Article

Analysis of Bi-LSTM CRF Series Models for Semantic Classification of NAVTEX Navigational Safety Messages

Changui Lee [1], Hoyeon Cho [2] and Seojeong Lee [1,*]

1   Division of Marine System Engineering, National Korea Maritime and Ocean University, Busan 49112,
     Republic of Korea; culee@kmou.ac.kr
2   Division of Maritime Information Technology, National Korea Maritime and Ocean University, Busan 49112,
     Republic of Korea; hoyeoncho@g.kmou.ac.kr
*   Correspondence: sjlee@kmou.ac.kr; Tel.: +82-51-410-4578

Abstract: NAVTEX is a key component in the Global Maritime Distress and Safety System (GMDSS) that automatically transmits urgent maritime safety information such as navigational and meteorological warnings and forecasts to vessels. For the safe navigation of smart ships, this information from different systems should be shared harmoniously in the Common Maritime Data Structure (CMDS). To share NAVTEX messages as CMDS, words in NAVTEX messages must be semantically classified and placed within the CMDS structure. While traditional parsing methods are typically used to understand message semantics, NAVTEX requires natural language processing methods with deep learning due to its unstructured messages. This paper applies six types of Bi-LSTM CRF-based deep learning models to NAVTEX navigational safety messages and analyzes the results to find the most suitable model for understanding the semantics of each word in NAVTEX messages. This technique can be applied to accurately convey the meaning of NAVTEX navigational safety messages to equipment that requires navigational safety information on smart ships without human intervention.

Keywords: NAVTEX; GMDSS; CMDS; S-100; machine readable; deep learning; smart ship; NLP

1. Introduction

For decades, the shipping industry has used various methods and communications technologies to protect ships and their crews. The NAVTEX system represents a core component of the Global Maritime Distress and Safety System, generally used for the automatic sending of maritime safety information to vessels, primarily related to navigational warnings and meteorological warnings and forecasts [1].

As NAVTEX messages are the key element of ship operation, it is traditionally interpreted by humans to reflect on navigational safety and make decisions. But in the environment of smart ship operations, human intervention is reduced, and machines take on this role instead. This means that instead of human interpretation, NAVTEX messages should be automatically exchanged with systems like the Electronic Chart Display and Information System (ECDIS) and autopilot to adjust routes according to navigational warnings or to assist human decision-making through display interfaces. For communication of safety information between smart devices without human intervention, it must be converted into a machine-readable format [2]. So, it means that NAVTEX should be integrated and shared through the Common Maritime Data Structure (CMDS), which is the standard provided by the International Maritime Organization (IMO) for the structural interoperability of data. A system that would correctly classify the meaning of each word and phrase of a NAVTEX message and properly place them should, therefore, be considered if the messages are to take the form of a CMDS.

An international standard exists for the form of NAVTEX messages, whereas the text itself, within a message body, is in unstructured form for expressing the different situations [3,4]. To understand the meaning of a structured message, parsing processes

have typically been used [5]. Since navigational safety messages are in an unstructured form, analysis with only the parsing method is quite hard. Thus, attempts to rightly understand the meanings of the messages using deep learning-based natural language processing techniques have risen. Especially, deep learning models like Bi-LSTM are claimed to be able to learn word representations in context and, therefore, are theoretically useful in processing NAVTEX messages [6].

In this paper, deep learning models based on Bi-LSTM CRF are applied to the classification of meanings for navigational safety messages in NAVTEX. We compared and analyzed different models based on Bi-LSTM CRF to find a suitable model for understanding the meaning of NAVTEX messages. This will enable more accurate and efficient interpretation of NAVTEX messages, facilitating their automated processing and conversion into the CMDS structure, thereby contributing to the safe navigation of smart ships.

2. Background

2.1. Navigational Telex (NAVTEX)

NAVTEX is a system that transmits navigational warnings, weather forecasts, search and rescue information, and other urgent maritime safety information to vessels. This system was introduced as a key component of the GMDSS by the 1974 International Convention for the Safety of Life at Sea (SOLAS Convention) [1]. NAVTEX operates on two main frequencies: one is the international NAVTEX service, which broadcasts on 518 kHz and is used globally; the other is the national NAVTEX service, which operates on 490 kHz and is broadcast in local languages by individual countries.

The NAVTEX system is designed to automatically deliver information to ships within specific geographical areas. It primarily uses text-based messages to provide critical information related to maritime safety, and NAVTEX receivers installed on vessels automatically receive and display these messages without requiring manual intervention.

NAVTEX messages have a standardized format and are composed of three parts to provide essential information for the safe navigation of vessels at sea, as shown in Figure 1 [3].
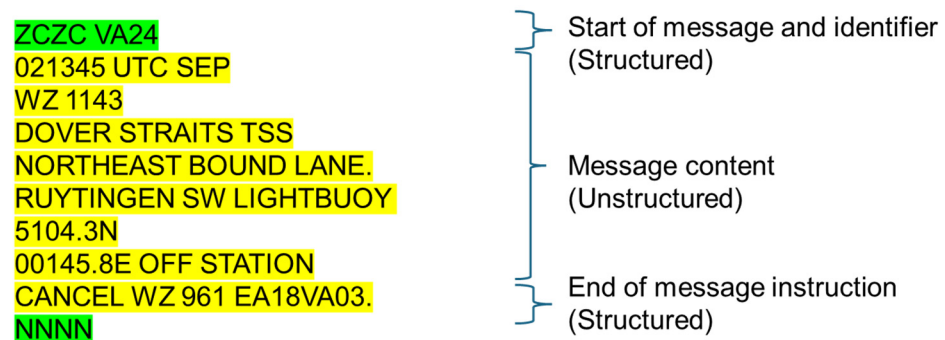

Figure 1. Structure of NAVTEX message.

The first part is the header of the message, which identifies the start of the NAVTEX message. All NAVTEX messages always start with the character "ZCZC", indicating the beginning of the message. In the header, there are technical codes that transmit the identification character (B1), the subject indicator character (B2), and the message numbering characters (B3B4). The second part is the main content of the NAVTEX message, including the transmission time, date of the message, navigational warnings, weather forecasts, and critical maritime safety information. This part contains various detailed information, depending on the situation, and provides clear and specific guidance. Especially, this part is made up of an unstructured text format to express various situations. The third part is the end of message instruction; all NAVTEX messages end with the word "NNNN" to indicate the end of the message.

2.2. Common Maritime Data Structure and S-100 Standards

Data compatibility is necessary for smooth data exchange, sharing, and utilization between ship and shore, shore and ship, shore and shore, and shipboard systems. The IMO enacted CMDS as a common data model that defines data structures and formats for data exchange, sharing, and utilization [7]. It has adopted the International Hydrographic Organization's (IHO's) S-100 standard as its base model. The IHO S-100 standard provides the complete framework necessary for the lifecycle management of maritime data such as data modeling, metadata, exchange formats, and data quality management. Through the S-100 standard, numerous maritime data can be generated, distributed, and utilized in a standardized format.

The S-124, a product of IHO S-100-based product specification provides navigational warnings and their information in a structured format [8]. And it certifies the consistency and accuracy of the information and intensifies interoperability between various maritime information systems. Through the S-124 product, seafarers can quickly and accurately access the necessary safety information.

NAVTEX messages are structured through the data structure of the S-124 standard. For example, NAVTEX messages about navigational warnings or weather forecasts are classified by the 'Warning Hazard Type' of the S-124 'Navigational Warning Feature' as shown in Figure 2. The detailed information of each message is broken down into various attributes of S-124, such as 'Geometry' and 'Restriction'. Through this, NAVTEX messages can be converted into a structured data format and integrated into standard data structures like CMDS.

2.3. Bi-LSTM CRF as Natural Language Processing Technology

Natural language processing (NLP) is one of the applications of artificial intelligence that makes computers understand, generate, and process human language [9]. It aims to extract meaningful information from text or speech data used by humans and performs numerous tasks. The problem of NLP is due to the fact that the main characteristic of natural language is hard for computers to understand due to numerous factors such as ambiguity, metaphors and similes, and irregular grammar. To solve this problem, NLP uses technologies and knowledge from various fields such as linguistics, computer science, and artificial intelligence to address these challenges of context dependency and ambiguity.

When it was a period of NLP in deep learning's infancy, recurrent neural networks (RNNs) were used. However, there were some problems in RNNs, such as gradient vanishing problems and long-term dependencies [10]. To solve that problem, LSTM, a variant of RNN, was used to introduce hidden state and cell state at one time to maintain long-term memory. Bi-LSTM is a model that stacked LSTM bidirectionally, allowing the capture of both past and future context in sentences [6]. Conditional random field (CRF) is a statistical model that deals with sequence labeling problems [11,12]. It predicts the optimal label sequence by considering the transition probability between labels and the emission probability between input features and labels. Figure 3, it shows the structure of the Bi-LSTM CRF Model. It is a model that combines Bi-LSTM and CRF as one, specialized for sequence labeling problems. This model extracts contextual information and model dependencies between labels from the input sequence using Bi-LSTM and CRF.

2.4. Begin, Inside, Outside (BIO) Tagging Scheme

The BIO tagging scheme is used very frequently in named entity recognition (NER) for solving the sequence labeling problem [13]. In this tagging scheme, tokens are labeled as being at the 'Begin', 'Inside', or 'Outside' of a named entity. Tags like 'B-' start with the entity name, 'I-' means inside of a named entity, and 'O-' means outside of any named entity. For example, the sentence "John Smith lives in New York" may be annotated with the use of the BIO tagging schema like that in Table 1.

The BIO tagging scheme allows for a clear indication of entity boundaries. Through this advantage, unstructured natural language can be easily converted into a structured machine-readable data structure with a clear indication of entity boundaries.

Table 1. Pairs of words and tags for an example sentence.

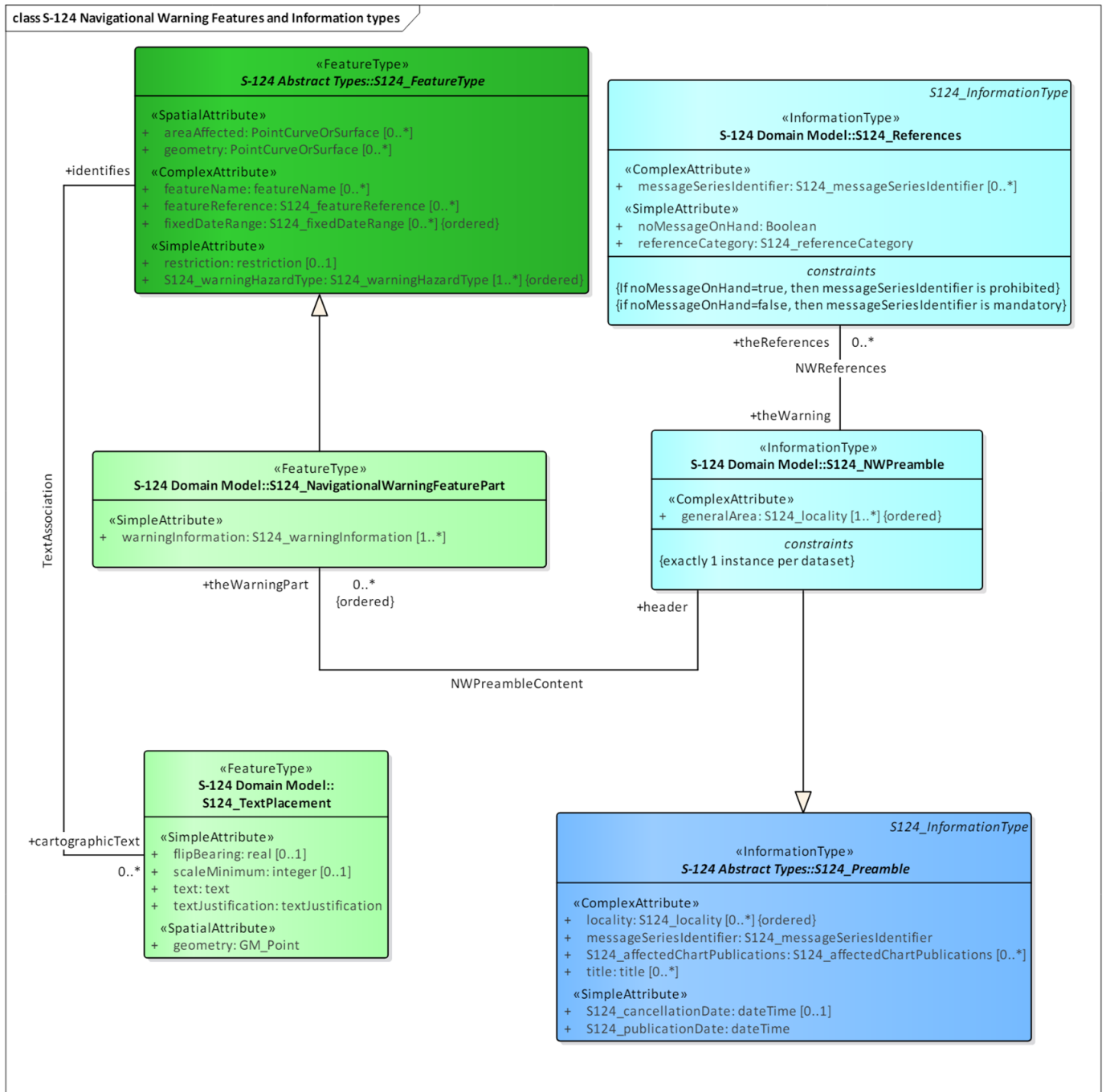| Word	Tag
John	B-PERSON
Smith	I-PERSON
Lives	O
In	O
New	B-LOCATION
York	I-LOCATION

Figure 2. Navigational warning features and information types in S-124 standard.
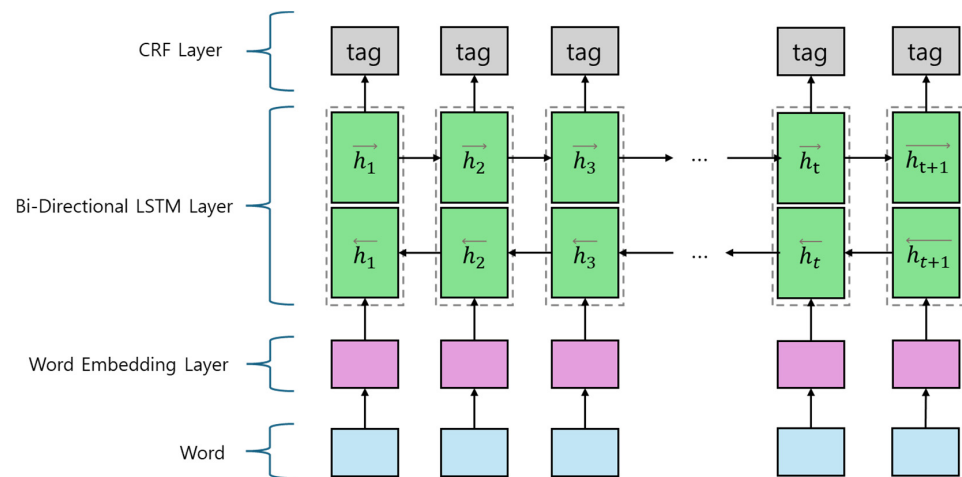
Figure 3. Structure of Bi-LSTM CRF model.

2.5. Literature Review

Recent studies have increasingly applied NLP technologies in the maritime sector, focusing on various aspects such as the classification of information on maritime accidents and NAVTEX message classification.

Mackenzie et al. [14] proposed a method for extracting information about piracy activities using deep learning techniques from unstructured maritime news articles. Their approach identifies maritime incidents in the articles, extracts information related to piracy using named entity recognition (NER), and analyzes the type and severity of these incidents. Jidkov et al. [15] developed a maritime event log (MEL) pipeline that collects information from various maritime documents and uses NLP and deep learning techniques, including NER, to automate the extraction of navigational risks and the classification of information.

Recently, some studies have also focused on classifying NAVTEX messages using machine learning. Sun et al. [16] proposed an adaptive-weighted TF-IDF-based classification model. In 2024, Sun et al. [17] further extended this field by developing a classification model for NAVTEX navigational warning messages using deep neural networks (DNNs) with an adaptive weighted TF-IDF approach. This study builds on earlier work on NAVTEX message classification and highlights the continued relevance of deep learning models in enhancing maritime safety information processing. Akyol and Keçeci [18] analyzed NAVTEX messages published for the Mediterranean region, categorizing them by annual message count, topic distribution, and type distribution to identify navigational risk trends.

Yan et al. [19] developed a content-aware, corpus-based model using BERT for the analysis of marine accidents, highlighting the effectiveness of BERT-based models in content recognition tasks within the maritime context. Shen et al. [20] explored the application of BERT in combination with Bi-LSTM and CRF for Chinese semantic named entity recognition (NER) in marine engine room systems, demonstrating the effectiveness of this layered approach in specialized maritime environments.

Ezen-Can [21] compared LSTM and BERT models for small corpora, showing that LSTM models could outperform BERT in terms of accuracy and performance when dealing with limited datasets. This study underscores the advantages of simpler models like LSTM when working with small datasets.

The study by Sun et al. focused on classifying the types of warnings in NAVTEX navigational messages, while Akyol and Keçeci examined navigational risk trends through manual literature analysis. This paper, however, focuses on the semantic classification of NAVTEX messages, an area not previously explored, and compares Bi-LSTM CRF models, which are advantageous for small datasets and limited-resource environments.

3. Experiment for Semantic Classification of NAVTEX Navigational Safety Information

For the harmonized sharing of NAVTEX messages to be machine-readable, the meaning of every word and phrase in the message must be put exactly into the structure of CMDS, and to achieve this, semantic classification of NAVTEX messages is required. Although recent NLP research predominantly utilizes transformer-based models such as BERT and GPT, this paper selects LSTM-based models because they can maintain performance even with the smaller size of the NAVTEX message dataset. Additionally, LSTM-based models are advantageous in industrial systems with limited resources due to their smaller model size. Therefore, in this study, we conducted experiments on converting NAVTEX navigational safety messages into machine-readable data using six different application models based on Bi-LSTM CRF and analyzed the results to find an optimal model.

Figure 4 shows the whole experimentation for finding the optimal model to semantically classify NAVTEX messages. First, this is a data collection stage where the NAVTEX messages are collected from different sources for the preparation of data that is to be analyzed. Afterwards, in the data preprocessing stage, irrelevant information is removed from the collected data and converted into a form suitable for learning. The data is then learned through six types of Bi-LSTM CRF models. After that, the semantics of the NAVTEX messages are classified through those trained models for their performance evaluation. Evaluation and model size comparisons are performed for all models, analyzing the optimum model.
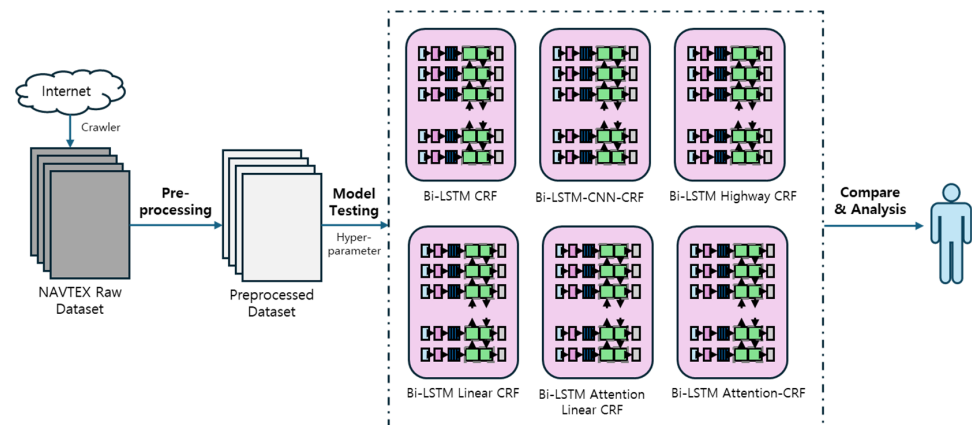

Figure 4. Flowchart of experiment procedure.

3.1. Preprocessing

For the experiment, 8541 international NAVTEX messages were collected from the NAVTEX message archive provided by Plovput in Croatia. These messages were transmitted by the split radio station from August 2020 to March 2024 and were written in English.

The preprocessing of the collected data involved three steps: filtering, cleansing, and semantic tagging. First, during the data filtering stage, messages relevant to the experiment were selected. Specifically, we chose only the navigational warning messages, identified by type code A, as these were the focus of the study. Among the selected messages, many were notices to maintain or cancel alerts, which could introduce bias in semantic analysis. Therefore, the number of messages was adjusted, resulting in 1634 messages remaining.

In the second stage of preprocessing, the cleansing process involved removing unnecessary symbols, spaces, and extraneous characters from the messages using regular expressions. First, symbols in parentheses containing letters or numbers were removed according to Equation (1) so that all the reference symbols not directly relevant to the message content are removed. This removed "(A)", "(B1)", etc. Then, alphabets followed by a period and number according to Equation (2) were removed so that the numbering that was irrelevant to message content had been deleted. Thus, for example, "A.1", "B.2",

etc. were deleted. Finally, Equation (3) was used for clearing special symbols unrelated to the core content of the message.

$$([A\text{-}Za\text{-}z0\text{-}9]) \tag{1}$$

$$[A\text{-}Za\text{-}z].[0\text{-}9] \tag{2}$$

$$[/,:;=\text{-}+\text{\textasciitilde}!\backslash[]?<>()] \tag{3}$$

Figure 5 shows an example of the original message (a) and the cleansed message (b) after going through the message cleansing process using regular expressions. For instance, dates were originally formatted with slashes, such as "27/03/2023", where the slashes were considered unnecessary symbols. These were removed to produce a standardized format like "27032023".

Original Message

```
ZCZC QA29
270640 UTC MAR 23
SPLIT RADIO
WEATHER BULLETIN FOR ADRIATIC ISSUED
BY THE MARINE METEOROLOGICAL SPLIT 0N
27/03/2023 AT 0400 UTC

...
NW 10-20 KT, AT FIRST UP TO 28, IN VELEBITSKI KANAL NE 10-20 KT,
OVERDAY SLIGHT DECR. IN THE EVENING AND OVERNIGHT
TURNING ON N/NE, QCKY INCR 14-28 KT, ON S-ERN PART 18-32, IN
VELEBITSKI KANAL UP TO 42 KT.
SEA 2-3, AT FIRST AND AGAIN IN THE EVENING 3-4, AFTER
MIDNIGHT POS 5.
VIS 10-20 KM.
VRB CLOUDY WITH RAIN, T-SHWRS, MOSTLY AT FIRST AND IN THE
EVENING.

...
NNNN
```

(**a**)

Cleansed Message

```
ZCZC QA29
270640 UTC MAR 23
SPLIT RADIO
WEATHER BULLETIN FOR ADRIATIC ISSUED
BY THE MARINE METEOROLOGICAL SPLIT 0N
27032023 AT 0400 UTC

...
NW 10 20 KT AT FIRST UP TO 28 IN VELEBITSKI KANAL NE 10 20 KT
OVERDAY SLIGHT DECR. IN THE EVENING AND OVERNIGHT
TURNING ON NNE QCKY INCR 14 28 KT ON S ERN PART 18 32 IN
VELEBITSKI KANAL UP TO 42 KT.
SEA 2 3 AT FIRST AND AGAIN IN THE EVENING 3 4 AFTER
MIDNIGHT POS 5.
VIS 10 20 KM.
VRB CLOUDY WITH RAIN T SHWRS MOSTLY AT FIRST AND IN THE
EVENING.

...
NNNN
```

(**b**)

Figure 5. Example of original message and cleansed message by RegEx Rules: (**a**) original message and (**b**) cleansed message by RegEx rules.

In the third stage of preprocessing, semantic tagging was performed on each token of the messages to facilitate supervised learning. To train machine learning models on NAVTEX messages, semantic tags presented in Table 2 should be assigned to the tokens of NAVTEX messages and represented as pairs of tokens and semantic tags.

This task is normally performed manually by human beings, but it is very time-consuming; therefore, we prepared the training data through the prompt engineering technique using Claude, an LLM. Human experts were involved in the re-examination of the prepared training data according to rules and corrected the wrongly tagged semantic tags to check data integrity. Messages that were not fully received through radiocommunication and thus did not have a complete message format could not be processed, even with human expert intervention. Therefore, these incomplete messages were removed, resulting in a final dataset of 1617 messages. Figure 6 illustrates token–semantic tag pairs annotated with the BIO scheme. In the BIO scheme, 'ZCZC' was given a tag meaning the start of an entity name and the start of the message header. These works were assigned a tag, meaning the inside of an entity name and an action to be taken by the recipient. Abbreviations commonly used in the maritime field to quickly convey information were also considered during the tagging process. For example, in Figure 6, 'VRB' is an abbreviation frequently used to convey 'Variable' quickly.

For model training, the 1617 messages were randomly divided into a train set, a validation set, and a test set at ratios of 80%, 10%, and 10%, respectively, resulting in 1293 messages for training, 161 for validation, and 163 for testing. The 80-10-10 split is

a prevalent approach in machine learning, providing sufficient data for training while minimizing the risk of overfitting and ensuring reliable evaluation of model performance.

Table 2. Semantic tags used for tagging NAVTEX Messages and their descriptions.

Semantic Tag	Description
METADATA_START	Start header of message. Always starts with characters, "ZCZC"
METADATA_CODE	The 4-digit message identification code that follows METADATA_START.
METADATA_DATETIME	Date and Time of Metadata.
METADATA_END	The termination string ending with "NNNN" at the end of the message.
CONTENT_EVENTS	Specific maritime or navigation-related events
CONTENT_INSTRUCTION	Instructions or actions the recipient should take
CONTENT_DATETIME	Date and time mentioned in the message
CONTENT_LOCATION_TYPE	Type of location mentioned (Point, Area, etc.)
CONTENT_LOCATION_POINT	Latitude and longitude coordinates
CONTENT_LOCATION_NUMBER	Numbers indicating distance or radius (e.g., "within a radius of 500 m" where 500 is the number)
CONTENT_LOCATION_UNIT	Units of distance or area
CONTENT_LOCATION_PLACENAME	Names of places, ship names, IMO numbers, etc.
CONTENT_OTHER	Other content that does not fall into the defined categories

```
ZCZC,B-METADATA_START
QA29,B-METADATA_CODE
270640,B-METADATA_DATETIME

            ⋮

VRB,B-CONTENT_EVENTS
CLOUDY,I-CONTENT_EVENTS
WITH,I-CONTENT_EVENTS
RAIN,I-CONTENT_EVENTS
T,I-CONTENT_EVENTS
SHWRS,I-CONTENT_EVENTS
MOSTLY,I-CONTENT_EVENTS
AT,I-CONTENT_EVENTS
FIRST,I-CONTENT_EVENTS
AND,I-CONTENT_EVENTS
IN,I-CONTENT_EVENTS
THE,I-CONTENT_EVENTS
EVENING,I-CONTENT_EVENTS
NNNN,B-METADATA_END
```

Figure 6. Pairs of token and BIO-schemed semantic tags.

3.2. Training of Bi-LSTM CRF-Based Application Models

In this study, six model types for the semantic classification of NAVTEX navigational safety messages were prepared. As shown in Figure 7, the Bi-LSTM CNN CRF model builds on the Bi-LSTM CRF architecture by adding a convolutional neural network (CNN) layer between the Bi-LSTM and CRF layers. The word-embedding layer first converts input tokens into dense vector representations. These vectors are then processed by the Bi-LSTM layer, which captures contextual information from both directions in the sequence. The CNN layer follows, extracting local features from the Bi-LSTM outputs to enhance representation. Lastly, the CRF layer performs structured prediction, ensuring coherent output sequences that accurately reflect the semantic content of the NAVTEX messages.

In addition to Bi-LSTM CNN CRF, to improve performance, models were prepared by adding a highway network layer, linear layer, or attention layer instead of a CNN layer between the Bi-LSTM layer and CRF layer in Figure 7, or by stacking the attention layer and linear layer to create a Bi-LSTM attention linear CRF model. Table 3 represents the architecture, expected advantages, and disadvantages of Bi-LSTM CRF model and its variants.

We used WandB and the Bayesian optimization algorithm to select the hyperparameter combination based on previous step results for required training with hyperparameters. The AdamW optimizer, which applies weight decay to the Adam algorithm, and a loss function called negative log-likelihood were used. In this case, the batch size, learning rate, embedding dimension, hidden dimension, and dropout were specified with their respective ranges and values, as in Table 4, to find the hyperparameters corresponding to the lowest validation loss during 10 training epochs.
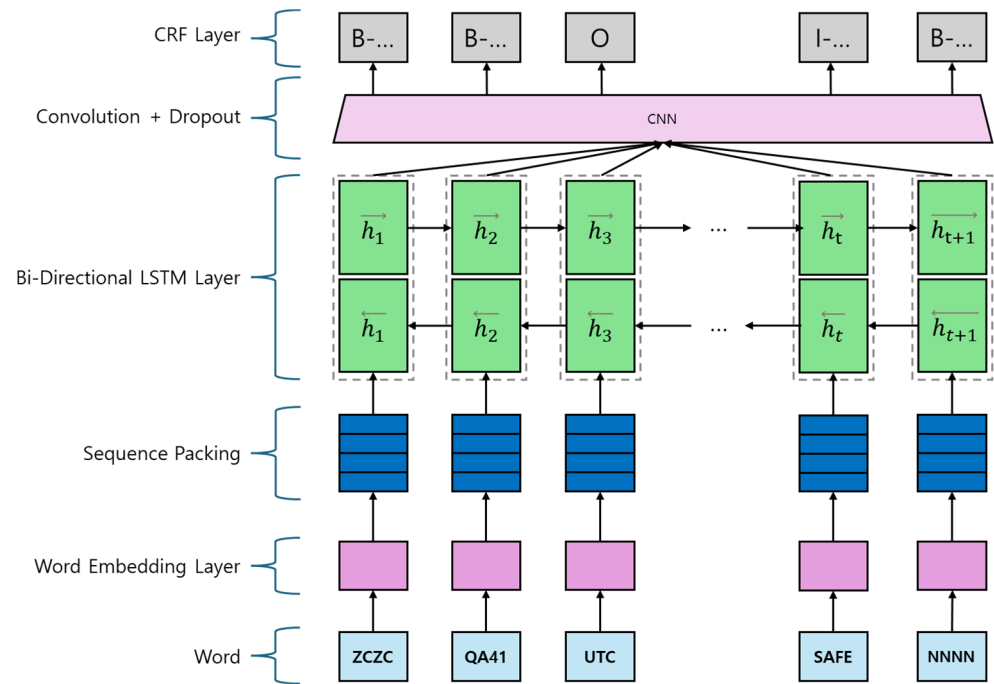
Figure 7. Architecture of Bi-LSTM CNN CRF.

Table 3. Pros and cons of Bi-LSTM CRF and its variant models.

	Layers	Expected Advantages	Expected Disadvantages
Bi-LSTM CRF	Embedding layer, Bi-directional LSTM layer, CRF layer	Captures contextual information and label dependencies well	Limited in learning complex patterns
Bi-LSTM CNN CRF	Embedding layer, Bi-directional LSTM layer, CNN layer, CRF layer	Enhances expressiveness by extracting local features through CNN	Increased computational load, leading to longer training time
Bi-LSTM Highway CRF	Embedding layer, Bi-directional LSTM layer, Highway network, CRF layer	Improves learning capability by controlling information flow with highway network	Potential decrease in training stability due to gate mechanism
Bi-LSTM Linear CRF	Embedding layer, Bi-directional LSTM layer, Linear layer, CRF layer	Optimizes CRF input by transforming LSTM outputs with linear layer	Increased model complexity, risk of overfitting
Bi-LSTM Attention CRF	Embedding layer, Bi-directional LSTM layer, Attention mechanism, CRF layer	Utilizes important contextual information for modeling label dependencies by directly connecting attention and CRF	Increased training time due to computational load of attention layer
Bi-LSTM Attention Linear CRF	Embedding layer, Bi-directional LSTM layer, Attention mechanism, Linear layer, CRF layer	Enhances expressiveness by assigning high weights to important contextual information through attention layer	Increased model complexity and computational load due to attention layer

Table 4. Hyperparameters and its ranges.

Hyperparameter	Searching Range
Batch size	8, 16, 32, 64
Learning rate	$1 \times 10^5 \sim 0.5$
Embedding dimension	16, 32, 64, 128, 256, 512, 768
Hidden dimension	16, 32, 64, 128, 256, 512, 768
Dropout	$1 \times 10^5 \sim 0.1$

Considering the sensitivity of the experiments, the hyperparameter values were set within a defined range rather than being fixed to specific values. In each iteration, the optimal hyperparameters were selected from within this range, taking into account their sensitivity. This approach was used to minimize the likelihood of obtaining favorable results merely by chance due to different hyperparameter settings.

For model training and testing, the hardware system used included an Intel® Xeon® Gold 5218R CPU, an NVIDIA A100 80 GB GPU, and 80 GB of memory, running on the Ubuntu 20.04 operating system. The software framework utilized for implementation included CUDA 11.8, CUDNN 8.7.0, and PyTorch 1.9.0.

4. Results and Analysis

The semantic classification of NAVTEX navigational safety messages was repeated 10 times for each of the six models: Bi-LSTM CRF, Bi-LSTM CNN CRF, Bi-LSTM highway CRF, Bi-LSTM linear CRF, Bi-LSTM attention CRF, and Bi-LSTM attention linear CRF.

4.1. Analysis in Aspects of Performance

Figure 8 illustrates the F1 score over 10 iterations for each model, allowing for the assessment of both the performance and stability of the models involved. The F1 score represents model performance, with variance being used as an indicator of stability.
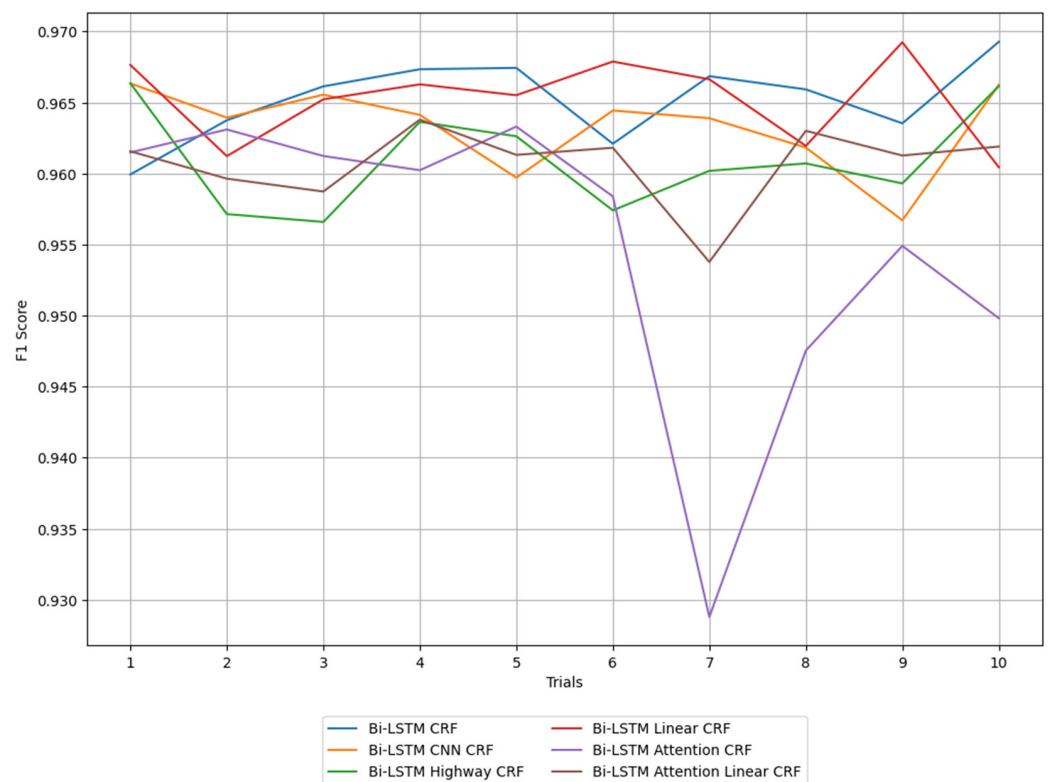

Figure 8. Performance results of Bi-LSTM CRF and its variant models.

Model performance is the most critical factor in determining its usefulness. To better understand the model's performance, it should be evaluated from multiple perspectives using various metrics. Beyond basic evaluation measures like accuracy and precision, advanced evaluation measures such as AUC-ROC, perplexity, and cosine similarity should also be utilized to evaluate the model from different angles. In this paper, we chose to evaluate model performance using the F1 score, which is a more practical metric for classification tasks. The F1 score represents the harmonic mean of precision and recall, meaning a high F1 score indicates a model that not only predicts accurately but also effectively identifies most of the actual positive classes. Equation (4) defines the F1 score

as follows: precision is the proportion of true positive classes among those predicted as positive by the model, while recall is the proportion of actual positive classes that the model correctly identifies. Thus, the F1 score serves as a key metric for comprehensively evaluating model performance, as it balances these two important indicators.

$$F1 \, Score = 2 \times \left(\frac{Precision \times Recall}{Precision + Recall} \right) \tag{4}$$

Variance of the F1 score for every model was computed to check on the stability of the models. Variance is a statistical indicator, depicted in Equation (5), which demonstrates how data points are spread out from the mean and becomes an important component in the assessment of how consistently the performance of model predictions are upheld.

$$Variance = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \tag{5}$$

Here, N is the number of data points, x_i is individual data point, and μ is the mean value of the F1 score. When the variance of the F1 score is low, it indicates that the model's performance is consistently maintained, suggesting that the model performs well across different datasets. High variance, on the other hand, means that performance fluctuates significantly with respect to the dataset used, indicating low stability and the unreliability of the model. Therefore, models with low variance are likely to show stable performance in real operational environments due to their high consistency in predictions.

Table 5 shows the minimum, maximum, average, and variance of the F1 score for each model. The minimum F1 score refers to a case where a model had the lowest performance, while the maximum value refers to when a model had the best performance. The average value describes the arithmetic mean of F1 scores over several iterations and gives knowledge regarding the general performance level of the model. The variance at the same time shows how far apart the values of F1 score move around the mean.

Table 5. Performance results of Bi-LSTM CRF and its variant models.

Models	F1 Score			Variance
	Minimum	Maximum	Average	
Bi-LSTM CRF	0.959932	0.969272	0.965227	0.000008
Bi-LSTM CNN CRF	0.956711	0.966339	0.963277	0.000009
Bi-LSTM Highway CRF	0.956591	0.966355	0.960198	0.000012
Bi-LSTM Linear CRF	0.960440	0.969239	0.965152	0.000010
Bi-LSTM Attention CRF	0.928789	0.963302	0.954988	0.000116
Bi-LSTM Attention Linear CRF	0.953776	0.963793	0.960680	0.000009

The evaluation of each model's performance in terms of the F1 score and its variance shows that while there is little difference in the overall F1 score, the significant differences in variance between the models suggest different levels of stability. Interestingly, the basic Bi-LSTM CRF model had the highest average F1 score and the lowest variance among all the models in the Bi-LSTM CRF series.

In the Bi-LSTM CNN CRF model, the CNN layer can effectively capture local patterns and may be useful with sequence data; however, its effect may not be prominent due to the characteristics of NAVTEX messages, where global contextual information may be more relevant. This suboptimal performance compared to the Bi-LSTM CRF model might be due to the CNN layer's bias toward local patterns, which limits its ability to capture the overall context of NAVTEX messages. In the Bi-LSTM highway CRF model, the highway network controls the flow of information between layers, making it easier to learn in deep networks; however, its influence might be limited for the task at hand, similar to other NAVTEX message analysis tasks where network depth might not be as crucial. The relatively lower performance

may be due to the fact that, although the highway network supports deep learning, a simpler structure might be more appropriate for the characteristics of NAVTEX messages.

The Bi-LSTM linear CRF model was high in average F1 score, and its variance was stable. This model converts the output of the Bi-LSTM via a linear layer by using the activation function ReLU before passing it into the CRF layer. The linear layer extracts core information and filters out unnecessary information, simultaneously increasing recall and precision. These characteristics further improved the model's structural performance and stability. The reason it performed just as high as the Bi-LSTM CRF is that the linear layer effectively transforms the output of Bi-LSTM, thus optimizing the tagging performance of CRF.

The Bi-LSTM attention CRF model shows relatively high variance compared to other models, indicating greater variability in performance. Although the optimal hyperparameters were selected based on sensitivity considerations within the range shown in Table 5, the F1 score dropped significantly in the 7th iteration due to the use of suboptimal hyperparameters. This means that finding the optimal hyperparameters requires considerable effort, and to ensure stable model performance, it is important to consider variance when evaluating models. Additionally, the Bi-LSTM attention CRF model has structural limitations when processing NAVTEX messages. The attention mechanism gives extra weight to the important parts of the sequence. However, in NAVTEX messages, the content is typically spread out, and important information might not be highly concentrated in specific parts, which could limit the effectiveness of the attention mechanism. This is likely the reason for the inconsistent behavior of this model.

The Bi-LSTM attention linear CRF model maintained consistent performance and provided stable predictions compared to the Bi-LSTM attention CRF model, owing to the combination of the attention mechanism and the linear layer. This is because, even though important information in NAVTEX messages tends to be spread out rather than concentrated in specific parts, the linear layer effectively filtered this information, optimizing the model's performance.

4.2. Analysis in Aspects of Model Size

In real-world applications, such as on ships, not only is the performance of the model crucial, but also resource efficiency, including memory usage and inference speed, due to the limited computing resources available. In such constrained environments, it is important to evaluate how efficiently the model uses memory and how much its size fluctuates. Particularly in mobile environments like ships, a model that maintains stable and consistent performance while using minimal resources is essential. Figure 9 visually represents the size of each model over 10 iterations. This figure allows us to observe how each model's size fluctuates.

To comprehensively evaluate the model sizes, the minimum, maximum, average, and variance of each model's size are presented in Table 6. While the size of the model is important, the variability in model size is a crucial indicator of how consistently the model can operate in real-world applications. For instance, if the variability in model size is high, excessive memory usage under certain conditions could lead to a decline in system performance.

Analyzing the size of each model, we can see that the Bi-LSTM CRF model has a moderate size despite the absence of any layer between the LSTM and CRF layers. Additionally, the moderate variance among the models suggests some fluctuation in model size with different hyperparameter settings. As the base model of the Bi-LSTM CRF series, it appears to have a mid-level size and variability.

The Bi-LSTM CNN CRF model has a relatively small size, with both a low average and low variance. Although the CNN layer is added to capture local patterns, it does not significantly impact the model size. The Bi-LSTM highway CRF model shows the largest average size and the highest variance. The highway network facilitates learning in deep networks by controlling the flow of information between layers, but this significantly increases the model size. The high variance suggests that the model size can vary greatly with different hyperparameter settings.
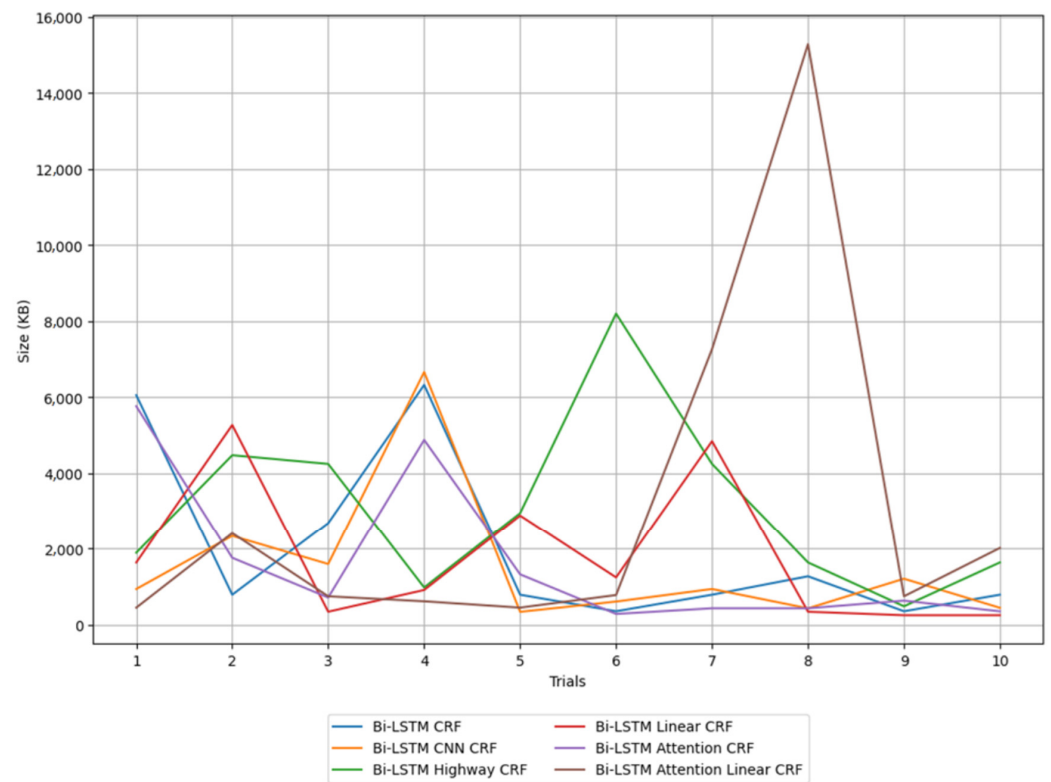
Figure 9. Size of Bi-LSTM CRF and its variant models.

Table 6. Size of Bi-LSTM CRF and its variant models.

Models	Size (Kbyte)			Variance
	Minimum	Maximum	Average	
Bi-LSTM CRF	354	6323	1882	3,855,208
Bi-LSTM-CNN-CRF	339	6662	1499	3,303,771
Bi-LSTM Highway CRF	484	8198	3042	6,274,171
Bi-LSTM Linear CRF	250	5268	1629	1,970,368
Bi-LSTM Attention CRF	284	5761	1956	2,849,085
Bi-LSTM Attention Linear CRF	449	15,282	3125	18,462,097

The Bi-LSTM linear CRF model has a relatively small size and low variance. This model transforms the output of the Bi-LSTM through a linear layer using the ReLU activation function before passing it to the CRF layer. The linear layer effectively filters the necessary information, reducing the model size without greatly increasing its complexity.

The Bi-LSTM attention CRF model is of medium size. The attention mechanism enhances model performance, but it also causes an increase in model size. The model's variability in size is relatively low, indicating that it maintains a stable size across various hyperparameter settings.

The Bi-LSTM attention linear CRF model has the largest size and the highest variance. Notably, in the 8th iteration, the model size increased significantly due to the selection of a large hidden layer dimension during the hyperparameter optimization process. This sudden increase in model size, resulting from suboptimal hyperparameters, can be considered a drawback. From a structural perspective, the combination of the attention mechanism and the linear layer enhances model performance, but it also significantly increases the model size. The high variance suggests that the model size can fluctuate considerably with different hyperparameter settings.

4.3. Comprehensive Analysis and Limitations

We analyzed the performance and size of each model in the semantic classification of NAVTEX messages, considering them comprehensively. The Bi-LSTM-CRF model, as the baseline model among the compared models, showed the highest average F1 score with low variance, indicating excellent consistency in performance. Additionally, the model size remained at a relatively moderate level with little variability, making it an efficient choice in terms of hardware resources.

The Bi-LSTM-CNN-CRF model, in terms of performance, effectively captures local patterns with its CNN layer, making it useful for sequence data analysis. However, due to the nature of NAVTEX messages, where global contextual information may be more important, its effectiveness may be limited. The Bi-LSTM linear CRF model showed performance and stability levels similar to those of the Bi-LSTM-CRF model, with a relatively small and efficient model size. The linear layer, using the ReLU activation function, effectively transforms the output of the Bi-LSTM and passes it to the CRF layer, efficiently extracting necessary information and filtering out unnecessary data. These characteristics provide a balanced outcome in both performance and model size.

On the other hand, the Bi-LSTM attention-CRF and Bi-LSTM attention linear CRF models use the attention mechanism to emphasize important information. However, due to the nature of NAVTEX messages, their effectiveness is limited. In particular, the Bi-LSTM attention linear CRF model shows high performance but has the largest model size and very high variance, making it less efficient in terms of hardware resource usage. Therefore, while these two models have high performance, they lack resource efficiency.

Considering the performance and model size together in semantic classification of the NAVTEX messages, the balanced choices are the Bi-LSTM-CRF model and the Bi-LSTM-Linear-CRF model. These two models maintained high performance, low variance, and an appropriate model size, effectively performing the semantic classification of NAVTEX navigational safety messages.

This paper's experiment, however, deals with navigational safety messages only among NAVTEX messages. Navigational safety messages are unstructured but consist of sentences with somewhat fixed formats centered on keyword enumeration; it doesn't have completely natural language features. Hence, the basic Bi-LSTM-CRF model and Bi-LSTM-Linear-CRF using a simple ReLU activation function showed good results. For the rest of the NAVTEX message types, the natural language characteristics would be more dominant and therefore would have to be considered all together to select an optimal model for semantic classification. Moreover, the dataset used in this study is limited to messages from the split radio station in Croatian waters, which may introduce bias related to specific location information and word choices commonly used in that region. This geographic and linguistic bias could limit the model's generalizability to other regions or languages, potentially leading to errors when deployed in different maritime contexts. To mitigate these risks, it is essential to train models on a more diverse dataset, encompassing a broader range of sea areas and message types.

It is important to recognize that relying on Bi-LSTM models for the analysis and classification of NAVTEX messages comes with certain risks. One potential risk is the occurrence of misclassifications or errors in the analysis, which could result from various factors such as insufficient training data, model overfitting, or inherent ambiguities in the unstructured text of NAVTEX messages. For instance, a misclassified navigational warning could lead to incorrect or delayed responses by automated systems like ECDIS or autopilot, potentially compromising the safety of the vessel.

To minimize the impact of these potential risks, future work should focus on integrating error detection mechanisms, improving model robustness through ensemble methods, and incorporating human-in-the-loop systems where critical decisions are cross-validated by human operators. This approach will help ensure that the safety-critical information processed by these models remains accurate and reliable, thereby supporting the safe navigation of smart ships.

5. Conclusions

NAVTEX is one of the major components of the GMDSS, transmitting important navigational and meteorological warnings, forecasts, and other urgent maritime safety information to ships. The automatic exchange of this information with systems like ECDIS or autopilot, without human intervention, to adjust routes or display messages is crucial for the safe navigation of smart ships. Therefore, this information should be consistently shared and integrated in line with the Common Maritime Data Structure (CMDS).

In this paper, we have implemented and compared six models: Bi-LSTM CRF, Bi-LSTM CNN CRF, Bi-LSTM highway CRF, Bi-LSTM linear CRF, Bi-LSTM attention CRF, and Bi-LSTM attention linear CRF for semantic classification regarding NAVTEX messages, with a particular emphasis on all navigation safety messages in the context of NAVTEX messages. We performed an analysis of the F1 Score and model size comprehensively to evaluate the performance and resource efficiency to derive an optimal model. Based on the experimental results, Bi-LSTM CRF and Bi-LSTM linear CRF have the most effective performance in semantic classification for messages in the NAVTEX messages. The Bi-LSTM CRF model achieved an average F1 score of 0.965227 with a variance of 0.000008 and an average size of 1882 KB with a variance of 3,855,208 KB. The Bi-LSTM Linear CRF model showed an average F1 score of 0.965152 with a variance of 0.000010 and an average size of 1629 KB with a variance of 1,970,368 KB. These models could demonstrate a high F1 score in addition to extremely low variance and prove to be very resource efficient with their relatively very small model size.

In the future, we will conduct studies that evaluate models by considering features not only of navigational safety messages but also of other message types, including those in the broader NAVTEX series, as well as data from more diverse sea regions. Furthermore, we will compare model performance using various evaluation measures from multiple perspectives and explore state-of-the-art techniques based on transformers.

In this paper, we compared and analyzed several models in the Bi-LSTM CRF series and implemented the one most suitable to understand the semantics of NAVTEX messages. The optimal model derived enables not only automated processing of NAVTEX messages but also conversion into a machine-readable format fitting the structure of CMDS. This would provide continuous and efficient sharing of maritime information, which would result in improved operational efficiency and maritime safety for intelligent ships, and further promote development within the entire maritime industry.

Author Contributions: Conceptualization, C.L. and S.L.; methodology, C.L.; software, H.C.; validation, C.L. and S.L.; formal analysis, C.L. and H.C.; investigation, C.L. and S.L.; resources, C.L. and H.C.; data curation, H.C.; writing—original draft preparation, C.L.; writing—review and editing, S.L.; visualization, H.C.; funding acquisition, S.L. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: https://www.plovput.hr/en/radio-service/navtex-system/navtex-archive-q (accessed on 31 August 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. International Maritime Organization. *MSC.1/Circ.1403/Rev.2 NAVTEX Manual—Section 1*; IMO: London, UK, 2022; p. 2.
2. International Hydrographic Organization. 15th Meeting of the Hydrographic Services and Standards Committee. In Proceedings of the Report of the MASS Navigation Project Team & Recommendations-Maritime Autonomous Surface Ships and S-100, HSSC-15, Helsinki, Finland, 5–9 June 2023; p. 13.
3. International Maritime Organization. *MSC.1/Circ.1403/Rev.2 NAVTEX Manual—Section 7*; IMO: London, UK, 2022; pp. 21–22.

4. Joint IMO/IHO/WMO. *Manual on Maritime Safety Information (MSI)*; International Maritime Organization: London, UK, 2015; pp. 1–76.

5. Grune, D.; Jacobs, C.J.H. *Parsing Techniques: A Practical Guide*, 2nd ed.; Springer: Berlin/Heidelberg, Germany, 2008; Monographs in Computer Science; Section: Preface.

6. Huang, Z.; Xu, W.; Yu, K. Bidirectional LSTM-CRF Models for Sequence Tagging. *arXiv* **2015**, arXiv:1508.01991. [CrossRef]

7. International Maritime Organization. *ANNEX 20 Strategy for the Development AND Implementation of e-Navigation*; MSC 85/26/Add.1; IMO: London, UK, 2008; p. 6.

8. International Hydrographic Organization. *Navigational Warnings Edition 1.0.0–May 2023*; IHO: Monaco, Monaco, 2023; p. 10.

9. Jurafsky, D.; Martin, J.H. *Speech and Language Processing*, 2nd ed.; Prentice-Hall, Inc.: Upper Saddle River, NJ, USA, 2009; Chapter 1 Introduction.

10. Pascanu, R.; Mikolov, T.; Bengio, Y. On the difficulty of training recurrent neural networks. In Proceedings of the 30th International Conference on Machine Learning (ICML 2013), Atlanta, GA, USA, 16 June 2013; Volume 28, pp. III–1310–III-1318.

11. Graves, A.; Schmidhuber, J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw.* **2005**, *18*, 602–610. [CrossRef] [PubMed]

12. Lafferty, J.D.; McCallum, A.; Pereira, F.C.N. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML), San Francisco, CA, USA, 28 June–1 July 2001; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 2001.

13. Ramshaw, L.A.; Marcus, M.P. Text Chunking Using Transformation-Based Learning. In *Natural Language Processing Using Very Large Corpora*; Armstrong, S., Church, K., Isabelle, P., Manzi, S., Tzoukermann, E., Yarowsky, D., Eds.; Springer: Dordrecht, The Netherlands, 1999; pp. 157–176.

14. Mackenzie, A.; Teske, A.; Abielmona, R.; Petriu, E. Maritime Incident Information Extraction using Machine and Deep Learning Techniques. In Proceedings of the 2021 IEEE Symposium Series on Computational Intelligence (SSCI), Orlando, FL, USA, 5–7 December 2021; pp. 1–6. [CrossRef]

15. Jidkov, V.; Abielmona, R.; Teske, A.; Petriu, E. Enabling Maritime Risk Assessment Using Natural Language Processing-based Deep Learning Techniques. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 1–4 December 2020; pp. 2469–2476. [CrossRef]

16. Sun, P.; Zuo, Y.; Wang, Y. Classification model for NAVTEX navigational warning messages based on adaptive weighted TF-IDF. In Proceedings of the 10th Multidisciplinary International Social Networks Conference, MISNC '23, New York, NY, USA, 4–6 September 2023; Association for Computing Machinery: New York, NY, USA, 2023; pp. 133–142. [CrossRef]

17. Sun, P.; Zuo, Y.; Li, X.; Wang, Y. Application of Deep Learning in the Classification of Maritime Safety Information. *Rev. Socionetwork Strat.* **2024**. [CrossRef]

18. Akyol, A.E.; Keçeci, T. Analysis of Navtex Messages Published for the Mediterranean Region in Terms of Safe Navigation of Ships. *Mersin Univ. J. Marit. Fac.* **2023**, *5*, 37–44. [CrossRef]

19. Yan, K.; Wang, Y.; Jia, L.; Wang, W.; Liu, S.; Geng, Y. A Content-Aware Corpus-Based Model for Analysis of Marine Accidents. *Accid. Anal. Prev.* **2023**, *184*, 106991. [CrossRef] [PubMed]

20. Shen, H.; Cao, H.; Sun, G.; Chen, D. Research on Chinese Semantic Named Entity Recognition in Marine Engine Room Systems Based on BERT. *J. Mar. Sci. Eng.* **2023**, *11*, 1266. [CrossRef]

21. Ezen-Can, A. A Comparison of LSTM and BERT for Small Corpus. *arXiv* **2020**, arXiv:2009.05451. [CrossRef]