Using a Social Network for Road Accidents Detection, Geolocation and Notification - A Machine Learning Approach

Samuel Pereira de Vasconcelos
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
email: samuel.vasconcelos@ccc.ufcg.edu.br

Cláudio de Souza Baptista
Information Systems Laboratory
Federal University of Campina Grande
Campina Grande - PB, Brazil
email: baptista@computacao.ufcg.edu.br

Hugo Feitosa de Figueiredo
Information Systems Laboratory
Federal Institute of Technology of Paraiba
Esperança – PB, Brazil
email: hugo.figueiredo@ifpb.edu.br

***Abstract* - The road system is the main means of transport used in Brazil. Traffic accidents are quite common in this mode of transport, incurring one of the biggest causes of death in the country. Profiles on social networks of the Federal Highway Police (PRF) and other sources of information, contribute to alert drivers as quickly as possible about road accidents that have occurred, in order to prevent other accidents from occurring. Also, such information can be used by drivers about possible delays, or even deviations in their paths. However, accessing such information via text while driving is illegal and further increases the risk of accidents. Therefore, this paper addresses a study about reliable posts in social networks, in particular Twitter, to create a supervised classification model, which is capable of classifying tweets about the occurrence or not of accidents. The results include the best induction model obtained for classifying tweets, among several analyzed, as well as the construction of a mobile application that can notify through audio drivers about accidents reported on their way, in real time.**

Keywords* - *Road accidents; Machine learning, Natural Language Processing; Geoprocessing; Mobile Computing; Social media.

I. INTRODUCTION

Road accidents have been one of the most important challenges of contemporary society worldwide, being the cause of concern and studies in several countries. Road accidents are the eighth leading cause of death in the world [1].

On the other hand, we live in a highly connected world. Social networks may help in disseminating traffic accident alerts in order to attract the attention of those who are traveling near the accident hotspot in order to prevent further accidents.

Thus, developing technologies that allow road accidents alerts to reach, in real time and with reliability, the drivers who will cross such an accident is a challenge of socio-economic and public health interest, which gives rise to the focus of this work.

There are several channels on the Internet to alert the population about accidents, most of them in the form of text (e.g., tweets), which makes it impossible for drivers while traveling to access them. Also, there are applications for smartphones such as waze that aims to alert drivers while traveling on the highways; however, such applications are not completely reliable, as it comes from crowdsourcing [2]. Hence, using official communication channels, such as the Twitter profiles from the government, avoids misinformation.

This work proposes a framework that can automatically identify, monitor and alert drivers through audio and in real time about the incidence of road accidents using machine learning, geoprocessing and Natural Language Processing techniques based on tweets from authorities, such as the Brazilian Federal Highway Police (PRF).

The remainder of this paper is structured as follows. Section II discusses some related works on road accidents. Section III presents the methodology used in this research. In Section IV, the experiments carried out to identify the best machine learning model for the purpose of this research are discussed, as well as the analysis of the results obtained. Finally, Section V concludes the paper and presents future guidelines to continue this research.

II. RELATED WORK

There are several works that address different aspects of road accidents around the world. There are studies on the social, economic and environmental impacts of road accidents [3][4]. Other studies are focused on victims of road accidents, whether fatal or even sequelae [5][6]. Some authors are concerned with detecting potential patterns in road accidents [7]-[11]. Also, there are studies aimed at analyzing the main causes of road accidents [12]-[18]. Finally, there are studies aimed at classifying and predicting the severity of accidents [19]-[27].

Katsoukis et al. [28] used data mining techniques to classify the risk of accidents in regions in Greece, considering the number of accidents in a given region. Ryder and Wortmann [29] proposed an approach to detect and classify places with a high rate of accidents; aiming to alert drivers in real time about imminent dangers on the highways, through a mobile application.

Ryder et al. [29] developed a decision support system to prevent road accidents. This system sends alerts to drivers when they are close to high accident risk areas.

Ren et al. [30] proposed a deep learning model, through a neural network Long Short-Term Memory (LSTM), based on spatiotemporal data correlation in order to predict traffic accident areas in Beijing.

Zhang et al. [31] proposed a method to identify key points where the incidence of road accidents is high in a given period of time. The dataset contains characteristics such as: holiday, day of the week, time, crash site type, and weather conditions. Yu et al. [32] proposed a neural network model to predict accidents on highways, called STEEN, which combines spatial distributions, temporal dynamics and external factors, such as Points of Interest (POIs).

To the best of our knowledge, there is no work that simultaneously uses machine learning techniques, natural language processing and geoprocessing for the problem of detecting road accidents and their communication to drivers. The BERT transformer obtained the best results in our experiments. We experimented several machine learning models based on supervised classification. We use NLP regular expressions and preprocessing techniques. Finally, we georeferenced tweets from texts and store them in a spatial database system.

This integrated solution consists of the main contribution of this paper.

III. Methodology

The methodology used in this research was based on the Cross Industry Standard Process for Data Mining (CRISP-DM) [33]. This methodology consists of six steps: (i) Business Understanding, (ii) Data Understanding, (iii) Data Preparation, (iv) Modeling, (v) Evaluation and (vi) Deployment.

The Business Understanding step was covered in Section I. The Data Understanding step consisted of identifying the PRF profiles on Twitter and analyzing the keywords to be used to gather the tweets. The Data Preparation step addressed preprocessing and data labeling to create an annotated corpus. The Model Induction step focused on deploying several induction models, using supervised classification, with hyperparameter tuning. The Evaluation step carried out the analysis of the results of the models that were generated in the previous stage, choosing the model that presented the best performance. Finally, the Deployment step consisted of implementing the model in the SafeTrip tool in a mobile platform.

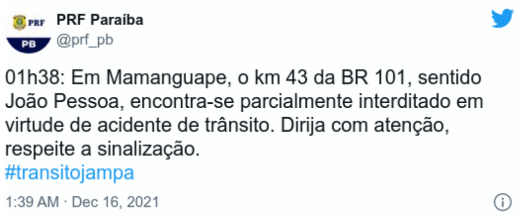
In the following, we detail the methodology steps.

A. Data Understanding

This step consists of selecting the Twitter PRF profiles; the choice of keywords used in road accidents; as well as the tweet gathering process.

1) Twitter Profile Selection: During the search for Twitter profiles, to avoid data poisoning, official PRF profiles from several states in Brazil were chosen, due to reliability on reporting road accidents. The chosen profiles were all active by November 2021 including: @PRFParana, @PRFCeara, @PRF191RJ, @PRF191PR @PRF191TOCANTINS, @PRF191SP, @PRF191SERGIPE, @PRF191RORAIMA, @PRF191RONDONIA, @PRF191PE, @PRF191PA, @PRF191MS, @PRF191ES, @PRF191AM, @prf_sc, @PRF191ACRE, @prf_rn, @prf_pi, @prf_pb, @PRF_MS, @prf_mg, @prf_df, @prf_ba e @prf_al.

2) Keyword Selection: We chose keywords commonly used in posts about road accidents in Twitter PRF profiles. The words chosen were: ’wounded’, ’accident’, ’death’, ’shock’, ’tragedy’, ’run over’, ’overturn’, ’victim’, ’collision’, ’turn’, ’turned’, ’ pileup’, ’fire’, ’crash’, ’tip over’, ’crash’, ’left the runway’, ’crashed’, ’dies’, ’passed away’ and ’fell over’. Figure 1 depicts an example of tweet from the Brazilian Federal Highway Police (PRF) of the State of Paraiba.



PRF Paraíba
@prf_pb

01h38: Em Mamanguape, o km 43 da BR 101, sentido João Pessoa, encontra-se parcialmente interditado em virtude de acidente de trânsito. Dirija com atenção, respeite a sinalização.
#transitojampa

1:39 AM · Dec 16, 2021

Figure 1. Example of a tweet with a highway accident alert.

3) Tweets Gathering Process: After choosing the profiles and keywords related to traffic accidents, requests were made to the Twitter API. The data returned by the Twitter API include: identification number, user, date, place of publication, text, and likes.

B. Data Preparation

This step carried out the manual data labeling and tweet pre-processing.

1) Data Labeling: We performed the manual labeling of tweets as follows. We label a positive class when the content contains information about traffic accidents, including the location where the incident occurred; whereas we label the negative class when the tweet did not contain data on traffic accidents. The resulting corpus has a structure similar to Table I.

TABLE I
Tweet classification example.

Tweet	Label
”the Dom Pedro road has stretches with rain this afternoon, drivers should redouble their attention, works that were carried out this friday have already been closed, there are no accidents and traffic flows well”	NEGATIVE
”attention br 116 km 643 jequie road partially closed due to accident reduce speed”	POSITIVE

At the end of the data labeling process, 3,311 tweets were obtained, but the dataset was quite unbalanced, the negative class occupied about 75% of the records, as shown in Figure 2.

2) Data Preprocessing: The collected data were preprocessed using several NLP techniques . First we converted text to lowercase, and removed non-relevant information such as links, emoticons and punctuation from the tweets using regular expressions. Also, we performed tokenization and removed stopwords.

We also performed corpus balancing to avoid possible bias in the classification models to be used. For that we used the statistical technique of subsampling [34]. Thus, the balanced set resulted in 832 instances for each class.

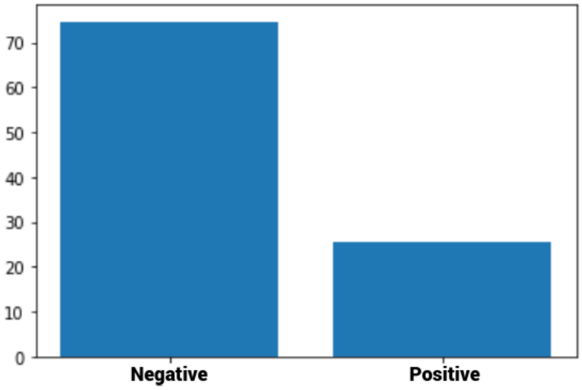


Figure 2. Labeled tweets distribution.

IV. Model Induction and Evaluation

In order to find out the best model that fits our solution, we ran several classification algorithms and compared their results.

A. Model Setup

To obtain the best supervised classification model for our proposed problem we used the following machine learning techniques: Naive Bayes, Logistic Regression, XGBoost, Support Vector Machine (SVM), Random Forest and the BERT transformer.

Our dataset was split into training and testing, 20% of the data were allocated to the test set, while 80% were allocated to training, with 10% of this training set being used during training as a validation set. To reduce the chances of overfitting, we used the k-folds cross-validation method with the parameter k = 10.

We used the sklearn GridSearchCV algorithm for tuning the model's hyperparameters o during training. We used accuracy, precision, recall and f1-score metrics to evaluate the models performance. Different combinations of hyperparameters were tested in order to find the model that best maximized the accuracy and f1-score metrics. The corpus and all the code used for training the models can be accessed through *google colaboratory* [35].

Concerning the transformer classifier, we used BERTimbau. BERTimbau is a pre-trained neural network to deal with Portuguese. For the model induction, pytorch tensors were used for 4 epochs, number recommended by the literature, using the same training set of the sklearn models. The code and corpus used can be accessed through *google colaboratory* [36].

B. Model Evaluation

In this subsection, we present and analyze the results of the classification models in order to choose the best induction model for the road accidents classification problem.

Table II presents the metrics from the several classification algorithms used. BertTimbau was the best induction model.

TABLE II
Models Evaluation.

	Métrics			
Model	Accuracy	Precision	Recall	F1
RandomForestClassifier	91,89%	87,36%	97,55%	92,17%
LinearSVC	92,19%	88,89%	96,39%	92,49%
MultinomialNB	93,69%	89,20%	98,74%	93,73%
LogisticRegression	94,89%	92,77%	96,86%	94,77%
XGBoost	95,50%	93,75%	97,63%	95,65%
BERTimbau	**97,29%**	**95,88%**	**98,79%**	**97,31%**

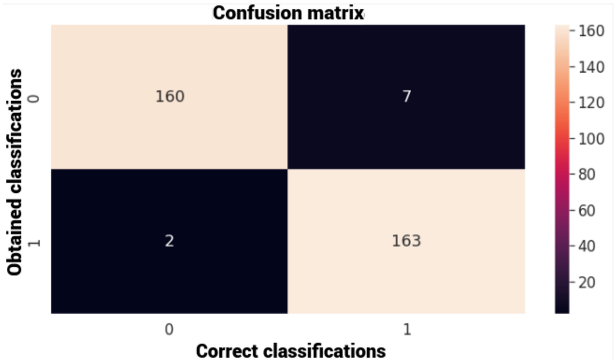


Figure 3. BERTimbau confusion matrix.

The confusion matrix for the BERTimbau model is presented in Figure 3.

To verify the absence of overfitting, Figure 4 plots the error behavior during training with training and validation data. Hence, we can observe that the validation data error decreases as the training one also decreases. Therefore, we can conclude that the model generalizes well.

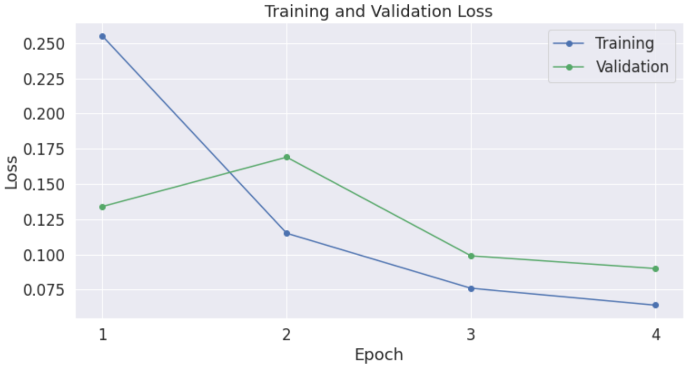


Figure 4. Overfitting test.

C. Deployment: Road Accidents System

The infrastructure of the accident alert system was developed with the help of the Python 3.10.2 programming language and the Postgresql 9.6 database for data management. In addition, the Spatial Postgis 3.2 extension was used.

The SafeTrip architecture is composed of three main components shown in Figure 5. The components are the crawler, the classifier and the mobile application.

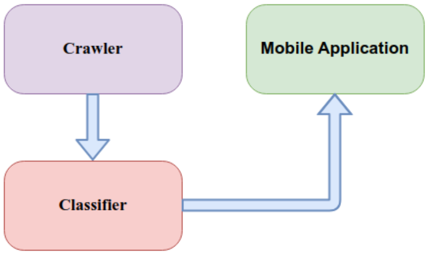


Figure 5. SafeTrip overall architecture.

1) Crawler: The crawler is responsible for tweets gathering.. With the help of the tweepy API, tweets were periodically collected, preprocessed and stored in the database, as shown in Figure 6.

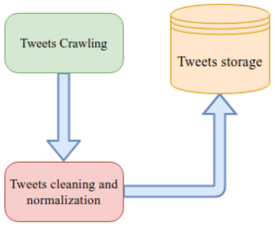


Figure 6. Crawler architecture.

2) Classifier: Figure 7 depicts the classifier module that is responsible for retrieving the tweet stored by the crawler and applying tokenization and data normalization to infer the tweet class according to the chosen classification model. When the tweets are classified as positive, a script based on regular expression is applied to extract information such as federal unit, code and kilometer of the highway to retrieve the accident geolocation. This geolocation consists of calculating the starting point of the kilometer where the accident occurred, so an algorithm was developed, described below and implemented in a stored procedure that, with the help of the database of Brazilian highways maintained by the Ministry of Infrastructure and the information extracted from the tweets, determines the latitude and longitude coordinates of the accident. At the end, this information, plus the time of occurrence and the text itself, are stored in the geographic database in the form of an alert.

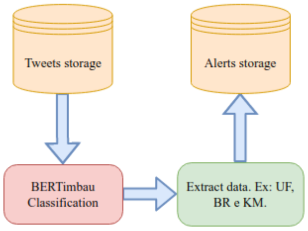


Figure 7. Classifier architecture.

3) The SafeTrip Mobile Application: In order to validate the results obtained and make them accessible to the community, a mobile application was developed to notify drivers by audio about accidents detected on Twitter. This mobile application was developed based on the client-server architecture model. According to Vaskevitch [37], in this architecture, processes are separated into independent platforms, allowing communication between processes while obtaining the maximum benefit from each different device. The server is responsible for monitoring PRF profiles, collecting published tweets, processing, classifying, extracting information, geolocating and sending alerts to users via websocket.

With the help of React Native, which is a javascript-based library that builds native code for Android and IOS applications, a mobile application was developed to act as a client, providing geolocation data and issuing alerts. This mobile application acts as a client in our architecture. Permissions are required to access the internet and also the location system of the device on which it is installed. In the mobile application home, users can choose one of all Brazilian cities as the destination city. The path between the user's current location and its destination will be displayed after the city is selected and the "Continue" button is pressed.

The client keeps the server informed about its location and when it detects an accident alert on the user's route, the server notifies the application that represents this information through an icon in the form of an exclamation mark, as shown in Figure 8. In addition, the textual content of the tweet that generated the alert is converted into audio.

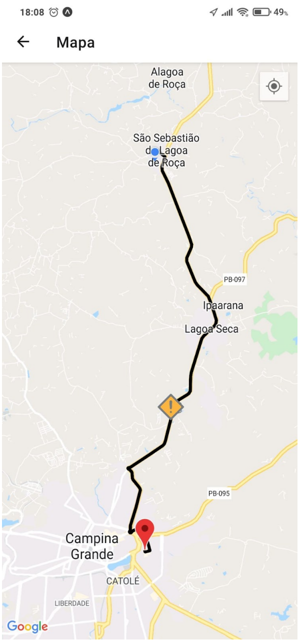


Figure 8. User´s route and accident allert.

V. Conclusion and Future Work

This article addressed the issue of road accidents and how, using technologies based on machine learning, natural language processing, geoprocessing and mobile applications, we can mitigate such accidents and improve drivers' travel safety through sound alerts of accidents that occurred on the respective routes. For that, tweets from the Brazilian Federal Highway Police were used, which report, among other things, the occurrence of road accidents, with a very high reliability, which differs from the information reported in traffic monitoring applications that are based on crowdsourcing.

One of the main challenges in carrying out this research was the manual labeling of the dataset for training the supervised classification models. Another major challenge was trying to understand the function of each hyperparameter in the classification models and choosing those that presented the best results. Finally, in the area of geoprocessing, we had the challenge of transforming an accident reported in the form highway kilometer to latitude and longitude coordinates. The BERT transformer obtained the best results in our experiments.

As future work, we intend to increase the corpus of accidents, using data from other secure sources such as tweets from highway concessionaires and the press, as well as carrying out usability tests of the mobile application. In addition, with reliable tweet classification models, it is possible to contribute as one more source for real-time updating of the Brazilian Federal Highway Police database referring to the history of accidents, in its open data portal, since there is a delay of about 2 months in the data update.

Acknowledgment

The second author would like to thank the National Council for Scientific and Technological Development (CNPq), Brazil, for partially funding this research.

References

[1] C. D. Mathers, T. Boerma, D. M. Fat, Global and regional causes of death, British Medical Bulletin, Volume 92, edition 1, pp. 7–32, 2009, https://doi.org/10.1093/bmb/ldp028.

[2] Google, Sobre o Waze, Available at https://support.google.com/waze/answer/6071177?hl=pt-BR&ref_topic=9022747, Accessed in 2023-03-09.

[3] A. Kocatepe, M. B. Ulak, E. E. Ozguven, and M. W. Horner,"Who might be affected by crashes? Identifying areas susceptible to crash injury risk and their major contributing factors", Transportmetrica A: transport science 15, pp. 1278–1305, 2019.

[4] C. Liu and A. Sharma, "Using the multivariate spatio-temporal Bayesian model to analyze traffic crashes by severity", Analytic methods in accident research 17, pp. 14–31, 2018.

[5] P. Tiwari, H. Dao, and G. N. Nguyen, "Performance evaluation of lazy, decision tree classifier and multilayer perceptron on traffic accident analysis", Informatica 41, 2017.

[6] P. Tiwari, S. Kumar, and D. Kalitin, "Road-user specific analysis of traffic accident using data mining techniques", In International Conference on Computational Intelligence, Communications, and Business Analytics, Springer, pp. 398–410, 2017.

[7] S. R. Santos, C. A. D. Junior, and R. Smarzaro, "Analyzing traffic accidents based on the integration of official and crowdsourced data", Journal of Information and Data Management 8, pp. 67–67, 2017.

[8] S. Kumar and D. Toshniwal, "A data mining framework to analyze road accident data", Journal of Big Data 2, pp. 1–18, 2015.

[9] S. Kumar and D. Toshniwal, "A data mining approach to characterize road accident locations", Journal of Modern Transportation 24, pp. 62–72, 2016.

[10] F. M. N. Ali and A. A. M. Hamed, "Usage Apriori and clustering algorithms in WEKA tools to mining dataset of traffic accidents", Journal of Information and Telecommunication 2, pp. 231–245, 2018.

[11] E. Turunen, "Using GUHA data mining method in analyzing road traffic accidents occurred in the years 2004–2008 in Finland", Data Science and Engineering 2, pp. 224–231, 2017.

[12] J. Abellán, G. López, and J. D. OñA, "Analysis of traffic accident severity using decision rules via decision trees", Expert Systems with Applications 40, pp. 6047–6054, 2015.

[13] H. İ. Bülbül, T. Kaya, and Y. Tulgar, "Analysis for status of the road accident occurance and determination of the risk of accident by machine learning in istanbul", In 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, pp. 426–430, 2016.

[14] S. Gao, R. Pan, R. Yu, and X. Wang, "Research on automated modeling algorithm using association rules for traffic accidents", In 2018 IEEE International Conference on Big Data and Smart Computing (BigComp), IEEE, pp. 127–132, 2018.

[15] Y. Guo, Z. Li, P. Liu, and Y. Wu, "Exploring risk factors with crashes by collision type at freeway diverge areas: Accounting for unobserved heterogeneity", IEEE Access 7 (2019), pp. 11809–11819, 2019.

[16] O. H. Kwon, W. Rhee, and Y. Yoon, "Application of classification algorithms for analysis of road safety risk factor dependencies". Accident Analysis & Prevention 75 (2015), pp. 1–15, 2015.

[17] Z. Li, X. Guo, and J. Sun. "Analysis and research on the temporal and spatial correlation of traffic accidents and illegal activities", In International Conference on Cloud Computing and Security. Springer, pp. 418–428.

[18] J. Wang and Y. Ohsawa, "Evaluating model of traffic accident rate on urban data". In 2016 Federated Conference on Computer Science and Information Systems (FedCSIS). IEEE, pp. 181–186, 2016.

[19] R. Bhavsar, A. Amin, and L. Zala, "Development of model for road crashes and identification of accident spots", International journal of intelligent transportation systems research 19, pp. 99–111, 2021.

[20] A. Iranitalab and A. Khattak, "Comparison of four statistical and machine learning methods for crash severity prediction". Accident Analysis & Prevention 108, pp. 27–36, 2017.

[21] S. Kumar, P. Tiwari, and K. V. Denis, "Augmenting classifiers performance through clustering: A comparative study on road accident data", International Journal of Information Retrieval Research (IJIRR) 8, pp. 57–68, 2018.

[22] N. Mor, H. Sood, and T. Goyal, "Application of machine learning technique for prediction of road accidents in Haryana-A novel approach", Journal of Intelligent & Fuzzy Systems 38, pp. 6627–6636, 2020.

[23] R. Richard and S. Ray, "A tale of two cities: Analyzing road accidents with big spatial data", In 2017 IEEE International Conference on Big Data (Big Data), IEEE, pp. 3461–3470, 2017.

[24] M. Sangare, S. Gupta, S. Bouzefrane, S. Banerjee, and P. Muhlethaler, Exploring the forecasting approach for road accidents: Analytical measures with hybrid machine learning, Expert Systems with Applications, pp. 113855, 2021.

[25] M. S. Satu, S. Ahamed, F. Hossain, T. Akter, and D. M. Farid, "Mining traffic accident data of N5 national highway in Bangladesh employing decision trees", In 2017 IEEE Region 10 Humanitarian Technology Conference (R10-HTC), IEEE, pp. 722–725, 2017.

[26] D. Singh and C. K. Mohan, "Deep spatio-temporal representation for detection of road accidents using stacked autoencoder", IEEE Transactions on Intelligent Transportation Systems 20, pp. 879–887, 2018.

[27] T. Tambouratzis, D. Souliou, M. Chalikias, and A. Gregoriades, "Combining probabilistic neural networks and decision trees for maximally accurate and efficient accident prediction", In The 2010 International Joint Conference on Neural Networks (IJCNN), IEEE, pp. 1–8, 2010.

[28] A. Katsoukis, L. Iliadis, A. Konguetsof, and B. Papadopoulos, "Classification Of Road Accidents Using Fuzzy Techniques", In 2018 Innovations in Intelligent Systems and Applications (INISTA), IEEE, pp. 1–5, 2018.

[29] B. Ryder, B. Gahr, P. Egolf, A. Dahlinger, and F. Wortmann, Preventing traffic accidents with in-vehicle decision support systems-The impact of accident hotspot warnings on driver behaviour, Decision support systems 99, pp. 64–74, 2017.

[30] H. Ren, Y. Song, J. Wang, Y. Hu, and J. Lei, "A deep learning approach to the citywide traffic accident risk prediction", 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, pp. 3346–3351.

[31] C. Zhang, Y. Shu, and L. Yan. "A Novel Identification Model for Road Traffic Accident Black Spots: A Case Study in Ningbo", China, IEEE Access 7, pp. 140197–140205, 2019.

[32] L. Yu, B. Du, X. Hu, L. Sun, W. Lv, and R. Huang, “Traffic Accident Prediction Based on Deep Spatio-Temporal Analysis”, IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI). IEEE, pp. 995–1002, 2019.

[33] P. Chapman, et al. Crisp-dm 1.0 Step-by-step data mining guide, 2000.

[34] D. N. Politis, J. P. Romano, M. Wolf, Subsampling, Springer, 1999.

[35] S. P. Vasconcelos, “Source code and corpus”, Available at: https://colab.research.google.com/drive/1bSWqT7AyHK5ZaprG7m_kf0fPkhcypPN_, Accessed in: 2023-03-10.

[36] S. P. Vasconcelos, “BERTimbau source code and corpus”, Available at: https://colab.research.google.com/drive/1h7RziptpLvCBSaJpiiq3ZbVH2N7Q6n33, Accessed in 2023-03-10.

[37] D. Vaskevitch, Client/server strategies, IDG Books Worldwide, 1995.