2024 IEEE Conference on Artificial Intelligence (CAI)

Maritime-Context Text Identification for Connecting Artificial Intelligence (AI) Models

Xiaocai Zhang[1], Hur Lim[2], Xiuju Fu[1], Ke Wang[1*], Zhe Xiao[1], and Zheng Qin[1]

***Abstract*—This study focuses on identifying texts related to maritime contexts using an advanced Large Language Model (LLM) and cost-sensitive approach for handling data imbalances. Firstly, a comprehensive dataset specifically for maritime-context queries is collected and augmented. Secondly, the dynamic contextual representations of input query considering the context of each word are obtained by a pre-trained LLM which incorporates Bidirectional Encoder Representations from Transformers (BERT) and Convolutional Neural Network (CNN). Thirdly, a Multi-Layer Perceptron (MLP) is constructed as the classifier to fine-tune the whole network on the newly collected dataset. Finally, the Focal loss is introduced for more effective parameter optimization to tackle the challenge of data imbalance between positive and negative samples, Extensive experiments have been conducted and the following promising results have been obtained: 1) The proposed approach achieves an impressive 99.97% F1 score in recognizing maritime-context texts; 2) The ConvBERT model, an enhancement over the original BERT, demonstrates superior performance in text representation while being more computationally efficient; 3) The Focal loss method outperforms other cost-sensitive learning strategies like class weighting and oversampling techniques; and 4) the proposed method surpasses other deep learning and BERT-based methods in text classification tasks.**

***Index Terms*—Maritime, Large Language Model (LLM), ConvBERT, text classification, imbalanced**

I. Introduction

In recent years, the advancements in Artificial Intelligence (AI) within the maritime sector are indeed significant and transformative. The development of AI-driven tools like Estimated Time of Arrival (ETA) estimation [1], fuel consumption prediction [2], traffic hotspots forecasting [3], vessel trajectories prediction [4], [5], maritime risk assessment [6], and vessel recognition/identification [7], [8] shows a substantial shift in how maritime operations are managed and optimized. These AI practices not only enhance efficiency but also contribute to safer and more sustainable maritime practices. The notion of developing a dedicated AI model store or repository for the maritime sector is both necessary and beneficial. Such a repository would serve as a centralized platform where practitioners, researchers, and developers can access, share, and collaborate on AI models tailored to maritime needs. This would accelerate innovation, improve the quality of maritime AI applications, and potentially lead to standardized practices across the industry. The involvement of the Institute of High Performance Computing (IHPC) at the Agency for Science, Technology and Research (A*STAR), Singapore in developing this concept is promising. Their expertise and resources could be pivotal in realizing an effective and robust maritime AI model store/repository. This initiative could pave the way for broader use and continuous improvement of AI models in maritime transportation, ensuring that the industry remains at the forefront of technological advancement. As this concept evolves, it will be interesting to see how it shapes the future of maritime operations and how it might influence similar developments in other industry sectors. The focus on collaboration and shared development in the field of AI reflects a growing trend towards open innovation and could lead to significant breakthroughs in the application of AI technologies in maritime.

To address the need for improved connections between industry partners and AI models within the maritime ecosystem, we are inspired to develop a new Question and Answer (Q&A) system specific to this sector. A primary challenge in accomplishing this system is to accurately identify whether incoming queries or texts pertain to maritime scenarios. Relevant maritime texts trigger further processing, while non-maritime queries are disregarded as they fall outside the system's scope. In this work, we have proposed an imbalanced learning approach to detect maritime-context texts. The following are the key contributions to highlight:

- We established a new human-annotated text/query dataset within the maritime field, designed to support and advance future research endeavors;
- We implemented a cutting-edge contextual representation method for text classification in the maritime domain;
- We developed a deep learning approach tailored for imbalanced learning in maritime text classification;
- We conducted evaluations using the newly established dataset to demonstrate the effectiveness and superiority of our approach.

The subsequent parts of this article are organized as follows. Section II presents a comprehensive review of previous studies on text classification in transportation. In Section III, we explain the proposed method in detail. Section IV details the results of the experiments and analyses. Finally, in Section V, we conclude this study.

*Corresponding author: Ke Wang.

[1]Xiaocai Zhang, Xiuju Fu, Ke Wang, Zhe Xiao and Zheng Qin are with the Institute of High Performance Computing (IHPC), Agency for Science, Technology and Research (A*STAR), 1 Fusionopolis Way, #16-16 Connexis, Singapore 138632, Republic of Singapore `zhang_xiaocai@ihpc.a-star.edu.sg`; `wang_ke@ihpc.a-star.edu.sg`

[2]Hur Lim is with the School of Computing, Singapore Polytechnic, 500 Dover Road, Singapore 139651, Republic of Singapore

979-8-3503-5409-6/24/$31.00 ©2024 IEEE
DOI 10.1109/CAI59869.2024.00165

II. RELATED WORK

The text classification methods in transportation can be generally divided into five categories: rule-based methods, machine learning methods, neural network methods, large language models (LLMs), and hybrid methods. The machine learning methods in this section do not involve the concepts of neural networks and LLMs. Rule-based methods apply a collection of specific rules or patterns to analyze, interpret, or manipulate natural language data. In [9], a specialized dictionary was created, featuring unique keywords, to categorize tweets into seven distinct travel-related activities, including eating, entertainment, home activities, shopping, and more. For the machine learning methods, the most frequently used models in transportation-related text classification are the Support Vector Machine (SVM) and Naive Bayes (NB). In [10], the SVM was utilized to filter out texts not related to transportation. Specifically, the SVM was trained to categorize texts into two groups: positive and negative, with the positive group representing texts relevant to transportation. Styawati *et al.* [11] transformed the comment texts from Gojek and Grab into vector representations using Word2Vec, a method that clusters similar contexts in the corpus closely in the vector space [12]. Following that, they applied SVM for sentiment classification based on the text representations. Klaithin *et al.* [13] used the NB classifier to classify the event from Twitter comments into six categories, including accident, announcement, question, orientation, request and sentiment. In a different study by Mogaji *et al.*, the authors used TextBlob, a well-known Python library for machine learning, for sentiment analysis. This tool assesses the sentiment of input texts by assigning them polarity scores between -1.0 and 1.0, and subjectivity scores ranging from 0.0 to 1.0.

In this field, cutting-edge neural network methods are widely employed, with most studies favoring architectures based on Recurrent Neural Networks (RNN) and Convolutional Neural Networks (CNN). Chen *et al.* [14] applied the Continuous Bag of Words (CBOW) approach for word embeddings and developed a deep learning model that combines Long Short-Term Memory (LSTM) and CNN. This model was designed for binary text classification, effectively differentiating between texts relevant and irrelevant to traffic. In a similar work [15], the authors utilized CBOW for word embedding and a CNN model to identify traffic-related microblogs. Their approach demonstrated superior performance compared to methods based on SVM and Multi-Layer Perceptrons (MLP). Azhar *et al.* [16] utilized Global Vectors for Word Representation (GloVe) to embed words from tweets. Subsequently, they implemented RNN, Gated Recurrent Units (GRU), and LSTM networks for detecting traffic-related tweets and classifying the severity of traffic situations. The task of identifying traffic-related tweets was approached as a binary classification problem. In this context, the LSTM network achieved superior performance compared to GRU and RNN, recording an accuracy of 94.2%, against 93.7% for GRU and 91.6% for RNN. Xu *et al.* [17] combined Word2Vec with a CNN to categorize fault types in aircraft maintenance logs. Their study showed that the CNN classifier outperformed both MLP and SVM classifiers in terms of effectiveness. Su *et al.* [18] developed a model based on the GRU for classifying sentiments in comments from Chinese railway users, integrating a multi-feature fusion approach.

In transportation research and applications, LLMs like Bidirectional Encoder Representations from Transformers (BERT) and its variants are highly prevalent. Wan *et al.* [19] utilized BERT for categorizing tweets into traffic-related and unrelated groups. BERT has also proven effective in multi-class classifications, such as sorting tweets into six categories, including incident, road construction, road closure, traffic delay, public transportation and unrelated information, achieving an impressive accuracy of 99.37%, significantly outperforming traditional machine learning methods like NB, Decision Tree (DT), and SVM. Osorio *et al.* [20] applied BERT to analyze sentiments in tweets about Madrid Metro, classifying them into positive and negative emotions. Oliaee *et al.* [21] also employed BERT to classify traffic injury severity in crash reports. In another study by Babbar *et al.* [22], RoBERTa was used for sentiment classification, outperforming existing techniques like Word2Vec, GloVe, FastText, BERT, and XLNET, with 97% accuracy, 96% recall, and 95% F1-score.

The hybrid methods incorporate two or more categories of methods for modelling by leveraging the advantages of different methods. Jidkov *et al.* [23] introduced the BERT to obtain the vector representation of texts, and architectures of Artificial Neural Network (ANN), CNN and LSTM are then appended as the classifier for binary incident classification, i.e., if the text contains maritime incident or not. In [24], BERT was used for feature representation learning from tweets, followed by the integration of a CNN for classifying traffic events from these tweets. The study found that BERT, as a contextual word embedding tool, outperformed other models like ELMo and Word2vec. The results showed that the best-performing neural network architecture is LSTM, with an accuracy of 94.4%. In the research work by Khodadadi *et al.* [25], TF-IDF, POS tagger and n-grames were employed for feature extraction from customer service report data. Then, the Chi-Squared method was utilized for dimension reduction. Furthermore, the LSTM-CNN model was developed for customer service claim validation, which classifies the claim into three types: valid, fake or vague. If the claim is predicted as valid, a machine learning method, namely Gradient Tree Boosting, is used to assign the corresponding request to sixteen different service departments.

III. PROPOSED METHOD

Fig. 1 outlines the structure of the maritime-context text identification framework and its integration into maritime AI models. Our approach to detecting maritime-context text is based on ConvBERT, an enhanced variant of BERT. This method utilizes ConvBERT to generate contextual text representations and subsequently employs an MLP for refining

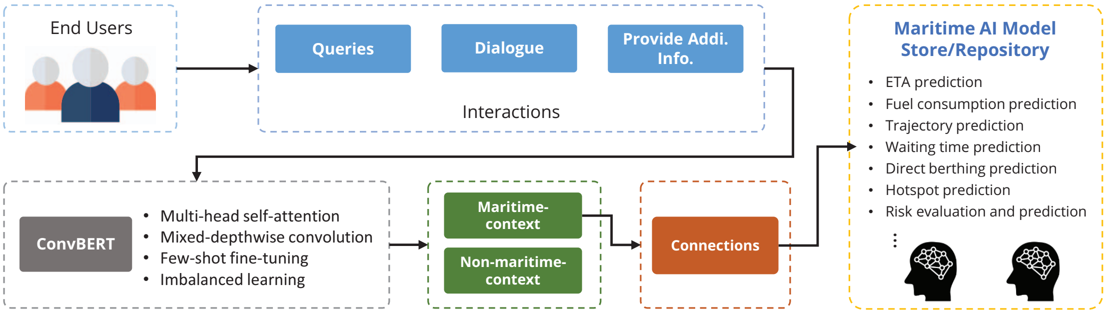


Fig. 1. The framework of maritime-context text identification and its application in connecting maritime AI models.

ConvBERT through fine-tuning with a focal loss. After identifying inputs with maritime context from end users, further steps are taken to establish connections with the designated AI models in the maritime AI model store/repository.

ConvBERT [26] can effectively capture both local and global contextual through the incorporation of convolutional layers. It employs the fundamental Transformer unit similar to BERT. This Transformer unit adheres to the standard architecture initially introduced by Vaswani *et al.* [27]. Each layer's self-attention mechanism can be characterized as

$$\text{Attention}\left(Q,K,V\right)=\text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V, \tag{1}$$

where Q, K, and V are the query, key, and value matrices respectively, and d_k is the dimension of the key vectors.

ConvBERT introduces a mixed-depthwise convolutional layer. The layer applies a convolution operation to the sequences after the multi-head attention mechanism. The convolutional layer can be formulated as

$$C\left(x\right)=\text{ReLU}\left(W*x+b\right), \tag{2}$$

where x is the input, W is the convolutional filter, b is the bias, and $*$ denotes the convolution operation.

In ConvBERT, the multi-head self-attention is modified to incorporate the grouped linear transformations, reducing the computational cost. The modified self-attention can be represented as

$$\text{G-Attention}\left(Q,K,V\right)=\text{Concat}\left(\text{head}_1,\cdots,\text{head}_g\right)W^O, \tag{3}$$

and

$$\text{head}_i=\text{Attention}\left(QW_i^Q,KW_i^K,VW_i^V\right). \tag{4}$$

Like BERT, ConvBERT is pre-trained using a Masked Language Model (MLM) objective, which can be formulated as

$$L_{\text{MLM}}=-\sum_{i\in M}\log P\left(w_i|w_{-i};\Theta_{ConvBERT}\right), \tag{5}$$

where M is the set of masked indices, w_i is the masked word, w_{-i} is the context, and $\Theta_{ConvBERT}$ represents the model parameters.

Given a sequence of tokens represented as a series of tuples, denoted as

$$X=\left\{x_t \mid t=1,2,\cdots,n\right\}, \tag{6}$$

this sequence X is input into the pre-trained ConvBERT model to obtain its vector representation, which is formulated as

$$y_{rep}=f_{ConvBERT}\left(X;\Theta_{ConvBERT}\right), \tag{7}$$

where $f_{ConvBERT}$ signifies the ConvBERT model that has been pre-trained. Following that, y_{pre} is connected to an MLP block, as represented by

$$y_{MLP}=f_{MLP}\left(y_{pre};\Theta_{MLP}\right), \tag{8}$$

where f_{MLP} signifies the MLP classifier constructed, and Θ_{MLP} stands the parameters of MLP. A Sigmoid activate function ($f_{Sigmoid}$) is then appended to predict the probability, as shown by

$$y_{oup}=f_{Sigmoid}\left(y_{MLP}\right). \tag{9}$$

The Focal loss is employed as the loss function for optimization, which is formulated as

$$\mathcal{L}=-\alpha\left(1-y_{oup}\right)^{\gamma} log\left(y_{oup}\right), \tag{10}$$

where α is a weighting factor and γ is the focusing parameter. The objective of network training is to learn the parameters of ConvBERT and MLP over the training data by minimizing the loss function, as indicated by

$$\left(\Theta_{ConvBERT},\Theta_{MLP}\right)=\underset{\Theta_{ConvBERT},\Theta_{MLP}}{\arg\min}\ \mathcal{L}. \tag{11}$$

In a similar way to the basic deep neural network training [28], the optimal parameters $\Theta_{ConvBERT}$ and Θ_{MLP} are obtained by a widely-used back-propagation algorithm with an Adam optimizer.

IV. Experiments and Evaluation

A. Dataset

The dataset in this study is sourced from two segments: a publicly available open-access dataset and data we gathered independently. For positive instances (maritime-related queries), we initially compiled 525 queries, which expanded to 38,885 samples after applying data augmentation techniques like synonym replacement, Round-Trip Translation (RTT), random swap, and introducing spell/keyboard errors. The negative instances (non-maritime-related queries) were derived from Google's Natural Questions dataset[1], encompassing 50,550 queries from various domains. The dataset exhibits a slight imbalance with a ratio of 1.3. Examples of queries from both positive (maritime-related) and negative (non-maritime-related) categories are illustrated in Fig. 2 for clarity.

- when the ves8e1 BUXLINK can reach the port of H0NC KONG?
- forecast the eta the next port of the ship with mmsi 564765123
- Check the estimated arrival tiern to the pilote boarding ground A. tho nsme of the vessle has D. K. I, are her MMSI ie 480128011.
- direct berthing prediction for ship who has come at the portholе of hamburg
- how long do the vas wait at the anchorage in porthole of Antwerp
- Oil consum estimation for wessel VIRTSU
- when can the pilot on board the trial on the terminal?

(a) Positive samples (maritime-related)

- who did the voice of the magician in frosty the snowman
- who said mangal bhawan amangal haari dravahu so dasrath ajir bihari to whom
- in order to be a new state a territory had to have a population of at least
- american competitiveness and workforce improvement act of 1998 (acwia) fee
- how many pages does all the harry potter books have
- when does the movie first reformed come out

(b) Negative samples (non-maritime-related)

Fig. 2. Examples of positive and negative samples.

B. Experimental Settings

We developed our model using Python and TensorFlow, and executed all tests on a system outfitted with an Intel(R) Xeon(R) Gold 6248 CPU, which runs at 2.50GHz, complemented by an NVIDIA Tesla V100 GPU with 32GB of memory. In terms of dataset division, we adhered to a training-validation-test split of 70%, 20%, and 10%, respectively.

C. Evaluation Metrics

To assess our model's effectiveness, we utilize various metrics including accuracy, precision, recall, F1-score, Matthews Correlation Coefficient (MCC), and the Area Under the Receiver Operating Characteristic (ROC) Curve (AUC). The ROC curve is derived from the True Positive Rate (TPR) and False Positive Rate (FPR), which are defined as

$$TPR = \frac{TP}{TP + FN}, \tag{12}$$

and

$$FPR = \frac{FP}{FP + TN}. \tag{13}$$

Additionally, the other five evaluation metrics are defined as:

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{14}$$

$$precision = \frac{TP}{TP + FP}, \tag{15}$$

$$recall = \frac{TP}{TP + FN}, \tag{16}$$

$$F1\text{-}score = 2 \times \frac{precision \times recall}{precision + recall}, \tag{17}$$

and

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{18}$$

where TP, FP, FN, and TN represent true positive, false positive, false negative, and true negative, respectively.

D. Results and Analyses

The performance of our proposed model on maritime-context text identification has been presented in Table I. Our method obtained a high and reliable performance on the test data. The F1-score, accuracy, precision, recall, MCC and AUC achieve high rates at 99.97%, 99.97%, 99.95%, 99.99%, 99.94% and 99.99%, respectively. The results are promising and encouraging. The standard BERT can achieve a very competitive performance to the ConvBERT we utilized. However, ConvBERT requires fewer computational resources than BERT, as demonstrated in Figure 3, making it an advantage. ConvBERT reduces 52.93% (i.e., 630 s v.s. 1339 s) the training time, as it is a lightweight neural network with much less parameters (i.e., 1.75×10^7 v.s. 10.95×10^7) than BERT. Furthermore, five deep learning models, including Encode-Decoder LSTM, Bidirectional Gated Recurrent Unit (BiGRU)-CNN, Attention Bidirectional LSTM (BiLSTM), and Temporal Convolutional Network (TCN) are constructed to compare with ConvBERT. For all of the deep learning methods, the input texts are first represented by a spare word embedding technique, namely the Global Vectors for Word Representation (GloVe), and then feed into the deep learning model for training and inference. Table I gives the comparison regarding the performance. The deep learning model with the spare word representation technique cannot outperform the BERT and ConvBERT, as BERT and ConvBERT are pre-trained on large corpora of text data which is better for Contextual Understanding. TCN performs the best among the deep learning models in the task of maritime-context text identification, with a high F1-score of 0.9705. ConvBERT surpasses it by at least 2.92% in terms of F1-score. Encode-Decoder LSTM and BiGRU-CNN perform the worst, obtaining inferior F1-scores of 0.4091 and 0.7234, respectively.

The ConvBERT model incorporates the Focal loss as an imbalanced learning strategy, and its performance is compared with other popular imbalanced learning approaches. These include oversampling techniques such as the Synthetic Minority Over-Sampling Technique (SMOTE) and the Adaptive Synthetic Sampling Approach (ADASYN), as well as cost-sensitive learning methods like class weighting. The comparative results are detailed in Table II, revealing that the Focal loss yields the best performance. In general, the cost-sensitive learning methods demonstrate superior performance compared to the oversampling techniques. Notably, class weighting outperforms SMOTE [29] and ADASYN [30] by a margin of

[1]https://ai.google.com/research/NaturalQuestions

TABLE I
PERFORMANCE COMPARISON BETWEEN OUR METHOD AND THE BASELINE METHODS

Model	F1-score	Accuracy	Precision	Recall	MCC	AUC
Encoder-Decoder LSTM	0.4091	0.6649	0.8768	0.2668	0.3489	0.8570
BiGRU-CNN	0.7234	0.8064	0.9545	0.5824	0.6299	0.9270
Attention BiLSTM	0.9229	0.9344	0.9441	0.9026	0.8661	0.9816
BiLSTM-CNN	0.9269	0.9392	0.9836	0.8747	0.8792	0.9897
TCN	0.9705	0.9749	0.9917	0.9501	0.9493	0.9974
BERT	0.9996	**0.9997**	**0.9995**	0.9997	0.9993	**0.9999**
ConvBERT	**0.9997**	**0.9997**	**0.9995**	**0.9999**	**0.9994**	**0.9999**

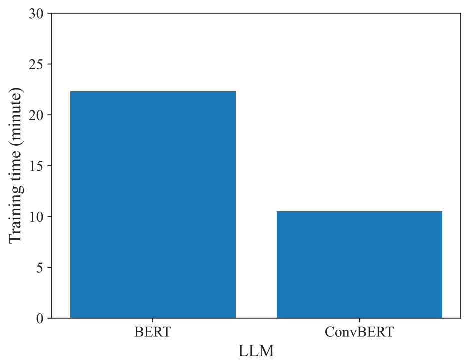
TABLE II
PERFORMANCE COMPARISON BETWEEN DIFFERENT LEARNING STRATEGIES

Model	F1-score	Accuracy	Precision	Recall	MCC	AUC
ConvBERT + SMOTE	0.9673	0.9715	0.9939	0.9949	0.9901	0.9998
ConvBERT + ADASYN	0.9660	0.9700	0.9509	0.9997	0.9816	0.9968
ConvBERT + Class Weighting	0.9988	0.9990	0.9981	0.9996	0.9980	0.9995
ConvBERT + Focal Loss	**0.9997**	**0.9997**	**0.9995**	**0.9999**	**0.9994**	**0.9999**

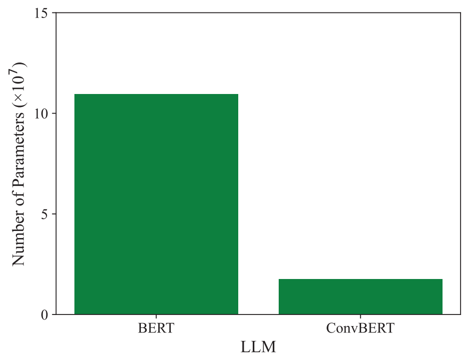
at least 3.15% (0.9988 vs. 0.9673) in terms of F1-score. Furthermore, the Focal loss surpasses them by at least 3.24% (0.9997 vs. 0.9673) in F1-score.

V. CONCLUSIONS

This paper introduces an LLM-based approach designed for detecting maritime-context texts. Initially, a self-collected dataset specific to maritime content is constructed, serving both as an experimental resource and a facilitator for future NLP research in the maritime transportation domain. Following this, an advanced LLM named ConvBERT is employed to capture the representation of the text data. To classify the text into either maritime-context or non-maritime-context, a classifier based on a multi-layer perceptron is constructed. Addressing the imbalanced nature of the data, the study utilizes the Focal loss. Experimental results indicate that the proposed approach is highly effective, achieving an impressive F1-score and accuracy of 99.97%. Notably, ConvBERT outperforms BERT, demonstrating slightly better performance while requiring less computational cost, reducing computational time by 52.93%. Furthermore, the proposed method surpasses other state-of-the-art deep learning models with a spare word embedding technique. These promising results suggest potential applications in the Q&A system, exhibiting the prospects of integrating comprehensive AI models in the maritime domain.



(a)



(b)

Fig. 3. Comparison between BERT and ConvBERT. (a) Training time comparison; (b) Number of parameters comparison.

REFERENCES

[1] X. Zhang, X. Fu, Z. Xiao, H. Xu, X. Wei, J. Koh, D. Ogawa, and Z. Qin, “Prediction of vessel arrival time to pilotage area using multi-data fusion and deep learning,” *arXiv preprint arXiv:2403.09969*, 2024.

[2] Z. S. Chen, J. S. L. Lam, and Z. Xiao, “Prediction of harbour vessel fuel consumption based on machine learning approach,” *Ocean Engineering*, vol. 278, p. 114483, 2023.

[3] Z. Xiao, X. Fu, L. Zhao, L. Zhang, T. K. Teo, N. Li, W. Zhang, and Z. Qin, "Next-generation vessel traffic services systems—from "passive" to "proactive"," *IEEE Intelligent Transportation Systems Magazine*, vol. 15, no. 1, pp. 363–377, 2022.

[4] X. Zhang, X. Fu, Z. Xiao, H. Xu, and Z. Qin, "Vessel trajectory prediction in maritime transportation: Current approaches and beyond," *IEEE Transactions on Intelligent Transportation Systems*, 2022.

[5] T. Liu, X. Xu, Z. Lei, X. Zhang, M. Sha, and F. Wang, "A multi-task deep learning model integrating ship trajectory and collision risk prediction," *Ocean Engineering*, vol. 287, p. 115870, 2023.

[6] T. Liu, S. Wang, Z. Lei, J. Zhang, and X. Zhang, "Trajectory risk cognition of ship collision accident based on fusion of multi-model spatial data," *The Journal of Navigation*, vol. 75, no. 2, pp. 299–318, 2022.

[7] X. Zhang, Z. Xiao, X. Fu, X. Wei, T. Liu, R. Yan, Z. Qin, and J. Zhang, "A viewpoint adaptation ensemble contrastive learning framework for vessel type recognition with limited data," *Expert Systems with Applications*, vol. 238, p. 122191, 2024.

[8] X. Zhang, Y. Liu, Y. Zheng, Z. Zhao, J. Li, and Y. Liu, "Distinction between ships and icebergs in sar images using ensemble loss trained convolutional neural networks," in *The 31st Australasian Joint Conference on Artificial Intelligence*. Springer, 2018, pp. 216–223.

[9] M. Maghrebi, A. Abbasi, T. H. Rashidi, and S. T. Waller, "Complementing travel diary surveys with twitter data: application of text mining techniques on activity location, type and time," in *2015 IEEE 18th International Conference on Intelligent Transportation Systems (ITSC))*. IEEE, 2015, pp. 208–213.

[10] F. Ali, S. El-Sappagh, and D. Kwak, "Fuzzy ontology and lstm-based text mining: a transportation network monitoring system for assisting travel," *Sensors*, vol. 19, no. 2, p. 234, 2019.

[11] S. Styawati, A. Nurkholis, A. A. Aldino, S. Samsugi, E. Suryati, and R. P. Cahyono, "Sentiment analysis on online transportation reviews using word2vec text embedding model feature extraction and support vector machine (svm) algorithm," in *2021 International Seminar on Machine Learning, Optimization, and Data Science (ISMODE)*. IEEE, 2022, pp. 163–167.

[12] X. Zhang, Z. Zhao, Y. Zheng, and J. Li, "Prediction of taxi destinations using a novel data embedding method and ensemble learning," *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 1, pp. 68–78, 2020.

[13] S. Klaithin and C. Haruechaiyasak, "Traffic information extraction and classification from thai twitter," in *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*. IEEE, 2016, pp. 1–6.

[14] Y. Chen, Y. Lv, X. Wang, L. Li, and F.-Y. Wang, "Detecting traffic information from social media texts with deep learning approaches," *IEEE Transactions on Intelligent Transportation Systems*, vol. 20, no. 8, pp. 3049–3058, 2018.

[15] Y. Chen, Y. Lv, X. Wang, and F.-Y. Wang, "A convolutional neural network for traffic information sensing from social media text," in *2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2017, pp. 1–6.

[16] A. Azhar, S. Rubab, M. M. Khan, Y. A. Bangash, M. D. Alshehri, F. Illahi, and A. K. Bashir, "Detection and prediction of traffic accidents using deep learning techniques," *Cluster Computing*, vol. 26, no. 1, pp. 477–493, 2023.

[17] Z. Xu, B. Chen, S. Zhou, W. Chang, X. Ji, C. Wei, and W. Hou, "A text-driven aircraft fault diagnosis model based on a word2vec and priori-knowledge convolutional neural network," *Aerospace*, vol. 8, no. 4, p. 112, 2021.

[18] C. Su, L. Jiang, and S. Pi, "Sentiment classification of chinese railway review text based on multi-feature fusion gated recurrent unit," in *2021 International Conference on Information Science, Parallel and Distributed Systems (ISPDS)*. IEEE, 2021, pp. 197–202.

[19] X. Wan, M. C. Lucic, H. Ghazzai, and Y. Massoud, "Empowering real-time traffic reporting systems with nlp-processed social media data," *IEEE Open Journal of Intelligent Transportation Systems*, vol. 1, pp. 159–175, 2020.

[20] J. Osorio-Arjona, J. Horak, R. Svoboda, and Y. García-Ruíz, "Social media semantic perceptions on madrid metro system: Using twitter data to link complaints to space," *Sustainable Cities and Society*, vol. 64, p. 102530, 2021.

[21] A. H. Oliaee, S. Das, J. Liu, and M. A. Rahman, "Using bidirectional encoder representations from transformers (bert) to classify traffic crash severity types," *Natural Language Processing Journal*, vol. 3, p. 100007, 2023.

[22] S. Babbar and J. Bedi, "Real-time traffic, accident, and potholes detection by deep learning techniques: a modern approach for traffic management," *Neural Computing and Applications*, vol. 35, no. 26, pp. 19 465–19 479, 2023.

[23] V. Jidkov, R. Abielmona, A. Teske, and E. Petriu, "Enabling maritime risk assessment using natural language processing-based deep learning techniques," in *2020 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2020, pp. 2469–2476.

[24] G. A. Neruda and E. Winarko, "Traffic event detection from twitter using a combination of cnn and bert," in *2021 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*. IEEE, 2021, pp. 1–7.

[25] A. Khodadadi, S. Ghandiparsi, and C.-N. Chuah, "A natural language processing and deep learning based model for automated vehicle diagnostics using free-text customer service reports," *Machine Learning with Applications*, vol. 10, p. 100424, 2022.

[26] Z.-H. Jiang, W. Yu, D. Zhou, Y. Chen, J. Feng, and S. Yan, "Convbert: Improving bert with span-based dynamic convolution," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 837–12 848, 2020.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[28] X. Zhang, H. Peng, J. Zhang, and Y. Wang, "A cost-sensitive attention temporal convolutional network based on adaptive top-k differential evolution for imbalanced time-series classification," *Expert Systems with Applications*, vol. 213, p. 119073, 2023.

[29] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[30] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in *2008 IEEE International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2008, pp. 1322–1328.