A Natural Language Processing Method For The Identification Of Critical Factors Influencing Road Safety

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

[1] *Energy Department, Politecnico di Milano, Milan, Italy*

dario.valcamonico@polimi.it
piero.baraldi@polimi.it
enrico.zio@polimi.it

[2] *Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy*

francesco.amigoni@polimi.it

[2] *MINES ParisTech, PSL Research University, CRC, Sophia Antipolis, France*

enrico.zio@mines-paristech.fr

ABSTRACT

Road safety analysis is typically performed by domain experts on the basis of the information contained in accident reports. The limitations in such analysis are in the fact of having to consider a large number of reports in textual form and in the subjectivity of the experts interpretation. To assist the expert in the analysis, this work develops a framework based on Natural Language Processing (NLP) for the automatic identification of those factors critical for accident severity. The framework combines the use of Hierarchical Dirichlet Processes (HDPs) for modelling the text of the reports and Sequential Forward Selection (SFS) for the identification of the critical factors. An application to a public repository of road accident reports is presented.

1. INTRODUCTION

Although road safety has improved with a reduction of the number of yearly fatalities in Europe from 54000 in 2001 to 28000 in 2013, the rate of road accidents with severe and fatal consequences is still exceeding the objectives stated by the European Commission in (European Commission, 2019). Road accidents remain the 8th leading cause of death worldwide, with the number of deaths peaking at 1.35 million people in 2018 (World Health Organization, 2018).

Road safety analysis is typically performed by experts through the use of large repositories of reports of road accidents collected by the public authorities and containing textual descriptions of the accidents and the results of post-event investigations. Road safety analysis profits from the identification of the factors influencing accident severity and frequency, with the objective of implementing preventive and mitigative solutions (Persia et al., 2016; Chao Wang et al., 2013). The main challenges encountered by domain experts are: 1) the poor quality and inhomogeneity of the textual content of the reports, which also typically do not contain enough information on the accident conditions (Imprialou & Quddus, 2019), 2) the hand-written nature of the reports, 3) the subjectivity of the post-accident assessments, which depends on the police officers writing the reports and 4) the large number of accident reports to be considered. To assist the experts in the analysis of the reports and alleviate somewhat the challenges just mentioned, this work develops a Natural Language Processing (NLP) framework for the automatic identification of factors influencing accident severity from the textual reports of road accidents.

NLP methods have been recently applied to repositories of accident reports in different areas with the objective of improving safety. In (Sarkar et al., 2016), a NLP technique, which considers the frequency of words used in occupational accident reports, is developed for the identification of the basic events causing accidents in steel plants. In (Williams & Betak, 2018), Latent Semantic Analysis (LSA) and Latent Dirichlet Allocation (LDA) are applied to railroad accident reports for their analysis. In (Bin et al., 2017), 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. In (Yang et al., 2020), an approach combining Convolutional Neural Networks (CNN) and Latent Semantic Analysis (LSA) is adopted for the classification of textual maintenance records,

with the objective of developing a stochastic multi-stage degradation model of excavators used in the mining industry. In (Guimarães et al., 2020), Principal Component Analysis (PCA) is combined with k-means clustering for analyzing accident classes in a corpus of occupational accident reports. In (Bezerra et al., 2020), 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. In (Morgan et al., 2021), Structural Topic Modeling (STM) is combined with network topology analysis to discover themes and their relationships in road crash narratives, and for the classification of road accident reports. In (Ansaldi et al., 2020), an ontology has been defined on the basis of safety documents and applied to the analysis of equipment ageing of a liquid fuel depot in an industrial establishment.

The method developed in this work combines Hierarchical Dirichlet Processes (HDPs), to convert the accident reports into numerical vectors, and a Sequential Forward Selection (SFS) technique, to identify the critical factors influencing the severity of the accident consequences. HDPs allow finding topics, i.e. distributions of words, which can be thought of as themes or concepts in the corpus of accident reports (Teh et al., 2006). In practice, each report is transformed into a vector whose elements are the degrees of membership of the report to the identified topics. Then, the problem of identifying the critical factors influencing the severity of the accidents is framed as a feature selection problem. The idea is to build a classifier which maps the reports, modelled as mixtures of topics, into the corresponding severity classes and to consider those topics that are more relevant for the classification of the reports. Artificial Neural Networks (ANNs) are considered to classify the reports, given their robustness and good performance in text classification problems (Mauni et al., 2020; Mishu & Rafiuddin, 2016; Zaghloul et al., 2009), whereas Sequential Forward Selection (SFS) is used as feature selection technique, given its capability of selecting, with limited computational effort, a small number of significative features (Marcano-Cedeño et al., 2010; Ververidis & Kotropoulos, 2005).

The developed framework is applied to a repository of accident reports provided by the US National Highway Traffic Safety Administration (NHTSA, 2015), containing the narrative of crash accidents recorded by police officers and the corresponding classification of the severity of the consequences.

The remaining of the work is organized as follows. In Section 2, the problem of identifying the critical factors is set and, in Section 3, it is formulated as a feature selection problem. In Section 4, the developed method is described. In Section 5, the case study is introduced and the obtained results are presented. Finally, in Section 6, conclusions, perspectives and future works are discussed.

2. PROBLEM STATEMENT

We consider a corpus of D road accident reports. Each report, $d_i, i = 1, \dots, D$, has been pre-classified by an expert according to the severity of its consequences, $l_i \in \{0, \dots, L\}$ where 0 is the label associated to the class of reports with associated the least impactful consequences and L to the most impactful consequences. The objective of the present work is to develop a method to automatically extract the critical factors influencing the severity of the accident consequences. The critical factors will be represented by a set of C tokens $F = \{t_c^{crit}, c = 1, \dots, C\}$, where a generic token t_c^{crit} is a combination of one or more adjacent words used in the reports.

3. PROBLEM FORMULATION

The problem of identifying the critical factors influencing the severity of the accidents is here framed as a feature selection problem (Fig. 1). The idea is to represent the reports as mixtures of topics, i.e. weighted distribution of tokens, and, then, to identify those topics which most contribute to the classification of the accident severity. Assuming that from the corpus of reports, $\{d_i, i = 1, \dots, D\}$, a dictionary, $\Delta = \{t_j, j = 1, \dots, T\}$, made by T tokens has been extracted, the problem formulation requires to: *a*) identify K topics $\{\phi_k, k = 1, \dots, K\}$, where a generic topic ϕ_k is represented by the set of weights $w_j^k \in [0,1]$ associated to the token t_j with $\sum_{j=1}^{T} w_j^k = 1$; then, each report d_i is represented by a vector γ_i, whose generic element $\gamma_i^k \in [0,1]$ is the degree of membership of the report d_i to the topic ϕ_k; *b*) consider the topics as features and solve the feature selection problem associated to the classification of a generic report d_i, represented by the vector γ_i, into its corresponding class of severity l_i. To do this, the feature selector is called to identify a minimal set of features which allows correctly classifying the reports. The selected $S < K$ features $\gamma^{k_s^{sel}}, s = 1, \dots, S$, are expected to correspond to the topics $\phi_{k_s^{sel}}, s = 1, \dots, S$ that most contribute to the severity of the accidents and the

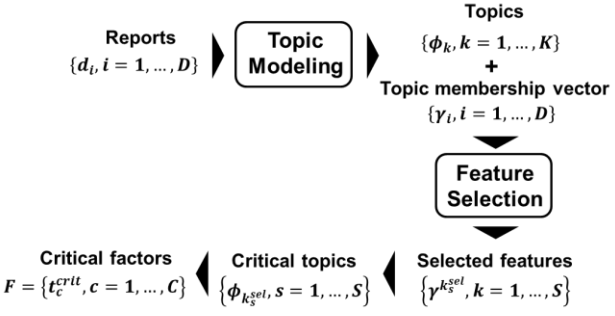

Fig. 1: Problem formulation.

critical factors are the tokens of the selected topics with the largest weights.

4. FRAMEWORK

The framework proposed for the extraction of the critical factors combines:

1) a model to transform the textual reports into numerical vectors, which is based on the sequential application of report preprocessing, Term Frequency Inverse Document Frequency (TFIDF) and HDP topic modelling (Fig. 2);
2) a procedure for the identification of the critical factors based on a wrapper feature selection with SFS as search engine and ANN as classification algorithm.

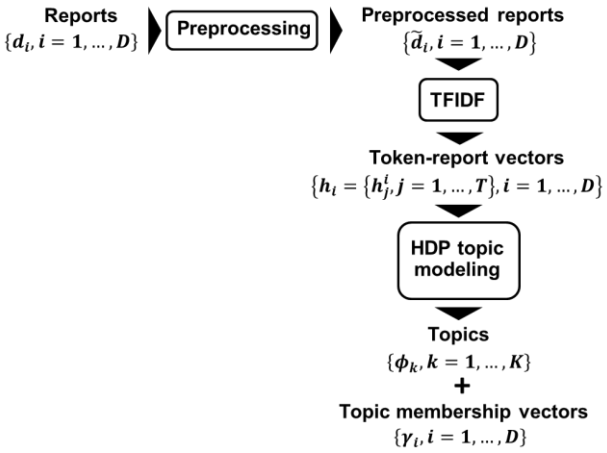

Fig. 2: Model developed to transform the textual reports into numerical vectors.

4.1 Model for the transformation of the textual reports into numerical vectors

The objective of this step is to transform each report d_i, $i = 1, \dots, D$, into a vector γ_i, $i = 1, \dots, D$, of real numbers. This is obtained by the sequential application of text preprocessing for the extraction of a dictionary Δ of T tokens, $\Delta = \{t_j, j = 1, \dots, T\}$, characterizing the corpus of the reports (Section 4.1.1), TFIDF for converting the preprocessed reports into numerical vectors, considering the frequencies of the tokens in the reports (Section 4.1.2), and HDP for topic modeling (Section 4.1.3).

4.1.1 Text preprocessing

The objective of text preprocessing is to convert each report of the corpus into a set of tokens and to define the dictionary $\Delta = \{t_j, j = 1, \dots, T\}$ of the corpus. It 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 (Sojka & Řehůřek, 2010) is used;
d) substituting each word preceded by the negation words '*no*' and '*not*' with "*no_word*" and "*not_word*", respectively, to account for negative sentences and improve the semantic interpretation of the results;
e) identifying frequent bigrams formed by pairs of contiguous words. This is performed by applying a procedure which associates to a generic bigram u_{nm} formed by the unigrams u_n and u_m, with $m \neq n$, the score s_{nm} (Mikolov et al., 2013):

$$s_{nm} = \frac{c(n,m) - a}{c(n)c(m)} \quad (1)$$

where $c(m)$ and $c(n)$ are the counts of the number of times that unigrams u_m and u_n appear in the corpus of reports, respectively, $c(n,m)$ is the count of the number of times that unigrams u_m and u_n appear contiguously in the corpus of reports, and a is a parameter used to establish a minimum number of times that a pair of words should contiguously appear in the reports to constitute a bigram. The list of frequent bigrams B is found by considering only the bigrams with s_{nm} larger than a preset threshold $s_{threshold}$ which allows controlling the total number of bigrams in the dictionary. 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 (Mikolov et al., 2013).

The dictionary Δ of the corpus is defined as the list of tokens $\{t_j : t_j \in U \cup B, j = 1, \dots, T\}$, where U is the list of the preprocessed unique words and B the list of the identified 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.

4.1.2 TFIDF

The corpus of preprocessed reports $\{\tilde{d}_i, i = 1, \dots, D\}$ is converted into a sparse matrix H of size $T \times D$, called token-report matrix, by applying the Term Frequency – Inverse Document Frequency (TFIDF) weighting procedure (Amati & Van Rijsbergen, 2002). The i^{th} preprocessed report is represented by the i^{th} column of the matrix H and the generic element h_j^i is set equal to the product of the frequency, tf_j^i, of the token t_j in the preprocessed report $\tilde{d}_i$, and a quantity inversely proportional to the frequency that the token t_j appears across all the preprocessed reports of the corpus, df_j:

$$h_j^i = tf_j^i \, log\left(\frac{D}{df_j}\right) \quad (2)$$

The second term of the product in Eq. 2 allows lowering the contribution of the most common tokens, which are those present in a large number of reports of the corpus and, therefore, do not provide specific semantic information about the report (Weiss et al., 2005).

4.1.3 HDP

Topic modeling algorithms are statistical methods that infer distributions of words called topics, which represent themes or concepts considering the co-occurrence of words in a corpus of reports (Blei, Carin and Dunson 2010; Griffiths, Steyvers and Tenenbaum 2007). In this work, we employ a topic modelling technique based on the Hierarchical Dirichlet Process (HDP), which allows that different reports share a common set of topics (Teh et al., 2006). The posterior inference of HDP is solved via an approximation algorithm based on the application of Variational Bayes (Wang, Paisley and Blei 2011). HDP receives in input the token-report matrix H (Eq. 2) and provides in output: *a*) the set of topic distributions $\{\phi_k, k = 1, \dots, K\}$, where a generic topic distribution ϕ_k is represented by a set of weights, $w_j^k \in [0,1]$ associated to the token t_j, $j = 1, \dots, T$, of the dictionary Δ, with $\sum_{j=1}^{T} w_j^k = 1$; *b*) the set of topic membership vectors $\{\gamma_i, i = 1, \dots, D\}$, where each vector γ_i is associated to the preprocessed report $\tilde{d}_i$ and whose generic element, $\gamma_i^k \in [0,1]$ with $\sum_{k=1}^{K} \gamma_i^k = 1$, is a measure of the contribution of topic ϕ_k to the description of the preprocessed report $\tilde{d}_i$.

4.2 Feature selection

We consider the feature selection problem associated to the classification of the preprocessed reports $\{\tilde{d}_i, i = 1, \dots, D\}$ into their corresponding severity classes $\{l_i : l_i \in \{0, \dots, L\}, i = 1, \dots, D\}$. As classification model we use an ANN, which receives in input the topic membership vector γ_i and provides in output the report severity class l_i (Fig.3).

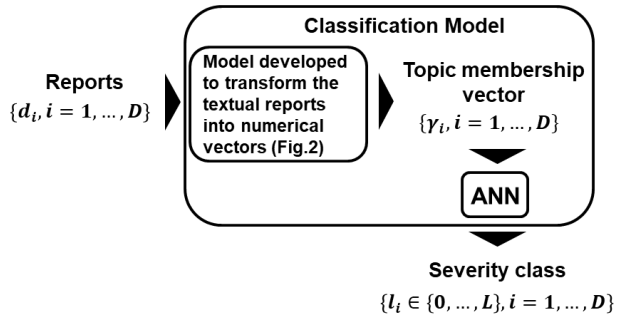

Fig. 3: Classification model.

We apply a wrapper feature selection approach in which the feature selector behaves as a wrapper around the classification algorithm. The feature subsets are compared using as criterion the classification accuracy achieved by the classification algorithm itself. Given the impracticality of an exhaustive search among the 2^K possible feature subsets, an heuristic search strategy is applied, which does not guarantee the identification of the best feature subset. This poses no problems in our case and is justified by the fact that our primary objective is the identification of the most relevant critical factors and not the development of the most accurate classification model. In particular, the Sequential Forward Selection (SFS) algorithm is used, which begins with no features and iteratively adds at each step the feature which allows obtaining the largest accuracy (Marcano-Cedeño et al., 2010; Ververidis & Kotropoulos, 2005). The classification accuracy is computed within a 10-fold cross validation approach to provide a more robust estimation of the accuracy and the search is stopped when adding a new feature does not increase the classification accuracy of more than 1%.

5. CASE STUDY

The developed framework is applied to a repository of accident reports provided by the US National Highway Traffic Safety Administration (NHTSA, 2015) made by 1274 reports, containing a total of 3685 unique words. The average report length is of 1930 words. Each report is formed by:

1) a text containing the narrative of the accident, organized in five sections:
   1.a) header, which reports the accident type and the involved vehicles;
   1.b) weather conditions;
   1.c) infrastructure conditions, such as the state of the pavement and of the illumination;
   1.d) accident dynamics;
   1.e) consequences to the people and vehicles;
2) a number in the set $\{0, \dots, 5\}$ associated to the accident severity, where 0 indicates little to no injury and 5 indicates very serious injury;
3) the total number of convalescence days of the people involved in the accident.

Table 1 reports the narrative, the accident severity and the number of convalescence days of two accidents described in the repository. The information in 1) and 2) is collected by a police officer after the event. A preliminary analysis of the repository has shown that there are significant inconsistencies among the text descriptions of the accident consequences reported in 1e), the classes of severity in 2) and the number of convalescence days in 3). For example, the class of severity 5 has been associated to several reports characterized by minor injuries of the driver according to the description in 1e) and the number of convalescence days. Therefore, the reports have been relabeled into the three macro classes of severity: “Minor to Moderate”, “Serious” and “Severe”, which were used in (Lee et al., 2016).

Table 1: Two examples of accident descriptions taken from the repository.

Narrative of the accident	Severity	Convalescence days
This case involves a non-horizontal impact between the case vehicle a passenger car stopped in traffic, and a passenger car in the process of rolling over. The driver and the right-front passenger the case vehicle were fatally injured. The driver of the vehicle is the case occupant. The case vehicle (V1), a 2009 four-door Nissan Maxima was in the westbound inside through-lane of a wet, concrete four-lane road, stopped at a controlled four-leg intersection. Vehicle two (V2) a 2004 four-door Chevrolet Malibu, was traveling in the outside northbound lane of the intersecting concrete roadway. It was dark but lighted and snowing. For unknown reasons V2 entered a counterclockwise rotation, crossed into the northbound bike lane and stuck the raised median at the gore of the outside northbound through lane and the channel for traffic turning right (east). The impact tripped V2 and it rolled right-side leading and struck a sign post, a traffic signal standard and a utility pole within the raised median before striking the case vehicle. The impact pushed the case vehicle into the traffic signal standard within the raised median for the traffic channel for westbound traffic turning right (north). The case vehicle came to rest facing south-southwest straddling the traffic channel for westbound traffic turning right and the outside westbound though lane. V2 came to rest in the outside on the north side of the intersection straddling northbound bike and through lanes facing south-southeast. The 61-year-old male driver and the 52-year-old female right-front passenger were the only occupants of the case vehicle. Both were restrained by three-point belts, and the steering-wheel, top instrument-panel mounted, both side curtain and seat-back mounted air bags deployed. Both occupants were dead on-scene. The driver was enrolled as a decedent.	5	0
Vehicle 1 (V1) is a 2004 Ford Explorer, four-door (case study vehicle). V1 was occupied by a male, age 30 as its driver and lone occupant. V1's occupant was wearing the available lap and shoulder belt and the frontal airbags deployed as a result of impact. The crash occurred during the hours of daylight. The weather at the time of the crash was clear and dry. The roadway surface was dry. The speed limit is not posted. The crash occurred at a T intersection on a two lane, two way roadway.. The roadway is level in the area of the crash. V1 came to the end of the roadway at a T intersection and departed the roadway, driving through a raised concrete curb (event 1) and a uphill grass embankment before striking another curb and entering a parking lot (event 2). V1 then struck a parked 2001 Chrysler Sebring (event 3). V1 then continued straight, striking a raised concrete curb with its left front tire (event 4) and then its left rear tire (event 5). Event four caused the left front tire to be ripped from the axle. V1 continued on past the curb and struck a tree head-on (event 6). V1 continued for a few meters until it struck a chain-link fence pole (event 7), where it came to final rest. The case study participant was transported by air to a trauma center.	3	15

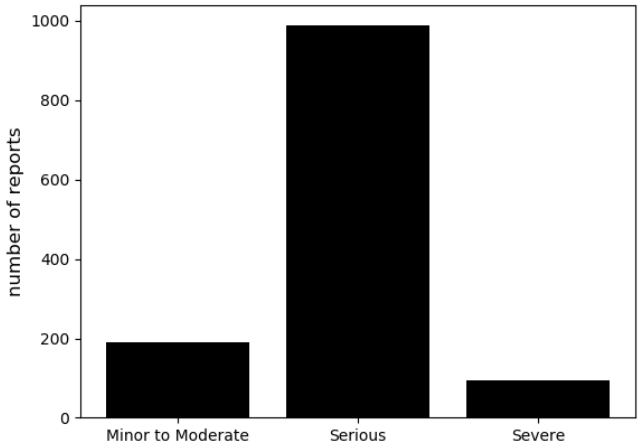

Fig. 4: Distribution of the reports in the three classes of severity.

Notice that the new classification of the events considers also the information about the number of convalescence days, which is less subjective than the label assigned by the police officer. The rules used for the assignment of the new classes to a generic report d_i are:

1) if the class assigned by the police officer is lower than 2 and the number of convalescence days are lower than 7, then the new class is “Minor to Moderate”;
2) if the class assigned by the police officer is 2 or 3 and the number of convalescence days is lower than 14, then the new class is “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: *a*) to reduce the subjectivity of the assignments made by the police officers by decreasing the number of classes from six to three and *b*) to be conservative by choosing the worst condition between the police officer assignment and the number of convalescence days. Notice

that the distribution of the reports in the three classes of severity is unbalanced (Fig. 4).

5.1 Framework development

In this case study, we focus on the identification of the factors which are critical for the accident severity considering the analysis of the two classes characterized by the mildest and the most severe consequences, i.e. the "Minor to Moderate" and "Severe" classes. This is motivated by the observation that the most critical factors are expected to cause major modifications of the accident consequence, i.e. from minor to severe. The class "Minor to Moderate" will be represented by the label 0 and the class "Severe" by the label 1. A side effect of considering only these two classes is that the total number of reports is reduced from 1274 to 285. Notice, however, that the intermediate class "Serious", which contains 989 reports, is expected to share common characteristics with the other two classes, and, therefore, to contain information on factors less critical with respect to the severity of the consequences.

The length of the reports has been reduced by considering only the weather (1b) and infrastructure condition (1c) subsections of the reports, which contain the most interesting part of the text with respect to the identification of the critical factors. The selected $D = 285$ reports are preprocessed following the procedure described in Section 4.1.1. The resulting dictionary contains $T = 97$ tokens. The preprocessed corpus of reports, $\{(\tilde{d}_i, l_i), i = 1, \dots, D\}$, where a generic pair $(\tilde{d}_i, l_i)$ is composed by the preprocessed report $\tilde{d}_i$ and its associated class $l_i \in \{0,1\}$, is divided into a training and a test set made of $D_{train} = 234$ and $D_{test} = 51$ reports, respectively. Data augmentation is applied to obtain a balanced and larger training set of reports, with the objective of improving the quality of the topics extracted from the repositories and the accuracy of the classification (Liu et al., 2020; Zhao et al., 2019). In particular, the data augmentation technique reported in (Wei & Zou, 2019) is applied in this work. It generates a new report from an old report 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 corpus of reports made by a similar number of reports of the two classes, the augmentation procedure is repeated four times for each report of class 0 and one time for each report of class 1. Eventually, a training set formed by $D_{aug} = 885$ reports (460 of class 0 and 425 of class 1) is obtained.

Class 0 and class 1 are characterized by 75 tokens out of 97 and 68 tokens out of 97, respectively, with 23 tokens being shared among them. Notice that data augmentation does not change the distribution of the number of tokens in a class, since by making copies of the reports we are also making copies of the tokens contained in them, and, therefore, the ratio of these two numbers remains constant in each class. Also, to allow identifying the critical factors by using original reports, data augmentation is not applied to the test set.

Bigrams have been identified by setting the parameters a and s_{thresh} introduced in Section 4.1.1 equal to 1 and 0.01, respectively, in accordance to (Mikolov et al., 2013). The number of topics, K, searched by the HDP has been set equal to 20 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 (Chen et al., 2011). A fully connected feedforward ANN, with an architecture characterized by twenty neurons in the input layer, ten neurons in the hidden layer and two neurons in the output layer, is developed for the classification of the reports severity class, l_i, from the 20-dimensional vectors, γ_i, provided by the HDP. 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 new test report, the ANN produces output values o_{l_i} between 0 and 1, which can be interpreted as the degree of confidence in the classification of the report to the corresponding class l_i (Nwankpa et al., 2018). Finally, the test report is assigned to the class with the associated largest degree of confidence.

5.2 Results

Fig. 5 shows the confusion matrix of the developed ANN classification model, obtained applying a 10-fold cross validation procedure. Table 2 reports the classification performance, in terms of accuracy (ratio between the numbers of correctly classified and tested reports) and $F_{measure}$:

$$F_{measure} = \frac{2}{\frac{TP+FN}{TP} + \frac{TP+FP}{TP}} \quad (3)$$

where TP, TN, FP and FN are the numbers of true positives, true negatives, false positives and false negatives, respectively.

Table. 2: Comparison of the classification performances of the ANN fed by all 20 features, the ANN fed by the 2 selected features and a classifier that assigns the class of severity "Minor to Moderate to all test reports. The classification accuracy and $F_{measure}$ as average $\pm$ standard deviation over the 10 folds considered in the cross validation.

Model	Classification accuracy	$F_{measure}$
ANN with inputs $\gamma_i^k, k = 1, \dots, K$	0.75 ± 0.03	0.71 ± 0.03
ANN with inputs $\tilde{\gamma}_i^4, \tilde{\gamma}_i^7$	0.88 ± 0.06	0.69 ± 0.04
Always class "Minor to Moderate"	0.67	0.65

The latter metric has been considered since it provides an accurate performance estimation in case of unbalanced datasets (Pereira & Saraiva, 2020). From Table 2, it can be seen that, despite the tendency of the classifier to assign the class of severity "Minor to Moderate" to some reports whose true class of severity is "Severe", the obtained classification performance confirms the capabilities of the HDP of extracting features $\gamma_i^k, k = 1, \dots 20$, providing an overall satisfactory representation of the reports and overperforming a classifier which assigns all reports to the majority class "Minor to Moderate".

Since the final objective of the work is the identification of critical factors and not the maximization of the accuracy in the report classification, other techniques which allow directly classifying documents without the intermediate step of topic modeling, such as TFIDF combined with logistic regression or random forest, are not considered, even if they can lead more accurate classification results (Altinel & Ganiz, 2018). The SFS-based feature selection is, then, applied. The ANN developed to classify the severity receives in input the features subset proposed by the SFS algorithm. In case of multidimensional feature subsets, the degrees of membership to the selected topics, $\gamma_i^{k_s^{sel}}, s = 1, \dots S$, are normalized by:

$$\tilde{\gamma}_i^{k_s^{sel}} = \frac{\gamma_i^{k_s^{sel}}}{\sum_{s=1}^{S} \gamma_i^{k_s^{sel}}} \quad (4)$$

so that that their sum is equal to 1.

In this case study, the search stops at the second iteration and selects the feature subset corresponding to topics 4 and 7. Fig. 6 shows the confusion matrix of the ANN fed by the 2-dimensional vector $\tilde{\gamma}_i = (\tilde{\gamma}_i^4, \tilde{\gamma}_i^7)$ formed by the degrees of membership to topics 4 and 7. Notice that, despite the significant reduction of the information provided in input, the

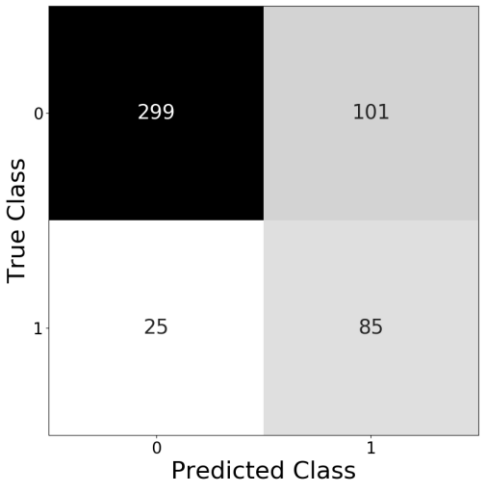

Fig. 5: Confusion matrix of the developed ANN receiving in input all 20 features $(\gamma_1, \dots, \gamma_{20})$.

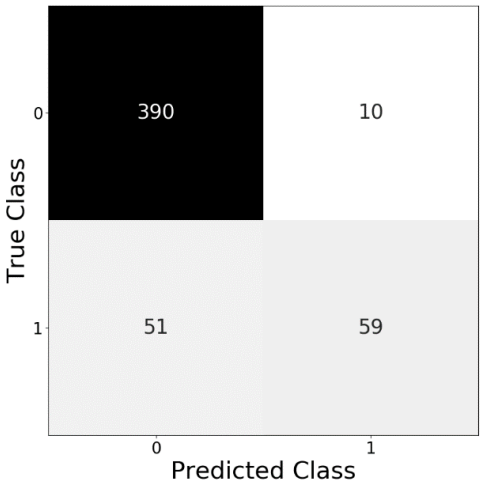

Fig. 6: Confusion matrix of the developed ANN receiving in input features $(\tilde{\gamma}_i^4, \tilde{\gamma}_i^7)$.

model fed by only two features provides classification performances similar to those of the model fed by all the twenty features (Table 2). In particular, the model fed by two features is very accurate in the classification of the class of severity "Minor to Moderate", whereas it tends to underestimate the severity of the reports whose true class is "Severe". Lack of conservativeness of the method, which indicates that the selected topics are not able to identify all tokens characterizing severe accidents, will be object of future research. A possible solution to overtake the problem is the use of a function which penalizes non-conservative classification as criterion for the feature selection.

Table 3 reports an example of accident whose true class is "Severe", which is incorrectly assigned to the class "Minor to Moderate": the reason of the misclassification is that the report contains tokens, such as "*dry*" and "*clear*", that typically characterize the "Minor to Moderate" class. Fig. 7 shows the ANN classification of a test set artificially generated, which contains 100 patterns $\tilde{\gamma}_i = (\tilde{\gamma}_i^4, \tilde{\gamma}_i^7) = [0,1], [0.01,0.99], [0.02,0.98], \dots, [1,0]$.

Table 3: Example of misclassified reports in the test set.

Report text	$(\tilde{\gamma}_i^4, \tilde{\gamma}_i^7)$	True class	Classification result
This two-vehicle collision occurred during the post-midnight hours (dark, streetlights present), of a winter weekend, at the intersection of a north/south trafficway and an east/west trafficway. At the time of the crash the weather was clear and the roadway surfaces were dry.	$(0.9371, 0.0629)$	1	0

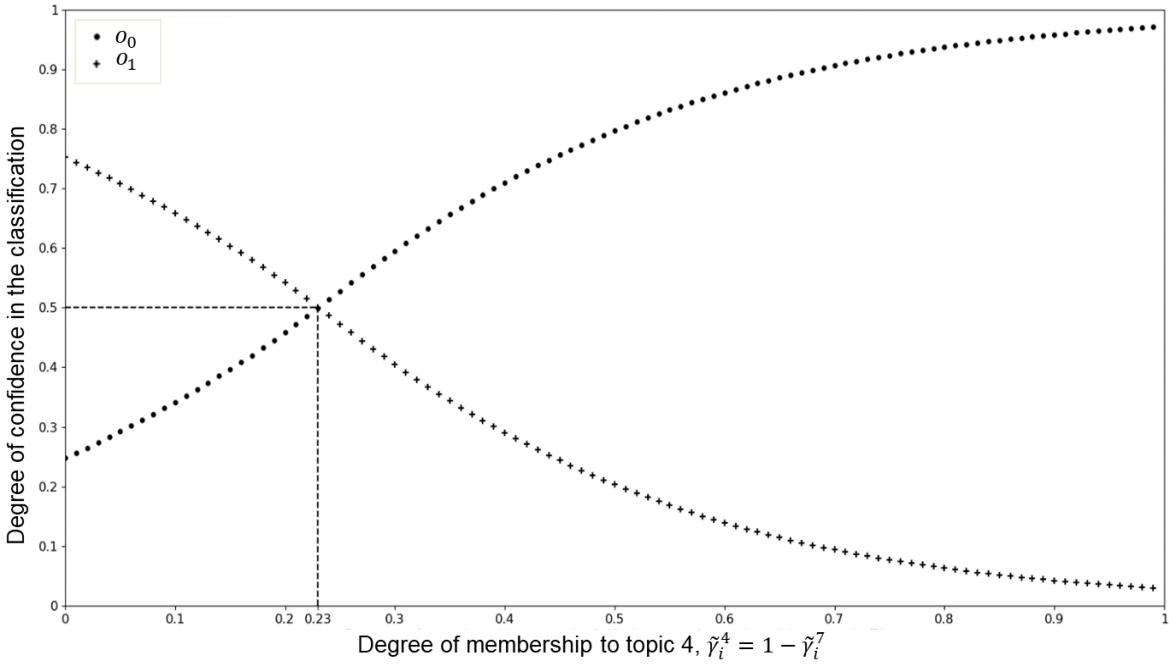

Fig. 7: Classification output for the artificial vector of degrees of membership to topics 4 and 7.

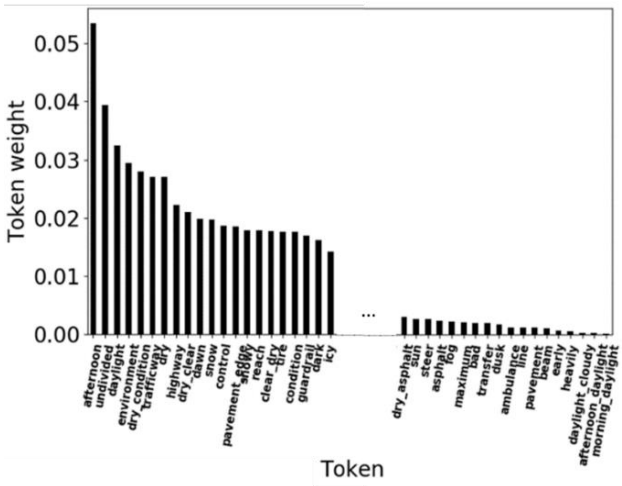

Fig. 8: distribution of topic 4

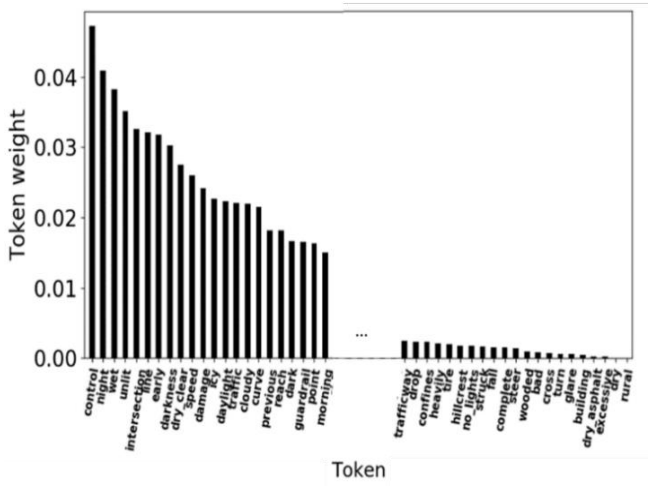

Fig. 9: distribution of topic 7

Table 4: Collection of the top fifteen words for the topic 4 and 7.

Topic	Top words
4	*afternoon, undivided, daylight, environment, dry_condition, trafficway, dry, highway, dry_clear, dawn, snow, control, pavement_edge, snowy, reach*
7	*control, night, wet, unlit, intersection, line, early, darkness, dry_clear, speed, damage, icy, daylight, traffic, cloudy*

Notice that when the degree of membership to topic 4, $\tilde{\gamma}_i^4$, exceeds 0.23, the degree of confidence in the class 0 assignment, o_0, becomes greater than 0.5: therefore, topic 4 can be associated to the class of severity "Minor to Moderate", i.e. a large degree of membership of a report to topic 4 indicates that its class is 0, and, on the opposite, topic 7 can be associated to the class of severity "Severe".

Table 4 reports the fifteen tokens with the associated largest weights of topics 4 and 7, and Figures 8 and 9 show the corresponding distributions of weights. As expected, tokens indicating favourable road and weather conditions, such as "*dry*", "*dry*" combined with "*condition*" and "*clear*", "*dry_clear*", "*afternoon*", "*daylight*", are associated to topic 4 and, therefore, characterize the severity class "Minor to Moderate". It is interesting to observe the presence of the tokens "*snow*", "*snowy*" and "*undivided*", which indicates the

absence of road surface markings, among those with the associated largest weights in topic 4. A possible interpretation is that in case of snow or undivided roads, drivers have a clear perception of the hazard and tend to be more cautious, e.g. by reducing the speed. Therefore, although accidents in case of snow or undivided roads occur, their consequences are likely to be not severe. With respect to the occurrence of severe conditions, the identified critical factors are associated to the speed of the car, the presence of intersections and other characteristics of the road infrastructure, such as the condition of the asphalt ("*wet*", "*icy*", "*dry_clear*") and of the illumination ("*night*", "*darkness*", "*unlit*", "*daylight*", "*cloudy*").

In conclusion, the developed framework applied to road accident reports provided four critical factors and a list of tokens associated to them, reported in Table 5. The semantic analysis of the identified critical factors, their relationship and their frequency in the accident reports can support domain experts in the identification and implementation of mitigative solutions. Nevertheless, since this activity would require the intervention of domain knowledge, it is out of scope of this work.

Table 5: Identified critical factors and associated tokens.

Critical factor	Associated tokens
Speed of the car	*speed, pavement_edge,*
Presence of intersections	*intersection, line, undivided*
Asphalt conditions	*dry_condition, dry, dry_clear, snow, snowy, wet, icy, damage, pavement_edge*
Illumination conditions	*afternoon, daylight, environment, dawn, night, unlit, early, darkness, daylight, cloudy*

6. CONCLUSIONS AND PERSPECTIVES

A framework combining Natural Language Processing (NLP) and Sequential Forward Selection (SFS) has been developed for the identification of critical factors influencing the severity of the consequences of car accidents. It has been applied to a public repository of road accidents collected from local authorities. The obtained results have shown that the topics identified by the feature selector contain tokens which refer to factors that allow classifying the reports in the correct class of severity and are expected to influence the consequences of the car accidents. Future work will be devoted to: *a*) improve the topic modelling algorithm to allow processing longer documents and repositories containing larger number of reports; *b*) improve the classification performance by exploring different classification methods within the developed framework for the identification of critical factors; *c*) improve the feature selection algorithm to allow exploring all the space of the possible features combinations, without suffering the "nesting effect" of the sequential forward selection search, which cannot remove features, and, therefore, could fall into local minima; *d*) improve the feature selection algorithm to penalize subset of topics which do not provide conservative classification of the accident severity; *e*) integrate the developed framework with a domain knowledge based analysis of the semantic of the critical factors, their relationship and their frequency in the accident reports; *f*) integrate the developed framework into a decision making scheme to allow public authorities identifying the protective and mitigation interventions needed to improve road safety. Also, the possibility of using NLP to automatically identify factors critical with respect to the frequency of occurrence of the accidents will be considered. The final objective is to develop a framework that informs the decision making on the proper intervention strategies needed to be implemented for improving road safety.

ACKNOWLEDGEMENTS

The participation of Piero Baraldi and Enrico Zio has been funded 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.

REFERENCES

Altinel, B., & Ganiz, M. C. (2018). Semantic text classification : A survey of past and recent advances. *Information Processing and Management*, *54*, 1129–1153. https://doi.org/10.1016/j.ipm.2018.08.001

Amati, G., & Van Rijsbergen, C. J. (2002). Probabilistic Models of Information Retrieval Based on Measuring the Divergence from Randomness. *ACM Transactions on Information Systems*, *20*(October), 357–389.

Ansaldi, S. M., Bragatto, P., Agnello, P., & Milazzo, M. F. (2020). An Ontology for the Management of Equipment Ageing. *The 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference*, 978–981. https://doi.org/10.3850/978-981-14-8593-0

Bezerra, C., de Santana, J. M. M., Moura, M. das C., & Lins, I. D. (2020). Automated classification of injury leave based on accident description and natural language processing. *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference, Venice, Italy*. https://doi.org/10.3850/981-973-0000-00-0

Bin, C., Baigen, C., & Wei, S. (2017). Text Mining in Fault Analysis for On-board Equipment of High-speed Train Control System. *Chinese Automation Congress (CAC), Jinan, China*, 6907-6911. https://doi.org/10.1109/CAC.2017.8244022

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine*, *27*(6), 55–65. https://doi.org/10.1109/MSP.2010.938079

Chen, M., Ji, X., & Shen, D. (2011). Short Text Classification Improved by Learning Multi-Granularity Topics. *Proceedings of the 22nd International Joint Conference of Artificial Intelligence*, 1776–1781.

European Commission. (2019). *EU Road Safety Policy Framework 2021-2030 - Next steps towards "Vision Zero"* (Issue COMMISSION STAFF WORKING DOCUMENT).

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in Semantic Representation. *Psychological Review*, *114*(2), 211–244. https://doi.org/10.1037/0033-295X.114.2.211

Guimarães, M. S., Gomes de Araújo, H. H., Lucas, T. C., Moura, M. das C., Lins, I. D., & Vilela, R. F. T. (2020). An NLP and Text Mining – based approach to categorize occupational accidents. *Proceedings of the 30th European Safety and Reliability Conference and the 15th Probabilistic Safety Assessment and Management Conference, Venice, Italy*.

Imprialou, M., & Quddus, M. (2019). Crash data quality for road safety research : Current state and future directions. *Accident Analysis and Prevention*, *130*, 84–90. https://doi.org/10.1016/j.aap.2017.02.022

Lee, J. S., Kim, Y. H., Yun, J. S., Jung, S. E., Chae, C. S., & Chung, M. J. (2016). Characteristics of Patients Injured in Road Traffic Accidents According to the New Injury Severity Score. *Annals of Rehabilitation Medicine*, *40*(2), 288–293. https://doi.org/http://dx.doi.org/10.5535/arm.2016.40.2.288

Liu, S., Lee, K., & Lee, I. (2020). Document-level multi-topic sentiment classification of Email data with BiLSTM and data augmentation. *Knowledge-Based Systems*, *197*. https://doi.org/10.1016/j.knosys.2020.105918

Marcano-Cedeño, A., Quintanilla-Domínguez, J., Cortina-Januchs, M. G., & Andina, D. (2010). Feature Selection Using Sequential Forward Selection and classification applying Artificial Metaplasticity Neural Network. *IECON 2010 - 36th Annual Conference on IEEE Industrial Electronics Society, Glendale, AZ*, 2845–2850. https://doi.org/10.1109/IECON.2010.5675075

Mauni, H. Z., Hossain, T., & Rab, R. (2020). Classification of Underrepresented Text Data in an Imbalanced Dataset Using Deep Neural Network. *IEEE Region 10 Symposium (TENSYMP), June*, 997–1000. https://doi.org/10.1109/TENSYMP50017.2020.9231021

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

Mishu, S. Z., & Rafiuddin, S. (2016). Performance Analysis of Supervised Machine Learning Algorithms for Text Classification. *19th International Conference on Computer and Information Technology (ICCIT)*, 409–413. https://doi.org/10.1109/ICCITECHN.2016.7860233

Morgan, K., Kwigizile, V., Lee, K., & Oh, J. (2021). Discovering latent themes in traffic fatal crash narratives using text mining analytics and network topology. *Accident Analysis and Prevention*, *150*(November 2020), 105899. https://doi.org/10.1016/j.aap.2020.105899

NHTSA. (2015). *Crash Injury Research (CIREN), Data*. https://one.nhtsa.gov/Research/Crash-Injury-Research-(CIREN)/Data:

Nwankpa, C. E., Ijomah, W., Gachagan, A., & Marshall, S. (2018). Activation Functions : Comparison of Trends in Practice and Research for Deep Learning. *ArXiv:1811.03378v1*, 1–20.

Pereira, J., & Saraiva, F. (2020). A Comparative Analysis of Unbalanced Data Handling Techniques for Machine Learning Algorithms to Electricity Theft Detection. *2020 IEEE Congress on Evolutionary Computation (CEC)*, 1–8. https://doi.org/10.1109/CEC48606.2020.9185822

Persia, L., Shingo, D., Simone, F. De, Feypell, V., Beaumelle, D. La, Yannis, G., Laiou, A., Han, S., Machata, K., Pennisi, L., Marchesini, P., & Salathè, M. (2016). Management of road infrastructure safety. *Transportation Research Procedia*, *14*, 3436–3445. https://doi.org/10.1016/j.trpro.2016.05.303

Sarkar, S., Vinay, S., & Maiti, J. (2016). Text mining based safety risk assessment and prediction of occupational accidents in a steel plant. *2016 International Conference on Computational Techniques in Information and Communication Technologies, ICCTICT 2016 - Proceedings*, 439–444. https://doi.org/10.1109/ICCTICT.2016.7514621

Sojka, P., & Řehůřek, R. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceeding of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, 1566–1581. https://doi.org/10.1198/016214506000000302

Ververidis, D., & Kotropoulos, C. (2005). Sequential Forward Feature Selection with Low Computational Cost. *13th European Signal Processing Conference*, 1–4.

Wang, Chao, Quddus, M. A., & Ison, S. G. (2013). The effect of traffic and road characteristics on road safety : A review and future research direction. *Safety Science*, *57*, 264–275. https://doi.org/10.1016/j.ssci.2013.02.012

Wang, Chong, Paisley, J., & Blei, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. *Journal of Machine Learning Research*, *15*, 752–760.

Wei, J., & Zou, K. (2019). EDA : Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, 6382–6388.

Weiss, S. M., Indurkhya, N., Zhang, T., & Damerau, F. J. (2005). *Text Mining. Predicitive Methods for Analyzing Unstructured Information*. Springer.

Williams, T., & Betak, J. (2018). A Comparison of LSA and LDA for the Analysis of Railroad Accident Text. *Procedia Computer Science*, *130*, 98–102. https://doi.org/10.1016/j.procs.2018.04.017

World Health Organization. (2018). *Global status report on road safety*.

Yang, Z., Baraldi, P., & Zio, E. (2020). A novel method for maintenance record clustering and its application to a case study of maintenance optimization. *Reliability Engineering and System Safety*, *203*(April), 107103. https://doi.org/10.1016/j.ress.2020.107103

Zaghloul, W., Lee, S. M., & Trimi, S. (2009). Text classification : neural networks vs support vector machines. *Industrial Management & Data Systems*, *109*(5), 708–717. https://doi.org/10.1108/02635570910957669

Zhao, H., Du, L., Buntine, W., & Liu, G. (2019). Leveraging external information in topic modelling. *Knowledge and Information Systems*, *61*(2), 661–693. https://doi.org/10.1007/s10115-018-1213-y

NOMENCLATURE

NLP	Natural Language Processing
HDP	Hierarchical Dirichlet Process
SFS	Sequential Forward Selection
ANN	Artificial Neural Network
TFIDF	Term Frequency Inverse Document Frequency
d_i	generic i-th report
D	Total number of reports
l_i	Label associated to the class of severity of the generic i-th report
L	Label associated to the class of reports with the most impactful consequences
Δ	Dictionary of the corpus of reports
t_j	Generic j-th token in the dictionary
T	Total number of tokens in the dictionary
t_c^{crit}	Generic c-th critical factor
F	Set of all critical factors
C	Total number of critical factors
ϕ_k	Generic k-th topic
K	Total number of topics
w_j^k	Weight of the j-th token of the dictionary in the k-th topic distribution
γ_i	Vector of the memberships of the i-th report to the topics
γ_i^k	Membership of the i-th report to the k-th topic
$\gamma^{k_s^{sel}}$	s-th feature of γ_i selected by the feature selector
$\tilde{\gamma}_i^{k_s^{sel}}$	Normalized degree of membership of the i-th report to the s-th topic
S	Number of features selected by the feature selector
$\phi_{k_s^{sel}}$	Topic corresponding to the s-th feature selected by the developed method
u_n	n-th unigram
u_{nm}	Bigram composed by the unigram u_n followed by the unigram u_m with $n \neq m$
s_{nm}	Score associated to the generic bigram u_{nm}
$c(n)$	Number of occurrences of the unigram u_n
$c(n,m)$	Number of occurrences of the bigram u_{nm}
a	Parameter used to establish the minimum number of times the unigram u_m should appear contiguously to the unigram u_n in a report, to constitute the bigram u_{nm}
B	Set of identified bigrams
U	Set of all unigrams
$c(n)$	Number of the occurrences of the unigram u_n
$c(n,m)$	Number of the occurrences of the bigram u_{nm}
$s_{threshold}$	Threshold on s_{nm} used to identify the bigrams
$\tilde{d}_i$	Generic preprocessed i-th report
H	Token-report matrix
h_j^i	TFIDF weight associated to the j-th token in the preprocessed i-th report
tf_j^i	Frequency of the j-th token in the preprocessed i-th report
df_j	Frequency of the j-th token across all the preprocessed reports of the corpus
D_{aug}	Total number of preprocessed reports forming the training set
D_{test}	Total number of preprocessed reports foming the test set
H_{aug}	Token-report matrix of the training set
H_{test}	Token-report matrix of the test set
o_{l_i}	Output of the ANN classifier associated to class l_i
$F_{measure}$	F-measure metric