Automating Maritime Risk Data Collection and Identification Leveraging Large Language Models

Donghao HUANG[1,2], Xiuju FU[3*], Xiaofeng YIN[3], Haibo PEN[4*], Zhaoxia WANG[1*]
School of Computing and Information Systems, Singapore Management University, Singapore[1]
Research and Development, Mastercard, Singapore[2]
Institute of High Performance Computing, Agency for Science, Technology and Research, Singapore[3]
School of Electrical and Information Engineering, Tianjin University, China[4]
dh.huang.2023@smu.edu.sg; fuxj@ihpc.a-star.edu.sg; yinxf@ihpc.a-star.edu.sg; penhaibo@tju.edu.cn; zxwang@smu.edu.sg*

***Abstract*—Maritime risk research is crucial yet challenging for improving safety, efficiency, and sustainability in maritime operations. This paper presents an innovative method for automating the collection and identification of risk data related to global maritime risks from news sources, addressing the limitations of traditional manual methods. To evaluate the proposed method, different learning-based models, including conventional machine learning approaches and advanced Large Language Models (LLMs) such as GPT-4 and LLaMA-3.1, are comprehensively studied for comparison. In addition, not only do we use popular evaluation metrics to assess the proposed method, but we also introduce a new evaluation metric, called the "Ratio of Valid Categories (RVC)," to evaluate model reliability. The merits of the proposed method are demonstrated across different evaluation metrics. The research results show that the proposed LLM-based methods, particularly the GPT-4-based method, consistently outperform traditional models, significantly improving both the efficiency and accuracy of maritime risk data collection and identification. Our findings contribute to the expanding literature on LLM applications in risk management, demonstrating their potential to transform data collection and identification practices.**

***Index Terms*—Maritime Risk, Automated Data Collection, Risk Identification, Large Language Models, Traditional Machine Learning, GPT-4o, Llama-3.1, Ratio of Valid Categories (RVC)**

I. Introduction

The collection and analysis of risk or anomaly data from online news sources are critical for effective risk management [1], [2]. Maritime risk management and research are especially essential. However, traditional methods for gathering and categorizing such data are often labor-intensive, time-consuming, and expensive [3]–[6]. Existing approaches rely heavily on manual effort for data labeling and annotation, which not only incurs significant operational costs but also limits scalability. For instance, using services like NewsAPI for comprehensive data collection can be prohibitively expensive[1], with search volumes insufficient to meet the needs of advanced risk information analytics.

Recent advancements in machine learning (ML) and natural language processing (NLP) have begun to address these challenges by automating parts of the data collection and classification pipeline [7]–[12].

Specifically:

- Teske et al. proposed a two-step NLP pipeline using traditional ML models like Logistic Regression and AdaBoost to classify maritime incident articles and extract relevant information [13].
- Jidkov et al. further enhanced this approach using deep learning techniques such as convolutional neural networks (CNNs) to identify specific types of incidents, although their models faced challenges like overfitting due to limited training data [14].
- Subsequent research by Mackenzie et al. introduced advanced metadata extraction techniques using models like CatBoost to improve the accuracy of information retrieval from unstructured news articles [15].

While these traditional ML and deep learning models have shown promise, they are often limited by the quality and volume of labeled data available for training. The emergence of Large Language Models (LLMs) like GPT-3 and its successors has marked a significant shift in NLP capabilities. LLMs such as OpenAI's GPT-4o can perform a wide range of NLP tasks, including text classification, summarization, and entity extraction, using few-shot learning—a technique that requires only a few examples to achieve high performance. Brown et al. demonstrated that scaling up language models improves their adaptability to new and diverse input data, which is crucial for maritime risk assessment models where real-time data is heterogeneous and constantly evolving [16].

Building on these advancements, this paper presents a novel pipeline that leverages LLMs to automate the collection and categorization of global incident news. Our approach integrates traditional ML models and state-of-the-art LLMs, such as GPT-4o, to evaluate their effectiveness in news classification tasks. We find that LLMs, particularly GPT-4o, consistently outperform traditional models, enhancing both the efficiency and accuracy of maritime risk data collection. The proposed pipeline not only streamlines the

[1]News API Pricing: https://newsapi.ai/plans

2375-9259/24/$31.00 ©2024 IEEE
DOI 10.1109/ICDMW65004.2024.00061

categorization process but also ensures the timely delivery of relevant insights, which is vital for proactive risk management.

This study contributes to the growing body of literature on the application of LLMs in risk management by demonstrating their superiority over traditional models in handling complex, real-world datasets. We also introduce a new evaluation metric, the "Ratio of Valid Categories," to measure the reliability of model outputs in classification tasks. This metric provides deeper insights into the performance of various models, highlighting the importance of model robustness and data quality in automated data collection pipelines.

Our main contributions are:

- Develop a new method for automated maritime risk data collection and identification.
- Propose a new evaluation metric: Ratio of Valid Categories (RVC).
- Analyze and compare the capabilities of the proposed method with traditional ML models and LLMs for maritime risk identification from news.
- Demonstrate the merits of the proposed method with LLMs in maritime risk data collection and identification across different evaluation methods.

II. Related Work

A. Traditional Machine Learning Approaches

Before the emergence of large language models (LLMs), traditional machine learning approaches, such as Support Vector Machines (SVM), Naïve Bayes (NB), and k-Nearest Neighbors (KNN), were widely utilized for text classification tasks [17]. Wang et al. also conducted a comprehensive analysis of various traditional machine learning methods, demonstrating their efficacy in terms of classification accuracy [18], [19].

B. Large Language Models and Their Impact

The advent of large language models (LLMs) has significantly transformed natural language processing methodologies. Brown et al. introduced GPT-3, a 175-billion-parameter autoregressive language model capable of performing various NLP tasks, such as translation, question-answering, and text completion, with strong performance in a few-shot learning setting, where the model is prompted with only a few examples instead of requiring task-specific fine-tuning [16].

C. Expanding Capabilities of LLMs

Further evaluations have demonstrated the expanding capabilities of LLMs across various domains:

- **Affective Computing:** Amin et al. (2023) evaluated the performance of ChatGPT models, including GPT-3.5, on affective computing tasks such as suicide tendency detection, personality assessment, and sentiment analysis [20], confirming the emerging capabilities of ChatGPT in these areas [21]. The study found that while ChatGPT models performed comparably to classical NLP methods like Bag-of-Words and Word2Vec, they still lagged behind fine-tuned language models such as RoBERTa for specific tasks.
- **Text Summarization:** Basyal and Sanghvi (2023) investigated the capabilities of LLMs in text summarization, conducting a comparative study across different LLMs. Their findings demonstrated that models like OpenAI's text-davinci-003 significantly outperformed others in generating concise summaries, as indicated by higher BLEU, ROUGE, and BERT scores [22].
- **Text Classification:** Chae and Davidson (2023) conducted a comprehensive study on the application of LLMs for text classification, from zero-shot learning to fine-tuning approaches [23]. Their research demonstrated how different LLM architectures, including GPT variants, perform in text classification tasks, highlighting their ability to generalize with minimal training examples. The study indicated that while larger models, such as GPT-3 and GPT-4, achieve higher accuracy, smaller models fine-tuned on specific datasets can also perform competitively at a fraction of the cost, making them an attractive option for many research applications. It further discussed the trade-offs between proprietary and open-source models, emphasizing the importance of evaluating models for bias and transparency.
- **Retrieval Augmented Generation (RAG):** Recent studies have explored the use of LLMs in Retrieval Augmented Generation (RAG) and in-context learning tasks. Huang and Wang (2024) evaluated Microsoft's Orca 2 in RAG applications, comparing it with models like GPT-4 and GPT-3.5-Turbo [24]. Their study found that Orca 2 excelled in generating high-quality responses efficiently on consumer-grade GPUs. Similarly, Huang et al. (2024) assessed Llama 2's performance in in-context learning using the MS MARCO dataset [25]. The results showed that Llama-2 models performed comparably to OpenAI's offerings, with Llama-2-13b-chat-hf slightly outperforming GPT-3.5-turbo in answer quality.

III. Dataset

The dataset used in this study, named the *Global Maritime Risk Incident Dataset (GMRID)*, consists of an extensive collection of global maritime risk data, with a primary focus on incidents and disasters. The original dataset contains 5,744 records, each characterized by 52 distinct attributes. A detailed examination revealed that most of these attributes were manually appended after the initial data collection process. This manual enrichment indicates that the core dataset initially consisted of two fundamental attributes: *Category* and *Details*.

Before applying the dataset to classification models, several preprocessing steps were undertaken to enhance the quality of the text data in the *Details* column. These steps included converting text to lowercase, tokenizing, removing punctuation, eliminating stopwords, and lemmatizing words. Such preprocessing was essential to standardize the data and reduce noise, thereby improving model performance.

The dataset includes a *Category* column summarizing the types of disruptions; however, this column was initially unstructured, containing 857 unique labels. To address this issue, the *Category* column was split using a comma delimiter, resulting in 111 unique labels. Recognizing that this number remained too large for effective classification, further consolidation reduced the number of labels to eight primary categories. Table I presents the mapping between the eight primary categories and the initial unstructured categories. This mapping is performed by four human annotators, who are researchers involved in this study.

During our experiments, we observed significant discrepancies between the labels generated through the manual process described above and those produced by the *gpt-4o* model. A manual review of these discrepancies showed a preference for the *gpt-4o* model's labeling in 9 out of 10 cases (refer to Table IV in Section V-A). As a result, we opted to use the *gpt-4o* model with a 5-shot prompting strategy to relabel all entries in the dataset. The revised dataset is referred to as *GMRID v2*, while the original dataset is referred to as *GMRID v1*.

IV. Methodology

A. Overview of the Proposed Method

The proposed method is designed to streamline the risk data collection and identification from news sources for users. This section details the proposed method as an operational pipeline from the perspective of a user using the system.

The proposed method, as illustrated in Fig. 1, automates the summarization, identification, and updating of disruption database based on the URLs submitted by users. The process is designed to be intuitive and user-friendly, allowing users to efficiently collect and analyze the risk data.

As shown in Fig. 1, the interaction begins when a user submits the URL of a disruption event. In the first key step (Step 1), the system extracts the content of the news and checks whether the event already exists in the database. If the event is new, the model will proceed with the following steps: summarizing the news, identifying the risk category, updating the database with the key elements, and presenting a summary of the information to users through visualization. This summary provides users with a quick overview of the news article's content. The prompt used for summarizing by LLM is shown in Listing 1.

Listing 1. Prompt for LLM Summarization

```
Summarize the following article in about 70 words,
focusing on what happened, where it happened, and
the consequences (economic loss, environmental
impact, etc.): {article}
```

Following the summarization (Step 2: Summarize the news using LLMs), the model identifies the risk category (Step 3: Identify the risk category of the news using LLMs). The results from both steps are used in the subsequent step to better organize the data into meaningful segments, enhancing usability for users. For this research, we evaluated the performance of traditional ML models and modern Large Language Models (GPT-4, Llama-3.1). Our comparative analysis focused on key performance metrics, such as accuracy and the ability to discern nuanced distinctions across diverse news topics.

Once the identification (Step 3) is complete, Step 4: Update the database will be conducted. This step involves updating and enriching the database by adding detailed records, including the event's headline, summary, risk category, URL link, and publication date. To further support user-friendly efforts, the model identifies and ranks related news events within the same category based on their recency. This functionality enables users to quickly access the most current and relevant information within their area of interest.

For enhanced risk data presentation, the system is designed to visualize what users get from it (Step 5). The system not only visualizes the summarization of the news and the identified risk category but also generates a summary table that distills the collected data into an easy-to-digest format, presenting key information such as headlines, summaries, categories, and publication dates.

B. Comparative Study of GMRID Identification: Traditional Models vs. LLMs

In this study, we evaluated the performance of both traditional classification models and modern Large Language Models (LLMs) such as GPT-4o and Llama-3.1. Our comparative analysis focused on key performance metrics, including accuracy and the ability to distinguish nuanced differences across diverse news topics.

We implemented five traditional machine learning models—Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors—using the scikit-learn library[2]. Training, evaluation, and hyper-parameter tuning were conducted with both the GMRID v1 and v2 datasets.

In addition to traditional models, we employed OpenAI's GPT-4o and GPT-4o-mini, as well as Meta's Llama-3.1-8B and Llama-3.1-70B models, for classification tasks. The dataset was split into two subsets, with 80% used for training and 20% reserved for testing.

To enable the LLMs to perform classification, we utilized few-shot prompting techniques. The system prompt template used for this classification is provided in Listing 2.

Listing 2. Prompt for LLM Classification

```
Task: You are a classifier. Your objective is to analyze
the given input and assign it to one of the predefined
categories: {categories}. Evaluate the content carefully
and use the defining characteristics of each category to
ensure an accurate classification.

Guidelines:
1. Understand the Categories: Familiarize yourself with the
 specific attributes of each category by referring to the
category descriptions provided in the JSON: {
categories_json}.
2. Contextual Analysis: Consider the broader context of the
 input. If an input could potentially fit into multiple
```

[2]https://scikit-learn.org/

TABLE I

MAPPING OF INITIAL UNSTRUCTURED CATEGORIES TO EIGHT PRIMARY CATEGORIES FOR MARITIME RISK CLASSIFICATION. THIS TABLE ILLUSTRATES THE CONSOLIDATION OF 857 UNIQUE LABELS INTO 8 PRIMARY CATEGORIES, DEMONSTRATING THE PROCESS OF DATA REFINEMENT FOR MORE EFFECTIVE CLASSIFICATION. THE MAPPING WAS PERFORMED BY FOUR HUMAN ANNOTATORS INVOLVED IN THE STUDY.

Primary Categories	**Initial Unstructured Categories**
Weather	Flooding, Severe Winds, Weather Advisory, Tropical Cyclone, Storm, Ice Storm, Earthquake, Tornado, Typhoon, Landslide, Water, Hurricane, Wildfire, Blizzard, Hail
Worker Strike	Mine Workers Strike, Production Halt, Protest, Riot, Port Strike, General Strike, Civil Service Strike, Civil Unrest Advisory, Cargo Transportation Strike, Energy Sector Strike
Administrative Issue	Port Congestion, Police Operations, Roadway Closure, Disruption, Cargo, Industrial Action, Port Disruption, Cargo Disruption, Power Outage, Port Closure, Maritime Advisory, Train Delays, Ground Transportation Advisory, Public Transportation Disruption, Trade Regulation, Customs Regulation, Regulatory Advisory, Industry Directives, Security Advisory, Public Holidays, Customs Delay, Public Health Advisory, Detention, Aviation Advisory, Waterway Closure, Plant Closure, Border Closure, Delay, Industrial Zone Shutdown, Trade Restrictions, Closure, Truck Driving Ban, Insolvency, Environmental Regulations, Postal Disruption, Travel Warning
Human Error	Workplace Accident, Individuals in Focus, Military Operations, Flight Delays, Cancellations, Political Info, Political Event
Cyber Attack	Network Disruption, Ransomware, Data Breach, Phishing
Terrorism	Bombing, Warehouse Theft, Public Safety, Security, Organized Crime, Piracy, Kidnap, Shooting, Robbery, Cargo Theft, Bomb Detonation, Terror Attack, Outbreak Of War, Militant Action
Accident	Hazmat Response, Maritime Accident, Vehicle Accident, Death, Injury, Non-industrial Fire, Chemical Spill, Industrial Fire, Fuel Disruption, Airline Incident, Crash, Explosion, Train Accident, Derailment, Sewage Disruption, Barge Accident, Bridge Collapse, Structure Collapse, Airport Accident, Force Majeure, Telecom Outage
Others	Miscellaneous Events, Miscellaneous Strikes, Outbreak of Disease

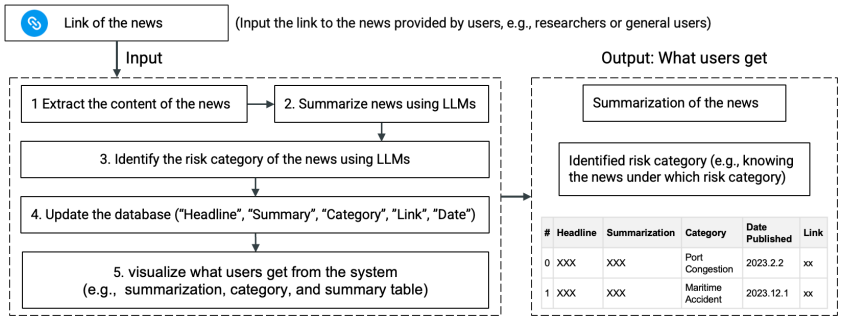

Fig. 1. Schematic Overview of the Proposed Method: Operational Pipeline for Risk Data Collection and Identification from News Sources. This figure illustrates the step-by-step process of the automated risk data collection and identification system, from URL submission by users to the generation of a summary table. Key steps include: 1. Extract the content of the news, 2.Summarize the news using LLMs, 3. Identify the risk category of the news using LLMs, 4. Update the database, and 5. Visualize what users get from the system.

```
categories, select the one that best reflects its primary
intent or focus.
3. Handling Ambiguity: For ambiguous inputs or those that
do not clearly align with any category, choose the category
 that most closely matches the content provided.
4. Ensure Accuracy and Consistency: Strive for consistent
and accurate classifications. Avoid arbitrary or random
assignments.
5. Provide Feedback: If the input cannot be classified into
 any of the given categories, return "Others".
{exemplars}
Output Format: Return only the name of the category that
the input belongs to. If uncertain, respond with "Others".
```

The system prompt template is designed as a versatile tool for text classification. In our study, the placeholders were populated with the following specific data:

- {categories}: A list of primary categories, as shown in Table I.
- {categories_json}: A JSON structure mapping primary categories to initial unstructured categories, as detailed in Table I.
- {exemplars}: For zero-shot classification, this field is left empty; for few-shot classification, exemplars are retrieved from the GMRID training set and formatted as shown in Listing 3.

Listing 3. Exemplars Used in Few-Shot Prompting

```
Example Inputs and Outputs:
- Input: local source reported operation at pier 1 and 2 of
 the container terminal at port of durban was suspended due
 to strong winds ...
- Classification: Weather
...
```

Using the above prompting techniques, the evaluation of GPT-4o and GPT-4o-mini models was conducted via the OpenAI API[3], while the evaluation of Llama-3.1 models was performed on Nvidia L40 GPUs using the HuggingFace Transformers library[4].

C. Evaluation Metrics

To evaluate the performance of each model, we calculated predictive accuracy on the held-out test data using the weighted F1 score. This metric provides an overall measure of the model's performance by taking into account both precision and recall, giving a balanced view of the model's ability to correctly classify data.

[3]https://platform.openai.com/docs/overview
[4]https://huggingface.co/docs/transformers/en/index

During the evaluation process, we observed that the Large Language Models (LLMs) occasionally produced outputs that did not correspond to any valid category labels defined in Table I. Table II presents a sample of classification results produced by the Llama-3.1-8B model, using different numbers of shots (0, 1, 3, 5, and 10).

TABLE II
SAMPLE CLASSIFICATION RESULTS FROM LLAMA-3.1-8B MODEL USING DIFFERENT NUMBERS OF SHOTS. THIS TABLE PRESENTS THE OUTPUTS OF THE LLAMA-3.1-8B MODEL FOR A SINGLE INPUT ACROSS VARYING NUMBERS OF SHOTS (0, 1, 3, 5, AND 10). IT ILLUSTRATES THE MODEL'S TENDENCY TO PRODUCE LENGTHY EXPLANATIONS AND JUSTIFICATIONS, PARTICULARLY WITH FEWER SHOTS, HIGHLIGHTING THE CHALLENGE OF GENERATING CONCISE, VALID CATEGORY LABELS CONSISTENTLY.

Details (Cleaned)	thursday reported operation u.s. air cargo company like fedex ups affected novel coronavirus covid-19 quarantine testing measure imposed official china report chinese official told flight crew would required quarantine ...
Ground Truth	Administrative Issue
0-shot(s) Result	Based on the provided input, the category that best fits the content is 'Administrative Issue'. The input discusses ...
1-shot(s) Result	Based on the provided input, the classification is: Others. *Reasoning*: The input discusses the impact of COVID-19 quarantine measures on air cargo operations in China ...
3-shot(s) Result	Based on the input, the classification is: Administrative Issue.
5-shot(s) Result	Based on the input, the classification is: Administrative Issue.
10-shot(s) Result	Administrative Issue

Based on the results in Table II, we observe that the Llama-3.1-8B model correctly classifies the input as 'Administrative Issue' when given 10 shots. However, in other cases, the model produces lengthy explanations and justifications for its classifications, resulting in outputs that do not directly align with the predefined categories. These outputs are deemed "invalid categories" and cannot be directly used for performance calculations. To address this issue, we developed a post-processing function to extract valid categories from the LLM outputs, ensuring accurate computation of the F1 scores.

We also noted that "invalid categories" were more frequent in the smaller models, which often lack the capability to strictly follow instructions, undermining the reliability of their performance in classification tasks.

To further quantify this issue, we introduced a new metric called the Ratio of Valid Categories (RVC):

$$RVC = \frac{\text{Number of Valid Categories}}{\text{Total Number of Entries}} \tag{1}$$

A "valid category" is any output generated by the model that matches the predefined set of category labels. This metric, RVC, provides insight into each model's reliability in producing valid classifications. A higher ratio indicates greater consistency in assigning entries to valid categories, which is crucial in classification tasks where the accuracy of labels directly affects the model's effectiveness.

V. RESULTS AND DISCUSSIONS

A. Performance Evaluation: GMRID v1 vs. v2

We conducted classification using various machine learning models, including OpenAI's models with few-shot prompting, to determine their optimal performance settings. The best results of this evaluation are presented in Table III.

TABLE III
CLASSIFICATION MODEL EVALUATION - RESULTS COMPARISON BETWEEN GMRID V1 (ANNOTATOR-LABELED GROUND TRUTH) AND GMRID V2 (GPT-4O LABELED GROUND TRUTH). THIS TABLE PRESENTS THE F1 SCORES OF VARIOUS MACHINE LEARNING MODELS, INCLUDING TRADITIONAL APPROACHES AND LLMS, ON BOTH DATASETS. IT DEMONSTRATES THE PERFORMANCE IMPROVEMENT OBSERVED WHEN USING THE GPT-4O LABELED DATASET (V2) COMPARED TO THE HUMAN-LABELED DATASET (V1), SUGGESTING POTENTIAL BENEFITS OF LLM-GENERATED GROUND TRUTH FOR TRAINING AND EVALUATION.

Model	F1 Score (GMRID v1)	F1 Score (GMRID v2)
Naive Bayes	79.93%	87.61%
Logistic Regression	80.39%	**90.74%**
Support Vector Machine	**81.63%**	90.08%
Random Forest	80.73%	87.80%
K-nearest Neighbors	79.81%	83.32%
GPT-4o-mini	62.85% (3-shot)	93.86% (1-shot)
GPT-4o	**65.06%** (5-shot)	**98.76%** (5-shot)

The lower performance of the OpenAI models on the GMRID v1 dataset warrants further investigation. To gain a deeper understanding of this issue, we conducted a manual inspection of the results from the top-performing run: *gpt-4o* with 5-shot prompting. Our analysis revealed that, in most instances where the model's output differed from the human-labeled ground truth, the model's classification appeared to be more accurate. Table IV presents a comparison of 10 such cases, where *gpt-4o* provided a more accurate label in 9 out of 10 instances.

The results in Table III highlight several key observations. Firstly, the v2 dataset, where labels were generated by the *gpt-4o* model, shows improved accuracy for nearly all machine learning models, suggesting that the generated labels are more consistent or better suited to the patterns these models learn.

Furthermore, the OpenAI models, particularly *gpt-4o* with a 5-shot strategy, significantly outperform traditional machine learning models on the v2 dataset, achieving the highest accuracy of 98.76%. This demonstrates the potential of few-shot prompting with advanced models like *gpt-4o* to deliver superior results, particularly when the ground truth is generated by a similar model.

The noticeable performance differences between the v1 and v2 datasets for the same models suggest potential biases or inconsistencies in human-labeled data that may not align with the patterns these models learn. Consequently, using models like *gpt-4o* for generating ground truth labels could provide a more consistent foundation for training and evaluating classification models.

Overall, these findings suggest that integrating advanced language models with few-shot learning strategies can significantly enhance classification performance, especially when the

TABLE IV

COMPARATIVE ANALYSIS OF ORIGINAL HUMAN-ANNOTATED LABELS AND GPT-4O GENERATED LABELS FOR MARITIME RISK CLASSIFICATION. THIS TABLE PRESENTS 10 SAMPLE CASES WHERE THE GPT-4O MODEL'S CLASSIFICATION DIFFERED FROM THE ORIGINAL HUMAN-LABELED GROUND TRUTH. BOLDED ENTRIES DENOTE THE LABEL DEEMED MORE ACCURATE AFTER MANUAL REVIEW, WITH GPT-4O BEING FAVORED IN 9 OUT OF 10 CASES.

Details (Cleaned)	**Original Label**	**GPT-4o Generated Label**
port captaincy cartagena dimar declared second time hotel la américas responsible for undue occupation of beach mangrove area; hotel owner comment.	Others	**Administrative Issue**
government source indicates casualties due to nationwide flooding in kenya. flooding caused by heavy rain impacting horn of africa. government coordinating with kenyan red cross and ngos.	Terrorism	**Weather**
fire in coal harbour impacted an underground homeless encampment. skytrain service disruptions, evacuation due to smoke. fire controlled, ongoing rail disruption.	Administrative Issue	**Accident**
three people killed due to poisoning/suffocation in a ship cabin belonging to new energy engineering company.	Terrorism	**Accident**
china gearing up for chinese new year; factory shutdowns and shipping disruptions expected. higher freight costs anticipated due to increased demand before the holiday.	Worker Strike	**Administrative Issue**
explosion and fire aboard a vessel at charleston harbor marina; three injured. emergency services on site.	Terrorism	**Accident**
tropical storm narda caused port closure in manzanillo, mexico; operations resumed.	Administrative Issue	**Weather**
caledonian sleeper service canceled due to planned industrial action by rmt union members.	Administrative Issue	**Worker Strike**
8 individuals arrested in new jersey for a multi-million-dollar cargo theft ring. stolen goods included various products from warehouses.	**Administrative Issue**	Terrorism
heavy rain and strong winds forecasted to hamper operations at the port of colombo.	Administrative Issue	**Weather**

ground truth data is generated or verified by the same type of models.

B. LLM Classification Performance: Open Source vs. Proprietary Models

We evaluated the performance of the Llama-3.1 models (8B and 70B) using few-shot prompting and compared their performance against the OpenAI models (GPT-4o and GPT-4o-mini) using the GMRID v2 dataset. Table V presents the F1 scores and the ratio of valid categories across different numbers of shots (0, 1, 3, 5, and 10) for each model.

TABLE V

COMPREHENSIVE PERFORMANCE COMPARISON OF LLAMA-3.1 MODELS (8B AND 70B) AND OPENAI MODELS (GPT-4O-MINI AND GPT-4O) ACROSS DIFFERENT FEW-SHOT PROMPTING SCENARIOS. THIS TABLE PRESENTS BOTH F1 SCORES AND RATIO OF VALID CATEGORIES (RVC) FOR EACH MODEL UNDER VARYING NUMBERS OF SHOTS. THE THE BEST PERFORMANCE FOR EACH MODEL HIGHLIGHTED IN BOLD.

Model	**Metric**	**0-shot**	**1-shot**	**3-shot**	**5-shot**	**10-shot**
Llama-3.1-8B	F1	71.51%	**87.08%**	86.52%	86.35%	77.28%
	RVC	49.43%	91.28%	92.94%	92.68%	**99.13%**
Llama-3.1-70B	F1	92.54%	93.20%	93.83%	93.95%	**94.01%**
	RVC	99.56%	99.91%	**100.00%**	**100.00%**	**100.00%**
gpt-4o-mini	F1	92.18%	**93.86%**	91.92%	92.54%	92.97%
	RVC	99.91%	**100.00%**	**100.00%**	99.91%	99.74%
gpt-4o	F1	94.97%	97.42%	97.46%	**98.76%**	97.12%
	RVC	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

The results reveal several key insights:

- **GPT-4o**: Consistently achieves the highest F1 scores across all numbers of shots, with a peak performance of 98.76% at 5-shot. It also maintains a perfect 100% ratio of valid categories across all scenarios, demonstrating the highest level of robustness and reliability among the models tested.
- **Llama-3.1-70B**: Shows competitive performance, particularly in maintaining valid categories. It achieves 100% validity from 3-shot onwards, matching GPT-4o in this aspect. Its F1 scores, while lower than GPT-4o, are consistently above 90%, peaking at 94.01% for 10-shot.
- **GPT-4o-mini**: Performs well overall but shows minor inconsistencies, particularly in the ratio of valid categories at 0-shot (99.91%), 5-shot (99.91%), and 10-shot (99.74%). Its F1 scores are generally lower than GPT-4o but competitive with Llama-3.1-70B.
- **Llama-3.1-8B**: This model exhibits the highest variability in performance. It demonstrates a substantial improvement in the ratio of valid categories, increasing from 49.43% at 0-shot to 99.13% at 10-shot. However, despite the enhancement in F1 scores with additional shots, its performance remains inferior to that of other models. The F1 score peaks at 87.08% for 1-shot, which is notably lower than the performance of most traditional machine learning models presented in Table III. This discrepancy underscores the challenges faced by smaller language models in achieving consistent performance across various few-shot configurations.

These findings highlight the impact of model size and architecture on classification performance in few-shot learning contexts. While the proprietary GPT-4o model maintains superior performance across all metrics, the larger Llama-3.1-70B model demonstrates competitive results, particularly in maintaining valid categories. This suggests that open-source models are closing the gap with proprietary ones, especially in terms of output consistency. The performance variations across different shot numbers also underscore the importance of selecting appropriate few-shot prompts for optimal model performance.

VI. Conclusion and Future Work

This study presents a novel approach for automating the collection and categorization of global maritime disruption data, utilizing the capabilities of Large Language Models (LLMs) like GPT-4o. The findings reveal that LLMs, particularly GPT-4o, significantly outperform traditional machine learning models in terms of both accuracy and consistency, proving to be more effective in classifying complex news data. The introduction of the "Ratio of Valid Categories" metric provided valuable insights into the reliability of each model, with GPT-4o achieving a perfect 100% ratio of valid categories across all scenarios. Although other models, such as GPT-4o-mini and Meta-Llama-3.1, demonstrated some variability, the results underscore the potential of LLMs in enhancing risk data collection and research capabilities. The findings also emphasize the importance of selecting robust models and highlight the role of LLMs in streamlining data-driven risk management.

While the results of this study are promising, several limitations must be acknowledged. The primary limitation is the quality of the available data, which inherently contains ambiguities and inconsistencies due to the diverse nature of news articles. These characteristics pose significant challenges for accurate categorization. Moreover, the use of GPT-4o for relabelling the data may introduce biases, particularly as the model is also evaluated on this dataset. To mitigate these potential biases and to ensure more robust validation, future work will incorporate comprehensive human evaluation. Moving forward, research will focus on improving data quality and refining evaluation methodologies. Exploring more advanced LLM models and fine-tuning them for better classification accuracy will be critical for advancing automated maritime risk management.

Acknowledgment

The authors express their sincere appreciation to the students from Singapore Management University (SMU) for their enthusiastic interest and invaluable contributions to this paper: LU Quanfang, WEI Yifan, YANG Xinyue, and LIAO Jiaxiong. Their dedication has significantly enriched our research. This work was supported by the Natural Science Foundation of Tianjin under Grant 19JCQNJC06000.

References

[1] Z. Wang, R. Goh, X. Yin, P. Loganathan, X. Fu, and S. Lu, "Understanding the effects of natural disasters as risks in supply chain management: A data analytics and visualization approach," in *2nd Annual Workshop on Analytics for Business, Consumer and Social Insights*, 2013.

[2] Z. Wang, V. Joo, C. Tong, X. Xin, and H. C. Chin, "Anomaly detection through enhanced sentiment analysis on social media data," in *2014 IEEE 6th international conference on cloud computing technology and science*. IEEE, 2014, pp. 917–922.

[3] D. Vilares, H. Peng, R. Satapathy, and E. Cambria, "BabelSenticNet: A commonsense reasoning framework for multilingual sentiment analysis," in *IEEE SSCI*, 2018, pp. 1292–1298.

[4] E. Cambria, N. Howard, J. Hsu, and A. Hussain, "Sentic blending: Scalable multimodal fusion for continuous interpretation of semantics and sentics," in *IEEE SSCI*, 2013, pp. 108–117.

[5] E. Cambria, B. Schuller, B. Liu, H. Wang, and C. Havasi, "Statistical approaches to concept-level sentiment analysis," *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 6–9, 2013.

[6] H. T. Nguyen, P. H. Duong, and E. Cambria, "Learning short-text semantic similarity with word embeddings and external knowledge sources," *Knowledge-Based Systems*, vol. 182, no. 104842, 2019.

[7] E. Cambria, *Understanding Natural Language Understanding*. Springer, ISBN 978-3-031-73973-6, 2024.

[8] E. Cambria, R. Mao, M. Chen, Z. Wang, and S.-B. Ho, "Seven pillars for the future of artificial intelligence," *IEEE Intelligent Systems*, vol. 38, no. 6, pp. 62–69, 2023.

[9] Z. Hu, Z. Wang, Y. Wang, and A.-H. Tan, "Msrl-net: A multi-level semantic relation-enhanced learning network for aspect-based sentiment analysis," *Expert Systems with Applications*, vol. 217, p. 119492, 2023.

[10] Q. Liu, S. Han, Y. Li, E. Cambria, and K. Kwok, "PrimeNet: A framework for commonsense knowledge representation and reasoning based on conceptual primitives," *Cognitive Computation*, 2024.

[11] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2021.

[12] Z. Hu, Z. Wang, S.-B. Ho, and A.-H. Tan, "Stock market trend forecasting based on multiple textual features: a deep learning method," in *2021 IEEE 33rd International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 2021, pp. 1002–1007.

[13] A. Teske, R. Falcon, R. Abielmona, and E. Petriu, "Automatic identification of maritime incidents from unstructured articles," in *2018 IEEE Conference on Cognitive and Computational Aspects of Situation Management (CogSIMA)*. IEEE, 2018, pp. 42–48.

[14] 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.

[15] A. Mackenzie, A. Teske, R. Abielmona, and E. Petriu, "Maritime incident information extraction using machine and deep learning techniques," in *2021 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, 2021, pp. 01–06.

[16] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in Neural Information Processing Systems*, vol. 33, 2020.

[17] I. Chaturvedi, Y.-S. Ong, I. Tsang, R. Welsch, and E. Cambria, "Learning word dependencies in text by means of a deep recurrent belief network," *Knowledge-Based Systems*, vol. 108, pp. 144–154, 2016.

[18] Z. Wang, V. J. C. Tong, and H. C. Chin, "Enhancing machine-learning methods for sentiment classification of web data," in *Information Retrieval Technology: 10th Asia Information Retrieval Societies Conference, AIRS 2014, Kuching, Malaysia, December 3-5, 2014. Proceedings 10*. Springer, 2014, pp. 394–405.

[19] Z. Wang, Z. Hu, F. Li, S.-B. Ho, and E. Cambria, "Learning-based stock trending prediction by incorporating technical indicators and social media sentiment," *Cognitive Computation*, vol. 15, no. 3, pp. 1092–1102, 2023.

[20] E. Cambria, X. Zhang, R. Mao, M. Chen, and K. Kwok, "SenticNet 8: Fusing emotion AI and commonsense AI for interpretable, trustworthy, and explainable affective computing," in *Proceedings of the International Conference on Human-Computer Interaction (HCII)*, Washington DC, USA, 2024.

[21] M. M. Amin, E. Cambria, and B. W. Schuller, "Will affective computing emerge from foundation models and general artificial intelligence? a first evaluation of chatgpt," *IEEE Intelligent Systems*, vol. 38, no. 2, pp. 15–23, 2023.

[22] L. Basyal and M. Sanghvi, "Text summarization using large language models: a comparative study of mpt-7b-instruct, falcon-7b-instruct, and openai chat-gpt models," *arXiv preprint arXiv:2310.10449*, 2023.

[23] Y. Chae and T. Davidson, "Large language models for text classification: From zero-shot learning to fine-tuning," *Open Science Foundation*, 2023.

[24] D. Huang and Z. Wang, "Evaluation of orca 2 against other llms for retrieval augmented generation," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2024, pp. 3–19.

[25] D. Huang, Z. Hu, and Z. Wang, "Performance analysis of llama 2 among other llms," in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 1081–1085.