang F, Fleyeh H, Wang X, et al. Construction site accident analysis using text mining and natural language processing techniques. *Autom Constr* 2019; 99: 238–248.
26. Heidarysafa M, Kowsari K, Barnes LE, et al. Analysis of railway accidents' narratives using deep learning. In: *17th international conference on machine learning and applications*, 2018. DOI: 10.1109/ICMLA.2018.00235.
27. Zhang F. A hybrid structured deep neural network with Word2Vec for construction accident causes classification. *Int J Constr Manag* 2022; 22: 1120–1140.
28. Rane A and Kumar A. Sentiment classification system of twitter data for US airline service analysis. In: *Proceedings of the 42nd IEEE international conference computer software and applications*, 2018, pp.769–773. New York: *IEEE.*
29. Sarkar S, Vinay S and Maiti J. Text mining based safety risk assessment and prediction of occupational accidents in a steel plant. In: *2016 international conference on computational techniques in information and communication technologies (ICCTICT)*, 2016, pp.439–444. New York: *IEEE.*
30. Williams T and Betak J. A comparison of LSA and LDA for the analysis of railroad accident text. *Procedia Comput Sci* 2018; 130: 98–102.
31. Kwayu KM, Kwigizile V, Lee K, et al. Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. *Accid Anal Prev* 2021; 150: 105899.
32. Limsettho N, Hata H and Matsumoto KI. Comparing hierarchical dirichlet process with latent dirichlet allocation in bug report multiclass classification. In: *2014 IEEE/ACIS 15th international conference on software engineering, artificial intelligence, networking and parallel/distributed computing, SNPD 2014*, 2014. New York: *IEEE.*
33. Tahvili S, Hatvani L, Felderer M, et al. Automated functional dependency detection between test cases using Doc2Vec and clustering. In: *IEEE international conference on artificial intelligence testing (AITest)*, 2019, pp.19–26. New York: *IEEE.*
34. Bragatto P, Ansaldi S, Agnello P, et al. Ageing management and monitoring of critical equipment at Seveso sites: An ontological approach. *J Loss Prev Process Ind* 2020; 66: 104204
35. Macêdo JB, das Chagas Moura M, Aichele D, et al. Identification of risk features using text mining and BERT-based models_ application to an oil refinery. *Process Saf Environ Prot* 2022; 158: 382–399.
36. Bin C, Baigen C and Wei S. Text mining in fault analysis for on-board equipment of high-speed train control system. In: *Chinese automation congress (CAC)*, Jinan, China 2017; pp.6907–6911. New York: IEEE.
37. Weiss SM, Indurkhya N, Zhang T, et al. *Text mining. Predicitive methods for analyzing unstructured information.* Springer, Sydney, Australia, 2005.
38. Pereira J and Saraiva F. A comparative analysis of unbalanced data handling techniques for machine learning algorithms to electricity theft detection. In: *2020 IEEE Congress on evolutionary computation (CEC)*, 2020, pp.1–8. New York: IEEE.
39. Sojka P and Řehůřek R. Software framework for topic modelling with large corpora. In: *Proceeding Lr 2010 Work new challenges NLP Fram*, 2010.
40. Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality. In: *Proceedings of the 26th international conference on neural information processing systems*, 2013, Vol. 2, pp.3111–3119. Red Hook, NY: *Curran Associates Inc.*
41. Amati G and Van Rijsbergen CJ. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans Inf Syst* 2002; 20: 357–389.
42. Griffiths TL, Steyvers M and Tenenbaum JB. Topics in semantic representation. *Psychol Rev* 2007; 114: 211–244.
43. Alghamdi R and Alfalqi K. A survey of topic modeling in Text Mining. *Int J Adv Comput Sci Appl* 2015; 6: 147–153.
44. Blei DM and Lafferty JD. Correlated topic models. In: *Proceeding of Advances in Neural Information Processing Systems 18 conference (NIPS 2005), Vancouver, Canada*, 2005, pp.147–154.
45. Blei DM and Lafferty JD. Topic models. In: Srivastava AN and Sahami M (eds.), *Text mining*. Chapman and Hall/CRC, New York, 2009, pp.101–124.
46. Wang C, Paisley J and Blei DM. Online variational inference for the hierarchical Dirichlet process. *J Mach Learn Res* 2011; 15: 752–760.
47. Almeida F and Xexéo G. Word Embeddings: A Survey. Computing research repository (CoRR), 2019.
48. Schakel AMJ and Wilson BJ. Measuring word significance using distributed representations of words. Computing research repository (CoRR), 2015.
49. Bengio Y, Ducharme R, Vincent P, et al. A neural probabilistic language model. *J Mach Learn Res* 2003; 3: 1137–1155.
50. Nwankpa CE, Ijomah W, Gachagan A, et al. Activation Functions : Comparison of trends in practice and research for Deep Learning. *arXiv:181103378v1* 2018; 1–20.
51. Zhang K, Su H, Dou Y, et al. Evaluation of the influences of hyper-parameters and L2 norm regularization on ANN model for MNIST recognition. In: *International conference on intelligent computing, automation and systems (ICICAS)*, 2019, pp. 379–386. New York: IEEE.
52. Conroy C, Tominaga GT, Erwin S, et al. The influence of vehicle damage on injury severity of drivers in head-on motor vehicle crashes. *Accid Anal Prev* 2008; 40: 1589–1594.
53. Augenstein J, Perdeck E, Stratton J, et al. Characteristics of the crashes that increase the risk of serious injuries. *Assoc Advacement Autom Med* 2003; 47: 561–576.
54. Shannon D, Murphy F, Mullins M, et al. Applying crash data to injury claims – an investigation of determinant factors in severe motor vehicle accidents. *Accid Anal Prev* 2018; 113: 244–256.
55. Li J, Chiu B, Shang S, et al. Neural text segmentation and its application to sentiment analysis. *IEEE Trans Knowl Data Eng* 2022; 34: 828–842.
56. Lee JS, Kim YH, Yun JS, et al. Characteristics of patients injured in road traffic accidents according to the New Injury Severity Score. *Ann Rehabil Med* 2016; 40: 288–293.
57. Chen M, Ji X and Shen D. Short text classification improved by learning multi-granularity topics. In: *Proceedings of the 22nd international joint conference of artificial intelligence.* AAAI Press, Barcelona, Spain, 2011, pp. 1776–1781.

58. Wei J and Zou K. EDA : Easy data augmentation techniques for boosting performance on text classification tasks. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing*, Association for Computational Linguistics, Hong Kong, China, 2019, pp.6382–6388.
59. Dogru HB, Tilki S, Jamil A, et al. Deep learning-based classification of news texts using Doc2Vec model. In: *2021 1st international conference on artificial intelligence and data analytics, CAIDA 2021*, 2021, pp.91–96. New York: IEEE.
60. Liu X, Wu J and Zhou Z. Exploratory undersampling for class-imbalance learning. In: *IEEE transactions on systems, man, and cybernetics*, 2009, Vol. 39, pp.539–550. New York: *IEEE.*
61. Palczewska A, Palczewski J, Robinson RM, et al. Interpreting random forest classification models using a feature contribution method. In: *Integration of Reusable Systems*. Springer, Cham, 2014, pp.193–218.
62. Loecher M. From unbiased MDI feature importance to explainable AI for Trees. *arXiv Prepr* (statistic, Machine Learning), 2021.
63. Marjai P, Lehotay-Kéry P and Kiss A. Document similarity for error prediction. *J Inf Telecommun* 2021; 5: 407–420.
64. Baraldi P, Compare M, Zio E, et al. Identification of contradictory patterns in experimental datasets for the development of models for electrical cables diagnostics. *Int J Performability Eng* 2011; 7: 43–60.
65. Vig J and Belinkov Y. Analyzing the structure of attention in a transformer language model. In: *Proceedings of the 2019 ACL workshop blackboxNLP: Analyzing and interpreting neural networks for NLP*, Florence, Italy, 2019, pp.63–76.

Appendix

Notations

NLP	Natural Language Processing
HDP	Hierarchical Dirichlet Process
TFIDF	Term Frequency Inverse Document Frequency
ML	Machine Learning
ANN	Artificial Neural Network
DT	Decision Tree
RF	Random Forest
LSA	Latent Semantic Analysis
LDA	Latent Dirichlet Allocation
BERT	Bidirectional Encoder Representation from Transformers
R	Repository of reports
D	Number of reports in R
D_{train}	Number of reports in the training set
D_{test}	Number of reports in the test set
D_{aug}	Number of reports in the training set after augmentation
d	Report
d_i	i-th report of R
$\tilde{d}_i$	List of tokens of d_i
L	Number of classes
l_i	Class of d_i
Δ	Dictionary of R
T	Number of tokens in Δ
T_{train}	Number of unique tokens in the training set
T_{test}	Number of unique tokens in the test set
t_j	j-th token of Δ
$\boldsymbol{h}_i$	TFIDF vector of d_i
h_{ij}	j-th element of $\boldsymbol{h}_i$
tf_{ij}	Number of times that the token t_j occurs in $\tilde{d}_i$
df_j	Number of reports in which the token t_j occurs
$P_i(t_j)$	Prior distribution of the token t_j in the i-th report
K	Number of topics
ϕ_k	k-th topics of R
w_{jk}	Weight of token t_j in the topic ϕ_k
$\boldsymbol{\gamma}^{HDP}$	HDP vector of d
$\boldsymbol{\gamma}_i^{HDP}$	HDP vector of the i-th report in the training set
$\boldsymbol{\rho}_i^{HDP}$	HDP vector of the i-th report in the test set
N	Length of vector representations provided by Doc2Vec
$\boldsymbol{\gamma}^{D2V}$	Doc2Vec vector of d
$\boldsymbol{\gamma}_i^{D2V}$	Doc2Vec vector of the i-th report in the training set
$\boldsymbol{\rho}_i^{D2V}$	Doc2Vec vector of the i-th report in the test set
U	Number of unigrams of R
u_p	p-th unigram of R
$c(u_p)$	Number of times that u_p occurs in R
B	Number of bigrams of R
u_{pq}	Pair of contiguous unigrams u_p and u_q
$c(u_{pq})$	Number of times u_{pq} occurs in R
c_{min}	Minimum number of times that a bigram must occur in R to be considered
s_{pq}	Score associated to u_{pq}
s_{thresh}	Threshold for the identification of bigrams
A	Classification accuracy performance metric
$F_{measure}$	F-measure performance metric
$c_f^{RF}(\boldsymbol{\rho}_{test})$	Contribution of the f-th feature to the prediction of the class of the input vector $\boldsymbol{\rho}_{test}$ by the RF classifier