

Analysing Maritime Piracy Reports Using a Text Mining–Based Frequency and Thematic Approach

Üstün ATA^{1*}

¹Bandırma Onyedi Eylül University, Maritime Vocational School. (uatak@bandirma.edu.tr). ORCID: 0000-0002-1513-7371

Article Info

Received: 30.01.2026

Accepted: 20.04.2026

Published: 20.05.2026

Keywords:

Text Mining, Maritime Piracy, Porter Stemming, Maritime Security, Incident Analysis.

Jel Code:

R49

Corresponding Author:

Üstün ATA

Research Article

<https://doi.org/10.5281/zenodo.20297788>

Abstract

Maritime piracy poses various risks to shipping operations and the efficiency of global trade. This study applies a text mining-based analysis to maritime piracy reports published between 2021 and 2023 by the International Maritime Bureau's (IMB) Piracy Reporting Centre to identify potential trends in piracy incidents. The Porter Stemming algorithm was used to preprocess unstructured textual data and extract high-frequency terms related to piracy incidents. Categories were created based on word frequency distributions and semantic similarity. These were grouped into thematic categories representing recurring terms, attack characteristics, locations, operational methods, and incident timelines. The results show that piracy reports primarily emphasize attack locations, threat execution, boarding actions, and the temporal progression of events. Furthermore, relatively less attention is paid to post-incident response measures. The study demonstrates that a purely data-driven text mining approach can effectively reveal structural patterns in maritime piracy reports and support situational awareness without relying on expert opinion or multi-criteria decision-making models.

Introduction

Maritime transport forms the backbone of global trade, offering a cost-effective method of transporting goods while generally producing lower greenhouse gas emissions compared to many alternative modes of transport. However, piracy remains a significant and ongoing threat to ship security, crew safety, and the reliability of global supply chains. With historical roots stretching back to antiquity, piracy continues to disrupt maritime transport routes despite significant technological, regulatory, and security advances. Beyond direct security implications, piracy can also influence operational decisions such as changing routes or opting for alternative modes of transport.

Therefore, a comprehensive understanding of the characteristics and trends of piracy incidents is of paramount importance for developing effective maritime security strategies. Traditional analyses of piracy often rely on statistical summaries, regional comparisons, or case-based assessments. While these approaches offer valuable insights, they are limited in capturing recurring linguistic patterns, thematic structures, and implicit meanings within large volumes of incident narratives. In this context, advances in text mining enable the systematic analysis of unstructured textual data and the extraction of meaningful information from piracy reports.

The Porter Stemming algorithm, also known as Porter Stemmer, is a widely used preprocessing technique in text mining and information retrieval. Developed in 1979 by computer scientist Martin F. Porter, this algorithm reduces English words to their roots by removing suffixes. It reduces the size of the vocabulary and increases computational efficiency in terms of storage, processing time, and retrieval performance. However, the stemming process has some limitations. In particular, Porter Stemmer can encounter the fit problem associated with over-stemming and under-stemming errors. Over-stemming occurs when words with different meanings or different roots are incorrectly reduced to the same root, resulting in a false positive association. In contrast, under-stemming occurs when semantically related words cannot be reduced to a common root leading to a false negative result.

This study is guided by the following research question: To what extent can a Porter Stemming-based text mining approach reveal dominant themes, operational patterns, and emerging trends in piracy incident reports published by the International Maritime Bureau between 2021 and 2023?

The rest of the paper is organised as follows: the theory is presented in Section 2; the methodology of the study is discussed in Section 3; the case study is defined in Section 4; the findings and discussions are presented in Section 5; and the conclusions are described in the final section.

Theory

Maritime piracy attacks adversely affect importer-exporters, shippers, ship owners, and international trade. Moreover, the shipping costs are dramatically increases because of the attacks. The International Maritime Bureau's (IMB) Piracy Reporting Centre stated that there is a significant increase in Q2 2023 events as 25% compared to the previous year report. The increase of the piracy attacks could affect global shipping in the future besides than the lowest attack record is dated in Q1 2023 as per IMB. The annual report of the IMB revealed that 120 incidents of maritime piracy and armed robbery occurred in 2023 compared to 115 incidents in 2022.

In this view, the literature section of the study is aimed to analyse previous research paper that are related to accident or incident reports using text mining techniques. The selected papers are investigated for the idea of finding any trend of previous reports or data.

S. Shi et al. (2019) analysed inland waterway ship collision for 419 collision accidents using text mining. The results revealed that human factors are the primary factors of ship collision. In similar, Zhong et al. (2020) is focused on learning from past accidents. The authors studied a novel framework that mixes deep learning and text mining technologies to analyse hazard records automatically. The results demonstrated that hazard records could be automated by a proper model. Moreover, insightful patterns of hazards could be presented to decision makers using a systematic and data-driven approach. Similarly, Ahadh et al. (2021) is proposed a semi-supervised keyword extraction approach for analysing accident reports. The proposed approach provided a domain specific predictive model that identify the cause of pipeline accidents. Besides, Chintalapudi et al. (2021) collected medical records from Italian Telemedical Maritime Assistance System (TMAS) for three years to analyse using lexicon and Naïve Bayes algorithms. The results indicated that visualization of symptomatic information is completed through word clouds and the sentiment analysis. In the same view, XU et al. (2021) studied text mining approach to identify safety risk factors from accident reports. The authors proposed an information entropy weighted term frequency to extract useful factors from accident reports. The results revealed that the proposed approach performs well to obtain importance degree of factors. In similar, Liu & Yang (2022) analysed historical railway reports to clarify risk factors using text mining. The results showed that potential relationship between hazards, faults, and accidents with proposed approach. It is important to note that, Bai et al. (2021) used a structural topic model (STM) to analyse research trends and themes in the maritime domain. As a text mining-based methodology, the proposed approach revealed that STM could provide more informed decisions. Besides, Tirunagari (2015) applied text mining methods to gather causal relations for maritime investigation reports. The authors employed pattern classification and connectives methods. The causal patterns were compared to human expert opinions. The results showed that text mining methods could be used to extract causal relations from maritime accident reports. Furthermore, He et al. (2023) analysed Piracy and Armed Robbery dataset to analyse the piracy behaviour patterns and characteristics of shipping risk using sentiment analysis, risk analysis, point/kernel density analysis, and word cloud analysis. The results provided that the proposed research could help to strengthen the monitoring of shipping risk.

On the other hand, Papadimitriou et al. (2018) aimed to address the strategic planning of short-haul maritime transport from a maritime cluster perspective and to examine the complementary relationships between these two areas. Paul (2012) studied to automatically classify a corpus of near-miss incidents consisting of 17,846 textual event descriptions. The findings showed that classification accuracy varies from 62% in the detailed event category to 95% in the high-level event category. These results demonstrated that automated text classification offers an effective method for structuring and analyzing large volumes of event data. Zhou et al. (2021) examined the implicit information in the sustainability reports of publicly traded container shipping companies using a hierarchical and unsupervised text mining method. As a result of the research, an integrated sustainability statement framework was developed consisting of three core dimensions: employee training and management, sustainable business management, and sustainable transportation operations.

Text mining studies are used for analysing accident reports and provided valuable insights regarding maritime domain. Moreover, artificial intelligence approach provided possible influential factors regarding insightful patterns of maritime security. Text mining approach could extract insights such as identifying common accident causes, extracting threat levels from incident reports, extracting threat levels from incident reports, prioritizing response strategies.

Method

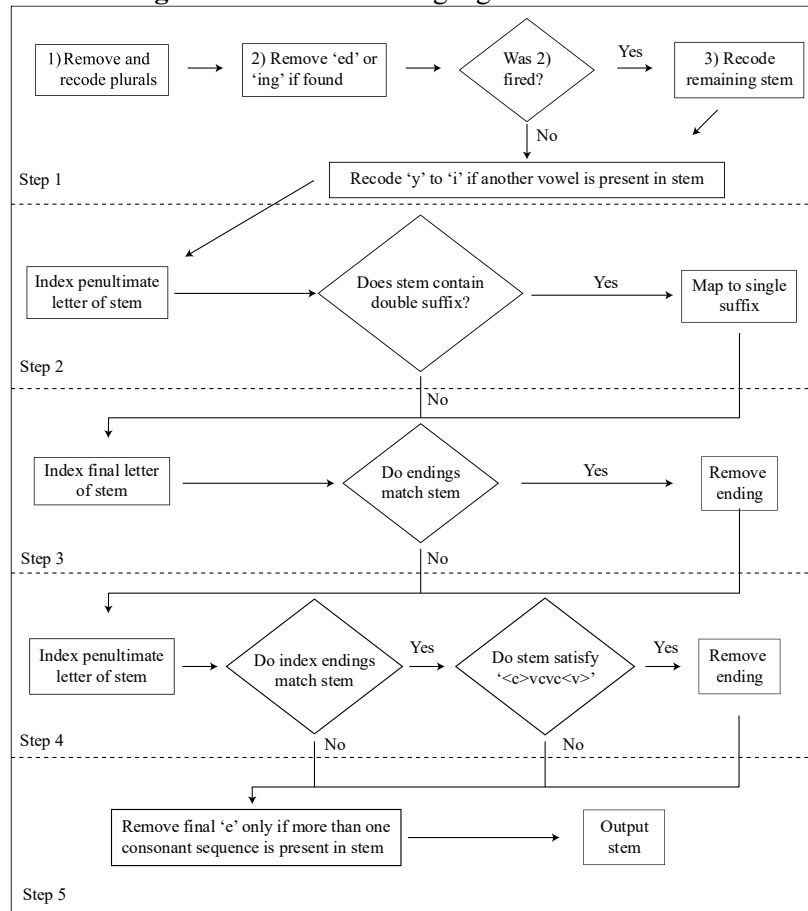
Porter Stemming

The Porter Stemming algorithm (referred to as the Porter Stemmer) is used to extract an English word's stem from its suffixes. The main idea of obtaining its stem is to reduce space and time complexity for Information Retrieval (IR). The algorithm was developed in 1979 by Martin F. Porter who is a Computer Scientist.

The main disadvantage of the porter algorithm is the fitting problem. Over-stemming and under-stemming are mainly two errors in stemming approach. If two words with different stems are stemmed to the same root, over-stemming or a false positive problem arises. On the other hand, under-stemming is not to reduce two words semantically related to the same root which called as a false negative.

The flowchart of the algorithm is as follows (Ali & Ibrahim, 2012):

Figure 1. Porter Stemming algorithm flow chart.



The Porter Stemming algorithm flowchart illustrates a rule-based process that reduces English words from their suffixes to their stem forms. The process begins by examining an input word according to predefined morphological rules. If the relevant suffix meets the necessary conditions, it is removed or transformed. Otherwise, the word is passed to the next processing step. The simplified stem of the word is obtained by progressively processing plural suffixes, tense suffixes, derivation suffixes, and suffixes. In this study, the flowchart represents the text preprocessing stage where word variations in piracy incident reports are standardized and pattern identification processes.

Case Study

The study data is collected from maritime events of 2021, 2022, and 2023 reported in web site of The International Maritime Bureau (IMB) Piracy Reporting Centre. Piracy reports are gathered monthly basis to create a one total year report. For three years, the reports are converted into “txt/pdf” file format to upload Google COLAB IDE.

The reports contain a comprehensive overview of piracy and robbery incident reported against ships from 2021 to 2023. The incidents are divided into three categories that international waters, territorial waters, and

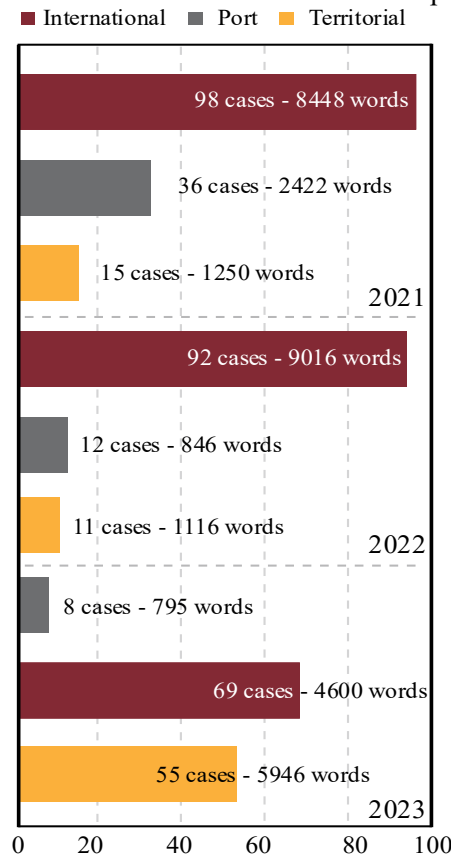
port areas. Moreover, geographic distributions and types of ships are provided. Nature of incidents, response actions, and consequences are other valuable information in piracy and robbery reports.

Total 396 piracy attacks which occurred in international waters, port area, and territorial waters are analyzed with Porter Stemmer for stemming using text mining approach. The reports are divided into three years to reveal any trend of frequencies of pirate attacks. The types of the vessels are ignored not to provide biased results, since the speed or freeboard height prevent vessels from attacks.

The incidents include a wide variety of ship types such as supply ships, tugboats, product tankers, bulk carriers, container ships, chemical tankers, general cargo ships, and fishing boats. Such incidents may occur in many different forms from large-scale hijackings to full-scale hijackings, often involving armed perpetrators. The examples indicate the variety of hazards that different kinds of vessels confront in different seaways around the world.

In 2021, total 98 different pirate attacks (8448 words in reports) occurred in international waters. Similarly, 36 pirate attacks (2422 words in reports) in port area and 15 pirate attacks (1250 words in reports) in territorial area occurred. In 2022, there was slight decrease by 6 percent (92 cases-9016 words in reports) in international area attacks. Furthermore, attacks in port area decreased by 67 percent (12 cases-846 words in reports) and attacks in territorial area decreased by 27 percent (11 cases -1116 words in reports). In similar, attacks in international waters decreased by 91 percent (8 cases -795 words in reports) in 2023. On the contrary, attacks in port area increased by 475 percent (69 cases -4600 words in reports) and attacks in territorial area increased by 400 percent (55 cases -5946 words in reports).

Figure 2. Accident numbers and word count in piracy reports



In the first step of the study, the text mining approach is applied to three years of pirate attack reports from 2021 to 2023 on yearly basis. The porter stemming algorithm is used for each year of pirate attack reports gathered from The International Maritime Bureau's (IMB) Piracy Reporting Centre portal to determine which word is used most by removing affixes from words. By removing unnecessary suffix or affixes, the algorithm could find meaningful insights regarding differences between year basis attacks. The stemming text mining approach is used to eliminate unnecessary information that jeopardize the result of the study. In this view, porter text mining algorithm is used to analyze piracy and robbery reports using default algorithm settings not to reveal biased results. The reason for selecting porter stemming algorithm is considered superior to other stemming algorithms in the scope of effectiveness, simplicity, efficiency, compactness, language independence, and customizability.

Top 100 words of text mining approach (and more than 25) are selected to analyze with experts' choices. As seen in Table 1, text mining results provided insightful results regarding piracy attacks with words as "alarm, inform, robber, perpetr, incid, escap, rais, duti, search, crew, board, boat, singapor, anchor, report, muster, etc."

Table 1. Text mining results.

	2021	2022	2023		2021	2022	2023
alarm	103	112	112	robber	86	104	61
alert	35	39	27	perpetr	137	93	150
inform	63	102	87	stolen	76	61	65
sight	85	89	74	incid	104	126	127
unauthoris	24	34	25	knife	0	35	31
investig	27	23	31	arm	62	46	48
escap	77	88	59	crew	300	238	242
rais	78	82	72	master	86	106	127
duti	91	74	60	person	73	76	62
search	90	67	79	offic	28	24	24
boat	65	110	64	onboard	52	23	37
barg	31	73	28	underway	77	73	59
tug	0	46	0	board	159	160	171
small	26	34	0	navig	36	40	36
coast	59	74	90	singapor	97	146	161
port	50	57	45	indonesian	24	29	36
anchor	79	53	71	republ	25	29	28
guard	66	77	90	broadcast	29	37	42
author	63	45	54	report	152	162	142
police	38	63	65	notifi	41	44	53
secur	51	40	52				
maritim	30	30	29				
safe	75	34	36	conduct	56	32	51
muster	55	60	57	initi	27	36	37
safeti	37	32	0	there	31	23	25
investig	27	23	31	share	0	25	36
while	67	47	66				
upon	36	32	37				
immedi	0	24	30				
further	31	29	46				
hr	0	0	38				

Findings and Discussions

A total of 396 incidents of piracy and armed robbery across international, port, and territorial waters were examined using a text mining methodology. Data were sourced from the International Maritime Bureau (IMB) Piracy Reporting Centre and categorised into three annual subsets to evaluate temporal developments.

The data reveal a significant shift in the geography of piracy. In 2021, activity was concentrated in international waters (98 incidents). By 2022, while international incidents saw a marginal 6% decline, port and territorial water incidents fell sharply. However, 2023 marked a dramatic pivot: attacks in international waters plummeted by 91%, while port-based and territorial water incidents surged by 475% and 400%, respectively. This suggests that perpetrators have modified their operational strategies, moving away from the high seas toward nearshore environments.

The Porter stemming algorithm was utilised to standardise the data, extracting the 100 most frequent stems. While core terms such as "crew", "board", and "robber" remained constant, specific yearly variations emerged:

- **Human Factor:** The persistence of "crew" and "master" underscores the continued focus on human targets.
- **Weaponry:** A notable increase in stems such as "knife" and "arm" after 2021 suggests either an escalation in violence or more granular reporting.
- **Location Focus:** In 2023, stems like "singapor", "anchor", and "port" trended upwards, mirroring the statistical shift toward anchorage zones.
- **Narrative Complexity:** Despite fewer high-seas attacks in 2023, the word count for port-side reports increased, indicating more detailed descriptions of complex encounters or multi-party response actions.

Overall, the findings indicate a redistribution of maritime security risk from international sea corridors to port and nearshore areas. This trend may be related to increased naval presence and surveillance activities in the open sea, as such security measures can direct perpetrators to less protected and more complex areas in terms of jurisdiction. Given that previous studies have mostly focused on geographical concentration areas or incident frequencies, this study reveals both spatial and narrative shifts in piracy and armed robbery incidents by evaluating statistical trends together with text mining outputs. In this respect, the study shows that modern piracy is becoming increasingly opportunistic and perpetrators can exploit lack of coordination between crews, port authorities, coastal security units, and particularly targeting vessels anchored at sea. The use of a data-driven text mining framework allows for the objective and scalable identification of latent patterns in maritime threats.

The findings suggest that maritime security policies should prioritize port area surveillance, monitoring of anchorages, and strengthening coordination between ship crews and port authorities. In this context, the study contributes to the existing literature by demonstrating that textual analysis of incident reports can complement traditional quantitative approaches and offer a deeper understanding of the changing nature of maritime security risks.

Conclusion and Recommendations

This research examines global piracy and armed robbery by applying a text mining framework to 396 incident reports spanning 2021 to 2023. By utilising the Porter stemming algorithm, the study systematically evaluates unstructured narrative data from international, port, and territorial waters to uncover latent temporal and spatial trends.

The analysis identifies a significant transition in the geographical distribution of maritime threats. While offshore incidents predominated in 2021 and 2022, a precipitous decline in international water attacks occurred in 2023. This was mirrored by a surge in incidents within port areas and territorial waters.

Linguistic evidence confirms that piracy remains a "crew-centred" phenomenon. Consistently high frequencies of stems related to boarding, onboard presence, and emergency protocols (e.g., alarms and mustering) highlight the human element of these encounters. Furthermore, an increase in weapon-related and location-specific terminology suggests either an escalation in the severity of threats or a shift toward more granular reporting standards.

Methodologically, this study demonstrates that narrative-based analytics can successfully augment traditional statistical counts. By deliberately omitting vessel-specific variables such as design or speed. The framework avoids potential biases, focusing instead on the raw dynamics of the incidents and subsequent response behaviours.

The results shows a shift in maritime security priorities, specifically toward enhanced anchorage surveillance, port-area access control, and crew readiness in jurisdictionally complex coastal zones. Text mining serves as a scalable, data-driven tool that enables stakeholders to detect emerging risk patterns that aggregate statistics might overlook.

Despite the contributions, the study has some limitations:

- Reporting Bias: Reliance on IMB data brings with it a sectoral limitation, particularly in small-scale incidents such as underreporting.
- Root Finding Sensitivity: Although the Porter stemmer method provides consistency, it can sometimes lead to the loss of semantic nuances by combining words with different contextual meanings under the same root.
- Scope and Time Interval: Excluding ship-specific technical data and limiting the analyses to a three-year time interval restricts the profiling of individual ship vulnerabilities, the identification of long-term cyclical trends, and the strengthening of the validity of the findings. Therefore, analyses covering longer-term data (5–10 years) would increase the validity and robustness of the findings.

Future work could consider integrating Advanced Natural Language Processing (NLP), such as transformer-based models or topic modelling, to capture deeper semantic layers. Additionally, cross-referencing text mining outputs with AIS (Automatic Identification System) data or port congestion metrics could provide a multi-dimensional view of the operational environments that facilitate piracy.

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Author Contributions

Author 1: Conceptualization; Methodology; Software; Validation; Formal analysis; Investigation; Resources; Data curation; Writing – Original draft; Writing – Review & Editing; Visualization.

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical approval

Not applicable.

Reference

- Ahadh, A., Binish, G. V., & Srinivasan, R. (2021). Text mining of accident reports using semi-supervised keyword extraction and topic modeling. *Process Safety and Environmental Protection*, 155, 455–465. <https://doi.org/10.1016/j.psep.2021.09.022>
- Ali, N. H., & Ibrahim, N. S. (2012). Porter Stemming Algorithm for Semantic Checking. *Iccit*, June, 253–258.
- Bai, X., Zhang, X., Li, K. X., Zhou, Y., & Yuen, K. F. (2021). Research topics and trends in the maritime transport: A structural topic model. *Transport Policy*, 102(November 2020), 11–24. <https://doi.org/10.1016/j.tranpol.2020.12.013>
- Chintalapudi, N., Battineni, G., Canio, M. Di, Sagaro, G. G., & Amenta, F. (2021). Text mining with sentiment analysis on seafarers' medical documents. *International Journal of Information Management Data Insights*, 1(1), 100005. <https://doi.org/10.1016/j.ijime.2020.100005>
- He, Z., Wang, C., Gao, J., & Xie, Y. (2023). Assessment of global shipping risk caused by maritime piracy. *Heliyon*, 9(10), e20988. <https://doi.org/10.1016/j.heliyon.2023.e20988>
- Liu, C., & Yang, S. (2022). Using text mining to establish knowledge graph from accident/incident reports in risk assessment. *Expert Systems with Applications*, 207(June), 117991. <https://doi.org/10.1016/j.eswa.2022.117991>
- Papadimitriou, S., Lyridis, D. V., Koliouisis, I. G., Tsioumas, V., Sdoukopoulos, E., & Stavroulakis, P. J. (2018). Strategic planning of short sea shipping within maritime clusters. In *The Dynamics of Short Sea Shipping: New Practices and Trends* (pp. 37-59). Cham: Springer International Publishing.
- Paul, K. (2012). Classifying maritime near-miss and injury report using text mining. Lamar University-Beaumont.
- Shi, H., Li, Y., Jiang, Z., & Zhang, J. (2021). Comprehensive power quality evaluation method of microgrid with dynamic weighting based on CRITIC. *Measurement and Control (United Kingdom)*, 54(5–6), 1097–1104. <https://doi.org/10.1177/00202940211016092>
- Tirunagari, S. (2015). Data Mining of Causal Relations from Text: Analysing Maritime Accident Investigation Reports. *ArXiv Preprint ArXiv:1507.02447*. <http://arxiv.org/abs/1507.02447>
- XU, N., MA, L., Liu, Q., WANG, L., & Deng, Y. (2021). An improved text mining approach to extract safety risk factors from construction accident reports. *Safety Science*, 138(June 2020). <https://doi.org/10.1016/j.ssci.2021.105216>
- Zhong, B., Pan, X., Love, P. E. D., Sun, J., & Tao, C. (2020). Hazard analysis: A deep learning and text mining framework for accident prevention. *Advanced Engineering Informatics*, 46(June), 101152. <https://doi.org/10.1016/j.aei.2020.101152>
- Zhou, Y., Wang, X., & Yuen, K. F. (2021). Sustainability disclosure for container shipping: A text-mining approach. *Transport Policy*, 110, 465–477. <https://doi.org/10.1016/j.tranpol.2021.06.020>