TAMIKA: Jurnal Tugas Akhir Manajemen Informatika & Komputerisasi Akuntansi
Vol. 4 No. 2 (Desember 2024) – ***Edisi Khusus SEMNASTIK 2024***
https://doi.org/10.46880/tamika.Vol4No2(SEMNASTIK).pp200-202 e-ISSN: 2828-1276

APPLICATION OF DECISION TREE ALGORITHM METHOD TO ANALYZE TRAFFIC ACCIDENT PATTERNS

[1]Rusmin Saragih✉, [1]Marto Sihombing, [1]Anton Sihombing, [2]Rivalri Kristianto Hondro
[1]Sekolah Tinggi Manajemen Informatika dan Komputer (STMIK) Kaputama, Binjai, Indonesia
Universitas Budi Darma, Medan, Indonesia
Email: evitha12014@gmail.com

ABSTRACT

Traffic accidents are complex problems that involve many variables such as weather conditions, vehicle type, location, and driver behavior. With the development of data processing technology, it is possible to analyze accident data in more depth to find significant hidden patterns. The Decision Tree algorithm is applied to predict the likelihood of an accident occurring and identify the factors that contribute most to the accident. The data used consists of accident records collected from various sources, including official reports and traffic statistics. The Decision Tree algorithm was chosen due to its ability to handle both categorical and numerical data, as well as the ease of interpretation of the analysis results. The results of this study show that factors such as vehicle speed, time of occurrence, and road conditions have a significant influence on the probability of an accident occurring. The results of news extraction are analyzed by creating decision rules to determine the pattern of accidents that occur. This decision rule is in the form of a decision tree with a dataset that uses data with the highest fatalities with the imputation feature mode by concept as a method of handling missing values and toll roads as attributes, resulting in an f1-score value of 60.00% and an accuracy value of 70.40%..

Keyword: Application, Decision Tree, Analysis, Accident Pattern, Traffic.

INTRODUCTION

Traffic accidents are a common occurrence on roads around the world, resulting in injuries, deaths, and property damage. These accidents not only have a significant impact on the individuals involved but also on their families and the community as a whole. In addition to the emotional toll, traffic accidents also have a financial burden, costing billions of dollars in medical expenses and insurance claims each year. It is essential to understand the causes of these accidents in order to implement effective preventive measures and reduce their frequency (Reason, 2017; Lucian-lonel, 2019).

Analyzing traffic accident patterns can help identify common factors that contribute to crashes, such as speeding, distracted driving, and impaired driving. By understanding these causes, authorities can develop targeted interventions, such as public awareness campaigns, improved road infrastructure, and stricter enforcement of traffic laws. This proactive approach can ultimately save lives, prevent injuries, and reduce the economic impact of traffic accidents on society (Setyabudi, 2021). In conclusion, by studying traffic accident patterns, we can work towards creating safer roads for everyone.

Decision trees can also be used to analyze the data collected from traffic accidents and identify the most common factors contributing to crashes (Zacharis, 2018).. This information can then be used to prioritize interventions and resources in areas where they are most needed. By combining decision tree analysis with online news information extraction, authorities can stay up-to-date on emerging trends and issues related to traffic safety, allowing them to make more informed decisions and prevent accidents before they happen. Ultimately, the goal is to create a comprehensive approach to road safety that addresses the root causes of accidents and creates a safer environment for all road users.

RESEARCH METHODS

Data and Sources

The data used in this research is data obtained from scraping on the detik.com website which is then extracted to retrieve the necessary information. The news site was chosen because there are quite a lot of articles related to traffic accidents on toll roads. The news text articles taken are news related to traffic accidents that occurred on roads in North Sumatra Province.

Online news article data obtained from scraping results are 5403 news articles about traffic accidents on roads in North Sumatra. Furthermore, the data was preprocessed in the form of data cleaning, text preprocessing, and filtering. The filtered data is 564 news articles and then information extraction is performed on the data.

Information extraction is done by utilizing one of the deep learning methods, namely Named Entity Recognition (NER) with Bidirectional Long Short Term Memory (Bi-LSTM) combined with Convolutional Neural Network (CNN) (Esther et al., 2021).

Data Preprocessing

Data cleaning is done to prepare the data by removing news with missing content, homogenizing the format, and removing duplicate data based on the similarity of news links or URLs. Then text preprocessing is done in the form of case folding, tokenizing, stopword removal, and stemming.

Filtering is done with the aim of reducing data that is not relevant to the topic and selecting similar data. In checking the relevance of the news, initially some data is taken and categorized into two categories, “relevant” and “irrelevant”, news that falls into the “relevant” category is labeled with the number “1” and “irrelevant” news is labeled with the number “0”. Then to automatically categorize the data, we can use Naïve Bayes Classifier Semi-Supervised Learning. Meanwhile, to find out similar content, the cosine similarity method can be used. Cosine similarity is done by vectorizing the content with TF-IDF, then measuring the similarity of each content with cosine value and storing it into a matrix. Similar news pairs are grouped together then deleted and left with the news with the latest upload date.

Information Extraction

Information extraction is performed using NER Bi-LSTM-CNN with a rule-based approach to obtain the required information in detail. NER is used to identify and extract accident information by categorizing entities into predefined classes. Then the rule-based system is used to retrieve the required attribute information. The entities extracted at this stage include CAUSE (cause of accident), DATE (date), DAY (day), LOC (location, area/place other than toll road), ORG (organization), PER (person/human), SINGLE (accident type), TIME (time), TOL (toll road), VEHICLE (vehicle), and MISC (others), each of which is given an IOB notation (Inside, Outside, Beginning) to mark the order.

RESULTS AND DISCUSSION

In the construction of the NER model, POS Tagging was first performed on the data with the help of the Flair library in Python programming. Then 30% of the data or 169 articles were manually annotated. This sample is used as a reference for training and evaluating the NER model. Annotations are given using the IOB (Inside, Outside, Beginning) approach, which indicates whether each token in the text is the beginning of the named entity, the inside of the named entity, or not part of the named entity at all.

IOB annotation in NER model building The data is separated into arrays per sentence, where each sentence is treated as a separate processing unit. This provides an advantage in the evaluation and use of the NER model, as the model can focus on each sentence individually to recognize and label the named entities contained therein.

The NER BiLSTM-CNN model produces an f1-score value of 81%, a recall value of 83%, and a precision value of 78%. Then from the classification results using NER BiLSTM-CNN, the extracted attributes are taken as information using a rule-based approach. The rules are used to identify and obtain certain attributes from the news text, such as highway, day, date, time, vehicle type, causal factor, accident type, number of injured victims, and number of dead victims.

From the results of information extraction, the similarity between the extracted data was checked again and 327 traffic accidents on the roads of North Sumatra Province were obtained which were considered single, with a summary of the data as follows:

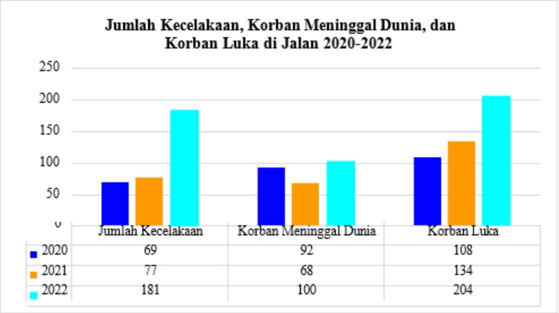

Figure 1. Result Traffic Accidents in North Sumatera 2020-2022

CONCLUSION

The results showed that the extraction of information on road traffic accidents in North Suamtera Province from detik.com news articles was carried out using the NER Bi-LSTM-CNN model which produced final data of 327 accident events. The results showed that CART can model the pattern of accidents quite well. The best CART model was found by using a dataset that only includes data on the three highways with the highest fatalities with the mode by concept imputation feature as a method of handling missing values. This model obtained an f1-score value of 60.00% and an accuracy of 70.40%.

DISEMINATION

This article has been disseminated at the National Seminar on Information and Communication Technology (SEMNASTIK) APTIKOM Year 2024 held by Universitas Methodist Indonesia on October 24-26, 2024.

REFERENCES

Esther, C., Eko, M., & Mauridhi. (2021). *Named entity recognition for extracting concept in ontology building on Indonesian language using end-to-end bidirectional long short term memory*. https://www.sciencedirect.com/science/article/pii/S0957417421002979

Lucian-Ionel. (2019). *Occupational accidents assessment by field of activity and investigation model for prevention and control.* https://www.mdpi.com/2313-576X/5/1/12

Reason, J. (2016). *Managing the risks of organizational accidents.* London: Routledge https://www.taylorfrancis.com/books/mono/10.4324/9781315543543/managing-risks-organizational-accidents-james-reason

Setyabudi, B. (2021). Kajian Peran Tempat Istirahat (Rest Area) Kendaraan Guna Menurunkan Tingkat Kecelakaan dan Kelelahan Pengemudi pada Jalan Tol Ruas Jakarta-Cikampek. *Warta Penelitian Perhubungan, 23*(4), 371-387.

Zacharis, N. Z. (2018). Classification and Regression Trees (CART) for Predictive Modeling in Blended Learning. *International Journal of Intelligent Systems and Applications, 10*(3), 1–9