MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

MARITIME ACCIDENT ANALYSIS USING TOPIC MODELING APPROACH

CEREN KESMEN

MASTER THESIS
Department of Industrial Engineering

THESIS SUPERVISOR
Asst. Prof. Dr. MERVE ER

CO-ADVISOR
Assoc. Prof. Dr. İLKER AKGÜN

ISTANBUL, 2024

MARMARA UNIVERSITY
INSTITUTE FOR GRADUATE STUDIES
IN PURE AND APPLIED SCIENCES

MARITIME ACCIDENT ANALYSIS USING TOPIC MODELING APPROACH

CEREN KESMEN
(524420023)

MASTER THESIS
Department of Industrial Engineering

THESIS SUPERVISOR
Asst. Prof. Dr. MERVE ER

CO-ADVISOR
Assoc. Prof. Dr. İLKER AKGÜN

ISTANBUL, 2024

ACKNOWLEDGMENTS

In this thesis, topic modeling approach is used to analyze the underlying causal factors of maritime accidents. I hope it will have a contribution to both theory and practice of maritime industry.

Asst. Prof. Dr. Merve ER always motivated me with her supportive role during the entire preparation process of the thesis and never gave up on me. Completion of the thesis would not have been possible without guidance and expertise of Asst. Prof. Dr. Merve ER and Assoc. Prof. Dr. İlker AKGÜN. I would not proceed and come to this point without them. I would like to express my endless respect and gratitude to them.

I thank my mom and father, Semra and Çetin, for their unconditional supports and love, with all my heart.

July, 2024 CEREN KESMEN

TABLE OF CONTENTS

ÖZET

KONU MODELLEME YAKLAŞIMI KULLANARAK DENİZ KAZASI ANALİZİ

Son yıllarda deniz taşımacılığında artan güvenlik kaygıları nedeniyle deniz kazaları literatürü giderek artmaktadır. Denizcilik en eski ulaşım yöntemlerinden biridir. Kullanıldığı yüzyıllar boyunca çok sayıda kaza meydana gelmiştir. Bu kazalar, insan ve deniz yaşamının tehlikeye girmesi, özellikle kimyasal tanker kazaları nedeniyle çevrenin ve ekosistemin bozulması, şirketlerin para kaybı, tedarik zincirlerinde nakliye gecikmeleri gibi istenmeyen sonuçlara yol açmaktadır. Teknik sorunlar, ekipman ve makinelerin arızalanması, çarpışma, karaya oturma, kötü hava koşulları gibi birçok faktörden dolayı kazalar meydana gelebilir. Son istatistikler, kazaların çoğunun insan faktörlerinden kaynaklandığını göstermektedir. Kazalarda farklı faktörlerin rolünü ölçeklendirecek ve ölçecek standart bir yöntem yoktur. Konu modelleme, kaza raporu metinlerinin analiz edilerek kazaların altında yatan faktörlerin ortaya çıkarılması için aday bir yöntemdir. Çok az sayıda araştırmacı konu modellemeyi deniz kazası raporlarına uygulamıştır. Bu çalışma, Birleşik Krallık deniz kazası raporlarına Gizli Dirichlet Algoritmasını (LDA) 2013- 2023 yılları arasında İngiltere'deki deniz kaza raporlarına uygulayarak bu önemli araştırma alanına katkıda bulunmaktadır. 242 kaza raporuna dayanarak dokuz ana konu belirlenmiştir. Sonuçlar kapsamlı bir yaklaşımla değerlendirilmekte ve ardından İnsan Faktörü Analizi Sınıflandırma Sistemi (HFACS) yapısı dikkate alınarak tartışılmaktadır. Böylece önerilen modelin denizcilik sektöründe uygunluğu ve kullanımı güçlendirilmektedir. Konu modelleme yönteminin sonuçları, farklı teknik, çevresel ve insanla ilgili faktörlerin kaza oluşumuna katkısını ölçmek için kullanılabileceğini göstermektedir. Önerilen yöntem, büyük miktarlardaki deniz kazası raporlarının kısa sürede yarı otomatik olarak analiz edilmesi ve sınıflandırılması konusunda başarılı sonuçlar sunmaktadır.

Anahtar kelimeler; Deniz Kazaları, Gizli Dirichlet Dağılımı, Konu Modelleme, Risk Yönetimi

ABSTRACT

MARITIME ACCIDENT ANALYSIS USING TOPIC MODELLING APPROACH

In recent years, maritime accident literature is growing due to the increasing safety concerns in maritime transportation. Maritime is one of the oldest transportation modes. Numerous accidents have occurred over centuries since it has been used. These accidents lead to unwanted consequences such as danger for human or marine life, disruption of the environment and the ecosystem, especially due to chemical tanker accidents, loss of money for corporations, and shipping delays in supply chains. Accidents may occur due to a number of factors, including technical problems, malfunction of equipment and machines, collision, grounding, and bad weather conditions. Recent statistics illustrate that most of the accidents are caused by human factors in the events. There is not a standart method to scale and measure the role of different factors in accidents. Topic modeling is a candidate method area for revealing the underlying factors of accidents by analyzing the accident report texts. Very few researchers have applied topic modeling to maritime accident reports. This study contributes to this important research field by applying the Latent Dirichlet Algorithm (LDA) to UK maritime accident reports between 2013 and 2023. Nine main topics are identified based on 242 accident reports. The results are evaluated with a comprehensive approach and then discussed by considering the Human Factor Analysis Classification System (HFACS) structure. Thus, it will strengthen the proposed model's appropriateness and usage in the maritime industry. Numerical results are used to quantify the contribution of different technical, environmental, and human-related factors to accident occurrence. Results of the topic modeling analysis are also discussed with the help of experts. The proposed model provides promising results in semi-automatically analyzing and categorizing huge amounts of maritime accident reports within minutes.

Keywords; Latent Dirichlet Allocation, Maritime Accidents, Risk Management, Topic Modeling

SYMBOLS

M **:** Number of documents

N **:** Number of words in a given document

α **:** Dirichlet parameter prior on the per-document topic distributions

ß **:** Dirichlet parameter prior on per-topic word distribution

ABBREVIATIONS

ANN : Artificial Neural Network

BoW : Bag of Words

CREAM : Cognitive Reliability and Error Analysis Method

DTM : Document Term Matrix

ECDIS : Electronic Chart Display and Information System

HFACS : Human Factors Analysis Classification System

IMO : International Maritime Organization

LDA : Latent Dirichlet Algorithm

LSA : Latent Semantic Allocation

MAIB : Maritime Investigation Branch

ML : Machine Learning

PFD : Personal Floatation Device

PLB : Personal Locator Beacon

STM : Structural Topic Modeling

SVM : Support Vector Machine

TF- IDF : Term Frequency -Inverse Document Frequency

VHF : Very High Frequency

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

Over the centuries, a lot of maritime accidents have occurred due to many different environmental, technical, and human-related factors. Some resulted in tragic consequences and left an unforgettable mark in history. One of the most known and popular ship crashes was the Titanic, the first luxury cruise ship that struck to an iceberg suddenly. After that accident, 1514 people lost their lives in 1912 (NBC News, 2012). In 1915, Lusitania, which was a giant English cruise ship, was hit by a torpedo launched by a German submarine off the coast of Ireland. The ship sank in 18 minutes. A Japanese ferry boat sank due to Typhoon Marie while crossing the Tsugaru Strait and thousands of people died. A Ro-Ro Estonian ship has sunk in the Baltic Sea and 857 people died in this accident. In 2008, a ship named Princess capsized due to Typhoon Fengshen, which was the fourth greatest typhoon ever, and approximately 700 people were died in this accident. Besides mishap casualties and injuries, Ever Green's blockage in the Suez Canal occurred in 2021 is a demonstration of maritime incidents' adverse side effects, which are shown as the cost (as freight, demurrage, and other related expenses), scarcity in transportation capacity (containers), and delays in punctual delivers (Özkanlisoy & Akkartal, 2021). The most significant ship accident in Turkish history was Ertugrul Fıkrateyni in 1890. 597 sailors died in this accident. Maritime accidents can lead to loss of human life and injuries, environmental damage, delays in customer orders, loss of money, and penalty costs (additional freight, demurrage, insurance, container, etc.). Especially container accidents may cause to severe pollution such as leakage of heavy metal, plastic fiber, fuel, and other toxic and harmful materials (Wan et al., 2022).

In order to reduce crashes and irregularities in the maritime industry, different standards, procedures, and prevention strategies are determined by the International Maritime Organization (IMO), which plays a crucial role in shaping naval safety regulations. IMO has established though measures, standards and a comprehensive regulatory framework to reduce the risk of maritime accidents and their environmental impact (IMO, 2024); e.g., International Convention for the Safety of Life at Sea (SOLAS), International Convention on Standards of Training, Certification and Watchkeeping for Seafarers (STCW), and the International Convention for the Prevention of Pollution from Ships (MARPOL). The International Safety Management (ISM) Code covers MARPOL,

SOLAS, and other international conventions that protect the safety of lives and properties, secure operation management, and the rest of marine habitats. In 2000, IMO issued Resolution A.884(21) to investigate the impact of the human role on fatalities, injuries, and mishap occurrences from a safety management perspective. From this point of view, IMO stimulates proactive prevention with this resolution. The human element consists of people factors, ship factors, external influences and environment, working and living conditions, shore-side management, and organization on board (IMO, A.884(21)). To measure the human role, Human Reliability Analysis (HRA) tools give a bright perspective to evaluate its effects; also, HRA has a supportive role in determining risk indicators. HRA aims to measure human reliability and detect possible poor relations due to human before an accident (Swain, 1990). Shi et al. (2021) listed used HRA methods in the last four decades including CREAM, STAMP, and HFACS, as given in section 2.3.

IMO and Shi et al. (2021) have presented the gap between maritime accident cause factors focused on Human Factor analysis and Classification System (HFACS) and searched by IMO. The motivation behind this analysis in the literature is that most studies are concentrated on either HFACS or IMO perspectives. There is no standard system or scale to assess maritime accidents as an intersection of IMO and HFACS. Also, studies on that area generally consider only one perspective and one type of vessel. IMO and Shi et al. (2021) have highlighted the gap between accident cause, interlinkage, and intervention for maritime accidents. This study aims to fill the gap between accident causes and interlinkages, generating procedures for all vessel types. Current literature and studies primarily work on specific vessel types and problems. Besides that, the studies on accident report analyses have not concentrated on merging Human Factor Analysis and Classification System (HFACS); generally, they identify causes of ship accident factors. Chen et al. (2022) used topic modeling to categorize accident reports according to accident types; they analyzed these factors under three subtitles: human-related, technical failure, and external failure factors. Chen et al. (2022) emphasized the need for a framework and method-based solutions to cover unclear nonlinear relations among accident causes. IMO has referenced their work and declared that human-related factors are critical for clustering. In order to contribute to this promising research area, this thesis aims to develop an HFACS-based semi-automated, unsupervised topic modeling system for maritime mishap reports.

This study focuses on maritime accidents and their triggering factors and attempts to answer the following three research questions to analyze maritime accidents more deeply:

i) What are the main accident factors that are observed in UK maritime accident reports between 2013 and 2023 years?
ii) What is the distribution of contributing factors of accidents? (especially with an emphasis on the accidents caused by HFACS)?
iii) Is there an association between different factors (e.g., ship type, region) and accidents? (Contributory factors)

A topic modeling approach is used to analyze 242 UK maritime accident reports between 2013-2023. Topic modeling is one of the text-mining methods that help to extract information, patterns, topics, keywords, latent associations, and classifications (Blei et al., 2003). The applications of topic modeling include marketing (Amodo et al., 2018), analysis of crash reports (Ahadh et al., 2021), social media (Karami et al., 2020), and research trends (Gupta et al., 2022). These studies employed one of the most popular topic modeling methods, Latent Dirichlet Allocation (LDA). Various topic modeling types exist, but mainly Structural Topic Modeling (STM) (Bai et al., 2021; Chen et al.,2022) and LDA (Shin et al.,2018) methods are utilized for maritime accident analysis in the literature. Accidents are evaluated from various points of view by researchers. For example, Chen et al. (2022) classified the influential impacts on incidents as social, technical, and human-sourced components. In this thesis, LDA approach is used to analyze maritime accident reports, and the results are assessed in the light of the well known HFACs framework.

In the first stage, accident reports are gathered from the UK accident database, and a deep preprocessing step is applied to clean the input text data. Then, LDA approach is performed. Outputs of LDA are evaluated with expert academics via brainstorming. The topics (words) extracted by LDA are evaluated by considering the components of the HFACS framework; e.g., unsafe acts, preconditions for unsafe acts, organizational influences, unsafe supervisions, preconditions for unsafe acts, and external factors.

Accidents may result in serious consequences such as loss of human life. Hence, identifying the source of accident factors, their frequencies, and possiae (domino) effects is vital for taking pre-actions before the accidents. Potential risks, severities, and

occurrences should be analyzed to configure the system properly to avoid unwanted outputs. Preparedness significantly affects the occurrence of accidents. Besides well-prepared pre-actions, on-time actions, evacuation plans and emergency management systems yield decrease in accident probabilities and consequences.

The remainder of the thesis is structured as follows: Section 2 presents recent maritime accident statistics and previous research in the maritime accident literature to provide a background for the study. Section 3 explains the methodology in detail. Section 4 gives the results, and finally Section 5 concludes the thesis.

2. LITERATURE

This section includes four subsections to provide a background for both the research problem and methodology part: i) maritime accident statistics, ii) academic literature on maritime accidents, iii) role of human in accidents, and iv) text mining and topic modeling.

2.1. Maritime Accident Statistics

This section summarizes maritime accident statistics based on the Marine Accident Investigation Branch (MAIB)'s Marine Accident Recommendations and Statistics 2022 report (MAIB, 2022). In 2022, the frequency of marine incidents was higher than other accident severity types for commercial vessels in the UK region. The occurrence rate of less severe accidents is higher in UK merchant ships with a gross tonnage under 100. Among the 80 vessels, 74 less serious incidents have been observed, which is twice the number of ships with a gross tonnage over 100. UK-registered less serious fishing vessel accidents were observed more than less serious non-UK commercial vessels. Also, the frequency of marine incidents in non-UK commercial vessels is higher than in UK fishing ships.

Between 2013 and 2022, there were only two merchant ship losses for above 100 gross tonnages in 2022 (which does not cover ro-ro passengers, offshore supply, and research ships). The frequency of collisions was 20, and it was the highest rate among other casualty events. The most frequent events can be listed respectively as in the following: collision, machinery, grounding, contact, fire/explosion, flooding and foundering. According to vessel types, casualties occurred in the ascendant to descendant order as follows: passenger, service, solid cargo, liquid cargo, and commercial recreational. These values are illustrated in Table 2.1 in more detail. After 2019, accidents did not result in fatalities. Additionally, the number of injured crew members has been decreasing over the past decade; in 2022, the number was 85. Table 2.2 explains the number of crew injuries and fatalities. Casualties and injuries were seen most frequently in accommodation places for ship crew. Besides that, upper and lower limb injuries were observed more frequently than other injuries. The number of injured and deceased passengers has increased within the last three years.

Table 2.1. Merchant vessels in casualties by nature of casualty and vessel category in 2022 (MAIB, 2022)

Casualty Event	Solid Cargo Ship	Liquid Cargo Ship	Passenger Ship	Service Ship	Commercial Recreational	Total
Collision	2	1	6	10	1	20
Contact	1	0	4	0	0	5
Fire/explosion	0	1	3	1	0	5
Flooding/foundering	0	0	1	1	0	2
Grounding	2	0	3	4	0	9
Machinery	4	0	4	3	0	11
Total	9	2	21	19	1	52

Table 2.2. Deaths and injuries to merchant vessel crew — 2013-2022 (MAIB, 2022)

Year	Number of Crew Injured	Of Which Resulted in Death
2013	134	1
2014	142	-
2015	141	2
2016	133	2
2017	153	0
2018	114	0
2019	105	3
2020	78	0
2021	74	0
2022	85	0

In 2022, accidents for UK merchant vessels under 100 gross tons were seen in mostly search and rescue (SAR) craft, sail craft, and other service ships. It is illustrated in Table 2.3. Grounding, collision, and machinery-related events occurred more than capsizing, flooding, fire, or contact. Other service ships, power crafts, sail crafts had higher injury numbers than sail craft, passenger ship, and other recreational crafts. On the other hand, the death rate for each passenger ship, power craft, and sail craft was 2, and there was not any life loss in other ship types.

Table 2.3. Merchant vessels < 100gt by nature of casualty and vessel category in 2022 (MAIB, 2022)

Casualty Event	Passenger Ship	Recreational craft | Power	Recreational craft | Sail	Recreational craft | Other	Service Ship | Search and	Serviceship|other	Total
Capsizing/listing	1	0	1	1	2	1	**6**
Collision	4	6	3	1	10	7	**31**
Contact	0	0	0	0	1	1	**1**
Fire/explosion	0	0	0	0	2	2	**2**
Flooding/foundering	0	0	0	2	0	0	**2**
Grounding	1	3	18	0	3	3	**43**
Machinery	2	3	1	0	5	5	**15**
Total per vessel type	**8**	**12**	**23**	**4**	**34**	**19**	**100**
Deaths	2	2	2	0	0	0	**6**
Injuries	2	9	4	-	7	10	**32**

During 2013-2022, total fishing losses showed a decreasing trend. Most grounding, collision, and machinery events are attributed to casualties. The death and injury rates change over the depending on the length of the fishing ships. In 2022, the crew fatality rate was lower than the previous year; besides, the number of injuries in the last decade fluctuated in manner but was nearly the same as in the previous decade. It is presented in Figure 2.1. However, the fatality rate had a more unstable and precise fluctuated manner when compared with the death rate in the last decade.

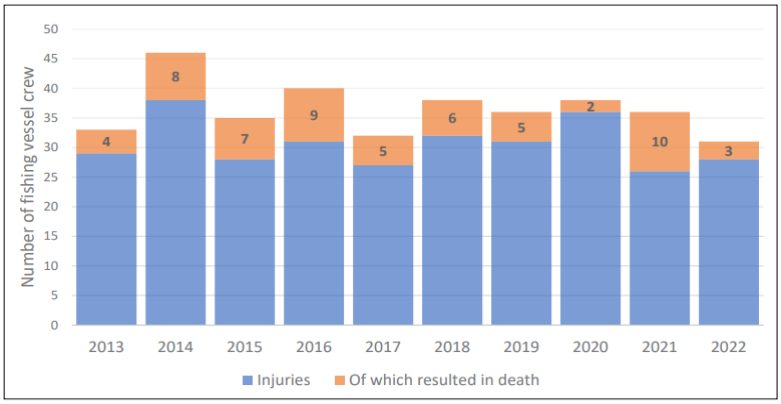


Figure 2.1. Deaths and injuries to fishing vessel crew by year — 2013-2022 (MAIB, 2022)

In 2022, among all non-commercial UK vessels exhibited the highest number of casualties compared to solid cargo vessels and other ship types. In the same year, the casualty frequency rate in ascending order according to event type is as follows: grounding 16, machinery 14, collision 8, contact 8, hull failure 2, and fire/explosion 1, as presented in Table 2.4. The highest injury rates occurred in passenger and solid cargo vessels. One fatality was seen in cargo and passenger ship accidents.

Table 2.4. All non-UK commercial vessels in UK waters — by vessel type and by nature of casualty in 2022 (MAIB, 2022)

Casualty Event	Passenger Ship	Recreational craft | Power	Recreational craft | Sail	Recreational craft | Other	Service Ship | Search and	Serviceship|other	Total
Capsizing/listing	1	0	1	1	2	1	**6**
Collision	4	6	3	1	10	7	**31**
Contact	0	0	0	0	1	1	**1**
Fire/explosion	0	0	0	0	2	2	**2**
Flooding/foundering	0	0	0	2	0	0	**2**
Grounding	1	3	18	0	3	3	**43**
Machinery	2	3	1	0	5	5	**15**
Total per vessel type	**8**	**12**	**23**	**4**	**34**	**19**	**100**
Deaths	2	2	2	0	0	0	**6**
Injuries	2	9	4	-	7	10	**32**

2.2. Academic Literature on Maritime Accidents

The number of research in maritime literature is increasing due to the growing risk concerns in maritime transportation. Especially, most of the studies focus on identifying the sources of accidents and developing proactive risk management strategies. Maritime accidents can occur due to several factors including environmental factors, social factors, human error, and technical errors. On the other hand, the type of accidents (e.g., grounding, capsizing, sinking, collision, etc.) and ship types are also associated with accidents. Some of the studies inspected in the literature review are given in Table 2.5.

Table 2.5. Overview of Maritime Accident Literature

Author	Used Method	Application Area
Eliopoulou et al. (2023)	Statistical analysis	Accident reports
Kim and Lim (2022)	Machine learning methods	Accident prediction
Wang & Fu (2022)	Dynamic Bayesian Network	Seafarers' impact on collision mishap
Yu et al (2023)	Text mining & Bayesian Network	Accident reports
Hwang & Youn (2022)	Text analytics	Collision mishap reports
Ma et al. (2024)	Dematel	Accident reports for collision case
Pilatis et al. (2024)	Statistical anlaysis	Various mishap event types
Rawson & Brio (2023)	Machine learning	Ais data and risk assessment
Kandel & Broud (2024)	Machine learning	Arctic mishap prediction
Cao et al. (2023)	Tree Augmented Naive Bayesian	Impacts determination for accident severity
Du et al.(2023)	Non-linear spatial multi-criteria decision method	Specific region for navigational accidents
Kasyk et al. (2024)	Ishikawa diagram	Human failures for navigational accidents

Several factors can lead to collisions. Wang & Fu (2022) state that low line of sight, unsafe acts of human, problems in very high-frequency communication, and bottleneck problems in maneuvering lead to it. Among 5560 accidents, the likelihood of the most frequent mishaps is found higher for collision, grounding, and fire accidents (Liao et al., 2023). Among the other event types, hull/machinery damage had an outstanding place, with a high seen frequency rate for passenger ships (Eliopoulou et al., 2023).

Vessels such as containers, general cargo, and chemicals have zero rates of accident fatality (Liao et al., 2023). A statistical analysis conducted for passenger ships by Eliopoulou et al. (2023) showed that the number of serious vessel mishaps decreased in the last decade when compared with the previous decade. The main problem leading to mishap occurrence is mostly sourced with organizational deficiencies or failures related to the organization, such as ship design, overburden, and rotational (Eliopoulou et al.,

2023). Besides historical statistical analysis, the use of this data to forecast future mishaps is vital to take preventive precautions. Kim and Lim (2022) used ML methods to account for variables such as ship properties, geographic information, weather, fishery information, number of mishaps by reasons, and mishap information. They created a system that anticipates mishaps and supervises the emergent process.

Cause analysis is essential to identify potential reasons for mishaps. It plays a vital role in making predictions and taking preventive actions. In the literature, various methods exist to extract the causes and make analyses, such as text mining (Yu et al., 2023; Hwang & Youn, 2022; Yan et al., 2023; Dominguez-Péry et al., 2023), association rule mining (Ma et al., 2024; Çakır et al., 2021; Changhai & Shenping, 2021; Özaydın et al., 2022; Lan et al., 2023; Lan & Ma, 2024; Wang et al., 2021), decision making/multi-criteria decision making (Shi et al., 2024; Du et al., 2023), Bayesian network (Fu et al., 2023; Cao et al., 2024) and statistical analysis (Eliopoulou et al., 2023; Pilatis et al., 2024). These analyses focus on various perspectives such as human factors (Ma et al., 2024; Lan & Ma, 2024), collisions (Yu et al., 2023), tugboat accident severity (Çakır et al., 2021), etc. Yu et al. (2023) evaluated causes concerning human, environment, ship, and management perspectives based on analysis using feature extraction from the reports (for collision accident base). On the other hand, Yan et al. (2023) approached maritime accident analysis holistically and developed a semi-automated system that employs different multi-topic models: LDA and BERT. According to their results, the primary three causes of accidents are low visibility, negligent outlook, and lack of a continuous supervision system (Yan et al., 2023). Some researchers employed a holistic analysis and gathered comprehensive results that demonstrate the most common mishaps occurring in different situations such as collisions, hull damage, and oil spills (Wang et al., 2021; Changhai & Shenping, 2019; Özaydın et al., 2021). According to Pilatis et al. (2024), the primary causes of collision and grounding mishaps are breaking rules, and navigation device breakdowns. Shi et al. (2024) used complex networks and DEMATEL and identified the primary causes of collision crashes as follows: improper ship handling, neglecting observations, deprivation of qualified crew, and misjudgments.

According to Rawson and Brito (2023), in the literature, more than 60% of maritime risk assessment studies are based on accident probability. They also stated that the top five most utilized ML methods were logistic regression, ANN, SVM, random forest, and

Bayesian Network. Bayesian Network is used for causal and scenario-based analyses (Zhou et al., 2024; Ma et al., 2024; Yu et al., 2023). Zhou et al. (2024) analyzed mishap data-based scenarios based on risk influencing factors, environment, ship, and severity type. In collision incidents, geographic zone and vessel type are the most impacted risk indicators (Antão et al., 2023). Lan et al. (2023) utilized random forest to forecast the mishap asperity and its impact on collisions. Liu et al. (2022) used Long Short-Term Memory (LSTM) method to predict ship trajectories.

Due to navigational risks in the Arctic region, several studies exist to determine potential mishaps and risk indicators (Kandel & Baroud, 2024; Fu et al., 2023; Yang et al., 2023). The probability of collision mishaps is high; in hot weather, grounding incidents are seen more frequently (Kandel & Baroud, 2024). Fu et al. (2023) stated that grounding crashes also occur due to ice collisions with the vessels. Unreliable speed and unreliable situations have the highest effects on mishaps occurring in this zone (Fu et al., 2023). Recently produced, and longer ships have the highest potential for ship-based deficiency (Kandel & Baroud, 2024). Incident impact factors such as gross tonnage, mishap type, geographic position, vessel type, and engine power have an effect on incident asperity level (Cao et al., 2023). Oil leakages in the engine room can lead to fire (or even worse thermal radiation), which can have catastrophic fatal effects depending on factors such as flame height, burning rate, and flame front (Liu et al., 2024).

Pilatis et al. (2024) stated that the main cause of accidents is human. The psychological situation of seafarers is one of the contributing human factors in mishaps and it is also classified under preconditions for unsafe acts in HFACS (Shi et al., 2021). Fan et al. (2023) studied this area and highlighted that the seafarer's distraction level increases with higher oxygen levels, and their experience level have an impact on collision accidents. Unsafe acts such as failure to provide the right sound and signal, deficiencies in the right proactive manner before the collision, and inaccurate emergency responses are critical to the occurrence of mishaps (Lan & Ma,2024). If current processes and their leading reflections are not well applied by the crew, it will also lead to navigational mishaps (Kasyk et al., 2023).

Navigational accidents occur due to different reasons. The reasons related to weather and geographical conditions have the highest impact. Du et al. (2023) identified these

decencies as “wind speed, wave height, sea fog, precipitation, water depth, and coastline distance”. Smart vessels are alternative preventive measures to avoid navigational accidents and collisions. The highest potential risks for collisions include unreliable velocity and distance, selection of the wrong rotation, failure in maneuvering, failure in emergency command, and early intervention in hazardous conditions (Zhang et al., 2024). Also, the possibility of human error depending on actions decreases (Zhang et al., 2024). Pan et al. (2024) suggested that real-time AIS data usage gives a proper and more precise perspective while evaluating collision likelihood.

2.3. Role of Human in Accidents

This section evaluates the Swiss Cheese Model, HFACS, and IMO’s approach to analyze the contributions of human factors with a general perspective. The Human Factor Analysis and Classification System (HFACS) is suggested by Shappell and Wiegmann (2000) and comes from the Swiss Cheese Model. It is a helpful method for risk analysis and management. It has four layers, and the first three layers prevent hidden failures. The fourth layer is preventing active failures within the holes. The first layer is concerned with organizational influences, the second represents unsafe supervision, the third layer is for preconditions for unsafe acts, and lastly, unsafe acts. However, IMO’s Classification differs from the Swiss Cheese Model-based HFACS with its five layers: diminished human performance, marine environment, safety administration, management, and mental action (Shi et al., 2021). The intersections between IMO’s classification and HFACS are illustrated in Figure 2.2 by Shi et al. (2021). They have highlighted a classification need to avoid the insufficient definition of causal factors. In 2000, IMO identified direct or indirect factors on human attitude, possible impacts such as people factors, ship factors, working and living conditions, external influence and environment, shore-side management, and organization on board as seen in Figure 2.3. These direct and indirect factors correlated with EMSA’s 2021 report, and its statistics validate IMO’s separation, which is presented in the statistics as shore management, shipboard operation, and external environment. Shi et al.’s (2021) model suggests that the relation between each criterion of IMO and HFACS also has a connection with their inner criterion.

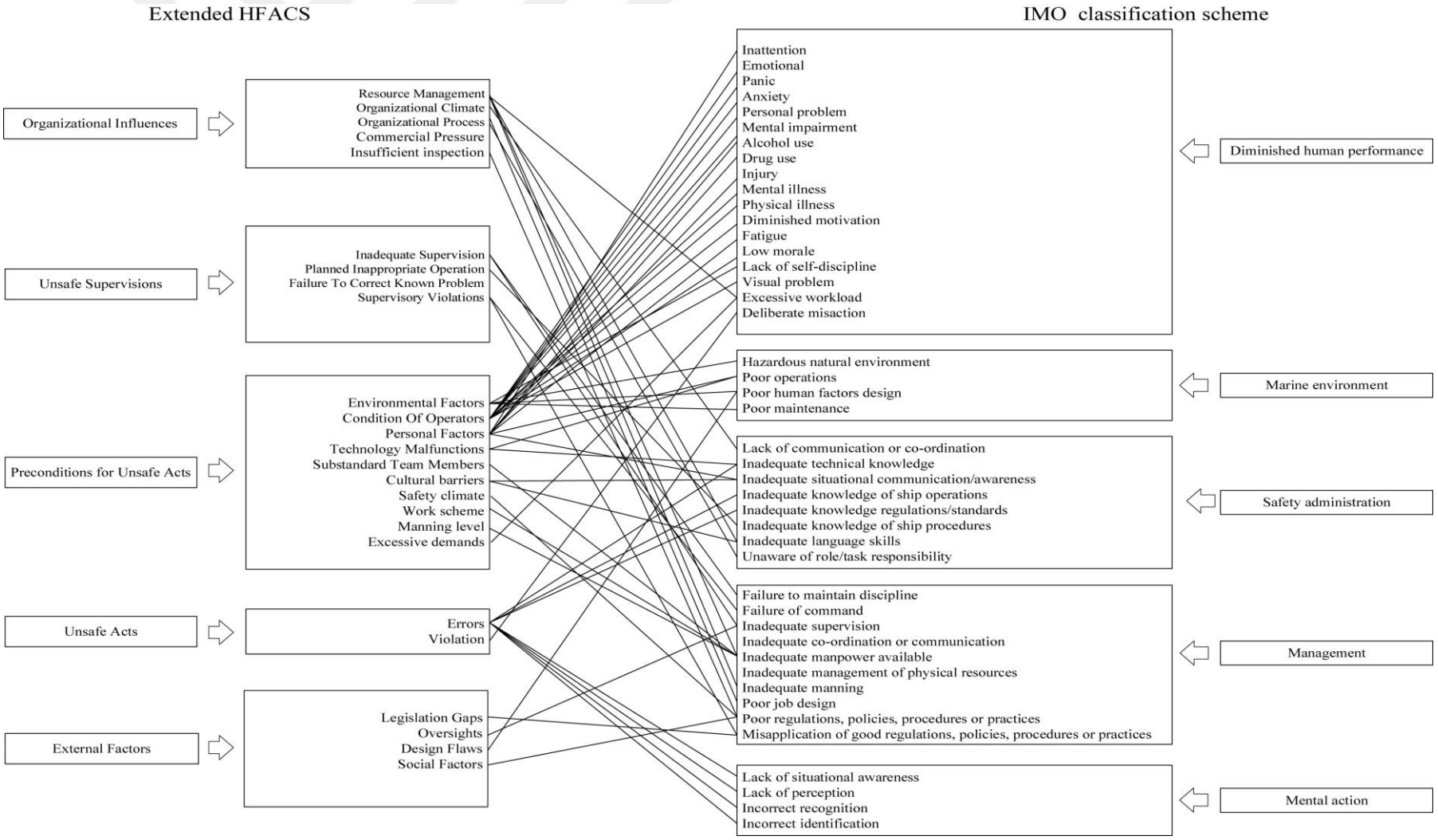


Figure 2.2. HFACS and IMO Human classification criteria (Shi et al., 2021)

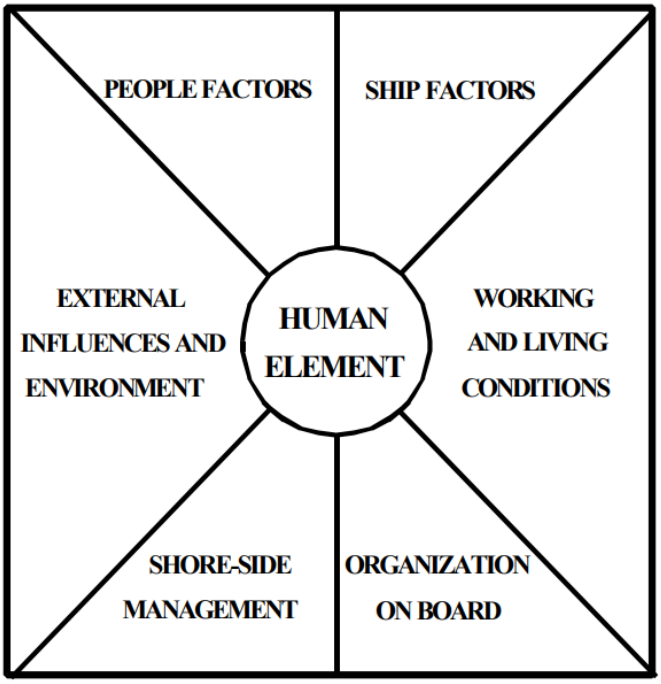


Figure 2.3. Direct and indirect factors (IMO, 2000)

Moreover, mismanagement of organizational influences related to resource management, organizational climate and process, commercial pressure, and insufficient inspection result in failures that are presented in Figure 2.2. These failures could be related to excessive overworking, miscommunication, insufficient procedure application, or lack of knowledge (Shi et al., 2021). Preconditions for unsafe acts are related to nature, technical and maintenance-based errors, conflicts in the crew, and person-based problems (which include cognitive, psychological, and drug addiction-based errors) (Shi et al., 2021). Unsafe acts and external factors are associated with in orderly errors, violations, and social influences as illustrated in Figure 2.2. Original HFACS Framework is illustrated in Appendix A.

Human reliability analysis (HRA) has progressed through three generations (the last phase is currently in progress). It encompasses the examination of human error from both retrospective and prospective standpoints, aiming to enhance comprehension of human error (Shi et al., 2021; Wu et al., 2017). HRA includes techniques such as CREAM (Ung et al., 2018; Wu et al., 2017), STAMP-CAST (Puisa et al., 2018), HFACS (Akyuz, 2017; Celik & Cebi, 2009; Yıldırım et al., 2009), HEART (Akyuz et al., 2016), Acci-Map (Fu et al., 2022), and SHELL (Shi et al., 2021).

2.4. Text Mining and Topic Modeling

Natural Language Processing (NLP) uses methods to extract meanings and sentiments from written or spoken language (Jackson & Moulinier, 2007). The ambiguous nature of NLP needs tools and methods to provide relations and patterns (Jackson & Moulinier, 2007). It is interested in information retrieval, text summarization, topic modeling, and opinion mining (Khurana et al., 2023; Chowdhary, 2020). It is possible to detect misinformation in social media, and the reinforcement learning method is helpful to proceed with it (Shahbazi & Byun, 2021). NLP can analyze a series of documents, and its techniques provide less unclear complexities using text preprocessing functions to get Information Retrieval (IR) using document retrieval (Subhashini et al., 2011). There are four models for IR: probabilistic, Boolean, vector space, and inference network model (Roshdi & Roohparvar, 2015). Numerous information retrieval systems rely on the vector space model (VSM) (Subhashini et al., 2011; Alves et al., 2008; Becker & Kuropka, 20003), and Term-Frequency-Inverse-Document-Frequency (HS & Shenoy, 2020; Meesad, 2021) wherein a document is depicted as a vector composed of index terms. Fake news detection is one of the areas that information retrieval is used (Meesad, 2021). Harisson et al. (2021) applied sentiment analysis and the Latent Dirichlet Algorithm (LDA) to patients' satisfaction and comments about similar drugs. They used machine learning methods such as SVM, ANN, and logistic regression to rank patients' satisfaction.

Text mining and text categorization do not cover the same areas, and categorization is not interested in producing new information from existing topics (Jackson & Moulinier, 2007). Text mining is critical to connecting syntactic and semantic perspectives and understanding and making relations between data and interested areas, such as detecting patterns, keywords, specified topics, or other data characteristics. Topic Modelling is one of the text mining methods used for hidden relations, trends, and categorizations of documents (Blei et al., 2003).

Topic Modeling works on giving an idea about the behavior of data/topic, commenting on documents/reports, and organizing the collection (The School of Informatics at the University of Edinburgh, 2017). It is one of the unsupervised machine learning techniques, and text-mining methods are not required to teach the patterns (Asmussen et

al., 2019). It means that there is no need for trained and worked datasets to start clustering data (Asmussen et al., 2019). Although topic modeling methods are used as unsupervised machine learning methods (Asmussen et al., 2019); sometimes they can be modeled as supervised methods (Mcauliffe & Blei, 2007) and semi-supervised methods (Karami et al., 2020). Unsupervised topic modeling is easier and faster than supervised methods (Asmussen et al., 2019). Figure 2.4 represents the general structure of topic models. Explanatory refers to unsupervised topic models, and the predictive section reflects supervised topic models.

Several types exist in topic models, such as the Latent Dirichlet Algorithm (LDA), Latent Semantic Analysis (LSA), Vector Space Model (VSM) -TFIDF, Correlated Topic Model, Latent Semantic Analysis/Indexing, Pachinko Allocation Model, Structural Topic Model (STM), Sparse Additive Generative (SAGE) topic model, Dynamic Topic Model, Continuous Time Topic Modeling, Self-Aggregating Topic Models, and BERTTopic (Blei et al., 2003; Blei and Lafferty, 2006a; Roberts et al., 2013; Subhashini et al., 2011; Thompson & Mimno, 2020; Roberts et al., 2013).

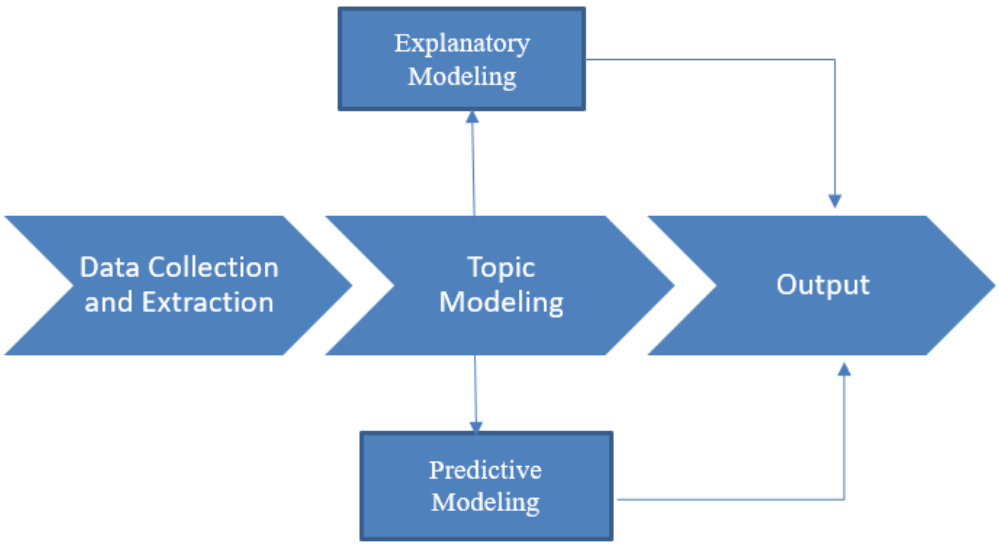


Figure 2.4. Overview of Topic Modeling and Explanatory and Predictive Modeling (adapted from Debortoli et al., 2016)

LSA (also known as Latent Semantic Indexing - LSI) uses the Singular Value Decomposition (SVD) technique due to SVD's adjustable, explicit, and traceable manners to ignore minor patterns and focus on the major concepts (Deerwester et al.,

1990). Latent Dirichlet Allocation (LDA) was introduced by Blei et al. (2003) as a "generative probabilistic model for collection of discrete data". Blei has stated that LDA is discovering themes' posterior inference (The School of Informatics at the University of Edinburgh, 2017). The study aims to enlighten people on making efficient and accurate short decisions with large documents when protecting proper correlations to cluster, detect innovation, summarize, and find relations (Blei et al., 2003).

The Correlated Topic Model (CTM) developed by Blei and Lafferty (2006a) is a more comprehensive method than LDA, and provides correlational relations with extra topics learned. As in LDA, mixture proportions come from Dirichlet, and the difference is made using the logistic normal distribution for conjugate binomial correlations (Blei & Lafferty, 2006a). PAM copes with multinominal complexities, even the restriction of binomial correlations in CTM (Li & McCallum, 2008). Li and McCallum generated the Pachinko Allocation Model (PAM) in 2006. A random and rare correlation of topics are covered, which includes words and inter-topic relations with the other topics. The difference is that multiple topics and their relations exist, in addition to inter-word correlations (Li & McCallum, 2006). PAM topics have relations with any distribution, but in the progress phase, they applied Dirichlet. Huynh-The et al. (2015) compared three different topic models to evaluate traffic behavior evaluation. PAM was selected as the best method to extract infrequent and complicated correlations. Its accuracy was better than LDA, and its performance in the multi-level analysis was better than the Markov clustering topic model (MCTM), which thrives on dual trait classification.

Dynamic Topic Modeling (DTM) is an extension of LDA, which employs the Kalman filter and wavelet regression time series to demonstrate topics' changeability over time (Blei & Lafferty, 2006b). Despite the dynamic topic model, continuous topic models use Brownian motion, and the model provides infrequent deviational inference for quick comparison (Wang et al., 2012). Wang and McCallum (2006) worked on a continuous topic over time model (TOT) that integrated word co-occurrences and temporal localization seamlessly without the Markov model. A self-aggregation-based topic model (SATM) acted as standard topic model behavior in the first phase, developing the document segment and adaptation to Gibbs sampling, which works better on short texts (Quan et al., 2015). Performance comparison is close to LDA and SATM, but pseudo-document numbers have a reverse effect on purity levels. However, SATM has more pure

output than LDA.

SAGE, developed by Eisenstein et al. (2011), considers log-frequency deflection that prevents too much rare word convergence using generative sides. SAGE has a more precise inverse proportion between perplexity (uses Chib-style estimation) and number of topics than LDA. BERTTopic modeling is one of the recent models that employed K-means clustering and vectors from BERT. According to the study of Thompson and Mimno (2020), BERT do not perform outstanding output compared to LDA at the levels associated with large-scale corpus.

All topic modeling methods that are mentioned above cannot deny LDA's generative and comprehensive ways except the continuous models. Vayansky and Kumar (2020) stated that LDA was the most used method. Some of them, such as CTM, DTM, and STM, are utilized as extensions of LDA or Dirichlet, as noted in the studies. Table 2.6 shows some of the topic modeling types used in the literature.

Table 2.6. Overview of Topic Modelling Types in the Literature

Topic Model Type	**Assumption**	**Author & Year**
LDA	Extracts words and frequencies with respect to Dirichlet	Blei (2003)
CTM	Extracts correlated topics from the binomial perspective	Blei and Lafferty (2006a)
LSA/ LSI	Extracts primary concepts ignore minor concepts	Deerwester et al. (1990)
PAM	Extracts rare and complex correlated topics for multinominal perspective	Li and McCallum (2006)
STM	Extracts topics while using a combination of SAGE, CTM, and Dirichlet multinominal regression	Roberts et al. (2013)
SAGE	Extracts topics while avoiding overfitting	Eisenstein et al. (2011)
DTM	Extracts predictive topic model	Blei and Lafferty (2006b)

The most used topic models in the literature are Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Analysis (LDA). The difference between LDA and PLSA is the distribution of the expected topic proportion of the document, which is θ distributes conditionally in LDA. On the other hand, in PLSA, it distributes discretely (Boyd-Graber et al., 2017). The distinction between PLSA and LSA is that LSA does not have any probabilistic behavior. In probabilistic topic models, generally, there are three distribution types which are preferred: Gaussian, Dirichlet, and discrete (Boyd-Graber et al., 2017). Boyd-Graber et al. (2017) have stated that LDA is the most demanded method over PLSA because LDA's adaptations have exceeding features and higher capability of adaptation compared with PLSA.

Benchimol et al. (2022) applied text mining methods using central bank data to extract appropriate texts while cleaning unnecessary texts and finding desired uses with R studio. Amado et al. (2018) used the LDA method to find recent tendencies in "Big Data in Marketing between 2010 and 2015," and a total of 1560 documents were analyzed.

Another usage of topic modeling is the application to mishap reports by Ahadh et al. (2021). They used two methods for determining specific keywords. The method of Yet Another Keyword Extraction (YAKE) is unsupervised and not restricted by tongue and area. Then, a supervised LDA was applied to two-phased semi-supervised key extraction for keyword detection to focus on certain areas. Karami et al. (2020) studied Twitter's widespread subject tendency using LDA. Gupta et al. (2022) used the LDA application via Phyton, stating that the untrained model can behave as a trained model in the output graphs. Some of the general LDA usage areas, type of topic modeling, author, and year information are given in Table 2.7.

Table 2.7. Example LDA Usage Areas in the Literature

LDA Usage Areas	Type of Topic Modeling	Author(s) & Year
Marketing	Unsupervised	Amado et al. (2018)
Crashing reports	Semi-supervised	Ahadh et al. (2021)
Social media	Supervised and Unsupervised	Karami et al. (2020)
Research trend	Unsupervised	Gupta et al. (2022)

Topic model evaluation is critical to understand the number, structure, and rationality of topics. In the literature, the interpretation of topic modeling is realized using different indicators such as quality, perplexity, stability, diversity, efficiency, and flexibility (Abdelrazek et al., 2023). Abdelrazek et al. (2023) stated that if the results are used by human, coherence is a logical option; otherwise, perplexity and flexibility work better. This interpretation for tuning of alpha, beta, and topic parameters is helpful to identify optimal parameter values (Panichella, 2021). The optimization of hyperparameters may be carried out utilizing machine learning techniques. Data can be separated into training and test set, and the trained model's achievement can be evaluated using indicators such as accuracy, precision, and cross-validation (Daud et al., 2023). Also, parameter tuning is important in supervised topic modeling. LDA needs more specific configurations than LSA (Alemayehu and Fang, 2024). Panichella (2021) applied automated tuning for topic modeling utilizing topic coherence, silhouette coefficient, and raw score for stability. In another study, Zhang et al. (2023) worked on different topic models, but they found the optimal topic number using the perplexity score. Perplexity measures how well the model predicts new or unseen data, so low perplexity demonstrates that the model's prediction is reliable and accurate (Lee et al., 2023; Zhang et al., 2023; Bâra & Oprea, 2024). In addition, Bâra and Oprea (2024) stated that perplexity does not adequately indicate an exact reflection of a human decision. Furthermore, coherence and perplexity have an inversely proportional relationship (Abdelrazek et al., 2023). The coherence value decreases as perplexity approaches to the ideal level (Abdelrazek et al., 2023). On the other hand, Lee et al. (2023) and Bâra and Oprea (2024) evaluated topic number quality with coherence and perplexity values. Coherence measures words' degree of relation with each other (Lee et al., 2023; Bâra & Oprea, 2024). Many coherence types exist, such as UCI, UMASS, word embedding-based coherence (Harrando et al., 2021), and CV

(Lamirel et al., 2024). CV coherence is a commonly utilized type. Lamirel et al. (2024) employed CV coherence in their study about philosophy. To do that, they used gensim library in Python. In addition to perplexity and coherence, Gan and Qi (2021) applied multiple evaluation methods like JS divergence, stability, and coincidence to find the optimal topic numbers due to the high volume of documents. To determine the ideal topic number, they created a topic score, including evaluation methods (except coherence), which is their novel approach. In another study, Twitter-based short-text topic modeling was evaluated for perplexity, coherence, stability, and topic diversity values (Mendonça & Figueira, 2024). Bayesian Optimization based on topic investigation and hyperparameter tuning proved that the topic quality and interpretability have relatively low indications and need further developments and observations (Terragni et al., 2024).

3. METHODOLOGY

This section explains the step-by-step procedure applied in the analysis part. It contains data collection, text preprocessing, Term Frequency and Inverse Document Frequency, LDA, and N-Gram models. The flowchart of the methodology is illustrated in Figure 3.1.

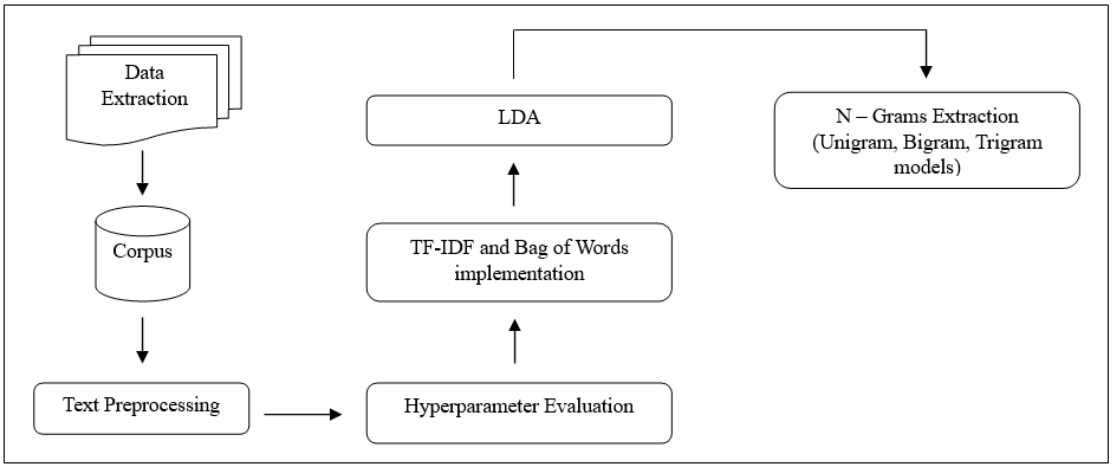


Figure 3.1. Flowchart of methodological application

3.1. Data Collection

Reports issued at MAIB[1] between 2013-2023 are searched, without any restrictions in vessel type (Merchant vessel 100 gross tons or over, Merchant vessel under 100 gross tons, fishing vessel, recreational craft – sail, recreational craft – power). Via excluding the repetitive reports, only the investigation reports, investigation reports on behalf of red ensign group, safety bulletins, complemented preliminary assessment, and overseas reports are taken into consideration. Totally, 242 reports are collected. MAIB-based and redirected reports that are in English are considered. The summary and actions taken are extracted from the online reports and converted to a CSV file.

3.2. Steps of Text Preprocessing

Text preprocessing steps provide instructions for reducing the text size and removing unnecessary characters, punctuation, stop words, white spaces, and lemmatization. In this

[1] Marine Accident Investigation Branch (https://www.gov.uk/maib-reports)

stage, text series are translated into minor units known as *tokens*. Unnecessary items such as punctuation, numbers, or the characters used in the text are removed. All letters are transformed into lower letters with lower casing. Figure 3.2 illustrates the utilized text preprocessing process. Text preprocessing supports dimensionality reduction and normalizing data into vector-based translations as tokens.

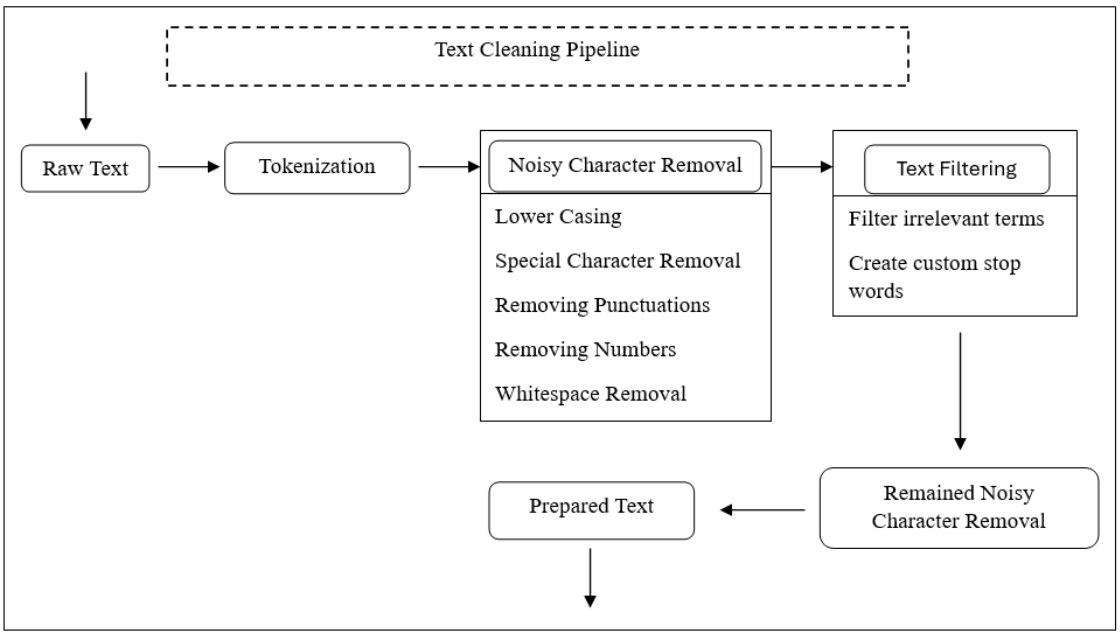


Figure 3.2. Text Preprocessing Process (adapted from Baydogan & Alatas, 2019)

3.3. Toolkits and Libraries used for Topic Modeling

There are various toolkits and libraries for Topic modeling. We use open source and most commonly used libraries in this study. NLTK and regex packages are used in the text preprocessing steps. Gensim and Sci-kit learn packages are used for LDA application. They apply LDA differently. Gensim uses the Bag of Words (BoW) feature, which assumes that each word has an equal frequency. Sci-kit learns behavior-based vector-based TF-IDF or Count Vectorizer act. Table 3.1 provides a summary of the toolkits used in preprocessing and topic modeling stages.

Table 3.1. Toolkits for Topic Modeling and Text Preprocessing

Toolkit	
NLTK	An open-source platform is used for classification, tokenization, stemming, tagging, parsing, and semantic reasoning.
Gensim	Open source platform is used for Word2Vec, FastText, Latent Semantic Indexing (LSI, LSA, LsiModel), Latent Dirichlet Allocation (LDA, LdaModel, LdaMultiCore).
Sci-kit learn	An open-source platform is used for classification, regression, clustering, dimension reduction, preprocessing, and model selection.
pyLDAvis	An open-source R package developed serves as a Python library for creating interactive visualizations of topic models for interpreting topics derived from a corpus of textual data to facilitate the creation of interactive web-based visualizations.
regex	An open-source platform to match text strings such as particular characters, words, or patterns of characters.

3.4. Coherence

Coherence provides the topic similarity evaluation (Zhang et al., 2023). Röder et al. (2015) presented a new coherence metric c_v and stated that it "combines the indirect cosine measure with the NPMI and the Boolean sliding window." Among the other types of topic coherence types, c_v has the best achievement (Röder et al., 2015); therefore, it is used to evaluate topic similarity Also, these coherence measures provide commenting opportunities for Dirichlet's prior parameters (Syed & Spruit, 2018).

3.5. Hyperparameter Tuning and Validation

Blei et al. (2003) introduced LDA and utilized "a Dirichlet prior to simplifying posterior inference.". They apply smoothing via a fuller Bayesian approach. In this way, they produce Ƞ (instead of alpha parameter) as a Dirichlet distribution with a single scalar parameter. Prevalently symmetric prior for setting parameters is employed for protecting

variation of topics (Pramanik & Jana, 2023). On the other hand, asymmetric prior produces better results for repeated and similar topics (Pramanik & Jana, 2023). All possible combinations of prior Dirichlet parameters are inspected (Syed & Spruit, 2018). Wallach et al. (2009) proved that asymmetric prior for Dirichlet parameters positively affects document-topic classification; on the other hand, it negatively affects document-word classifications. Dirichlet prior parameters are: 1) alpha parameter represents document-topic density, and 2) Beta represents topic-word density (Momtazi & Naumann, 2013).

In this thesis, specific different numerical values are assigned for both alpha and Beta parameters. The topic number is also determined from 2 to 10 for each parameter combination. Coherence values are calculated for each alpha-beta combination and topic number. The best combination is selected by assessing the coherence value of the output. Preprocessed data is separated into two groups, such as 100% corpus and 75% corpus. This separation provides validation for produced coherence scores. The flow of the hyperparameter tuning processes is illustrated in Figure 3.3. Additionally, the total number of iterations is equal to 540, which consists of the product of the number of alpha 6, number of beta 5, total count of topic range 9 (from 2 to 10), and total number of corpus sets 2 (75% and 100%).

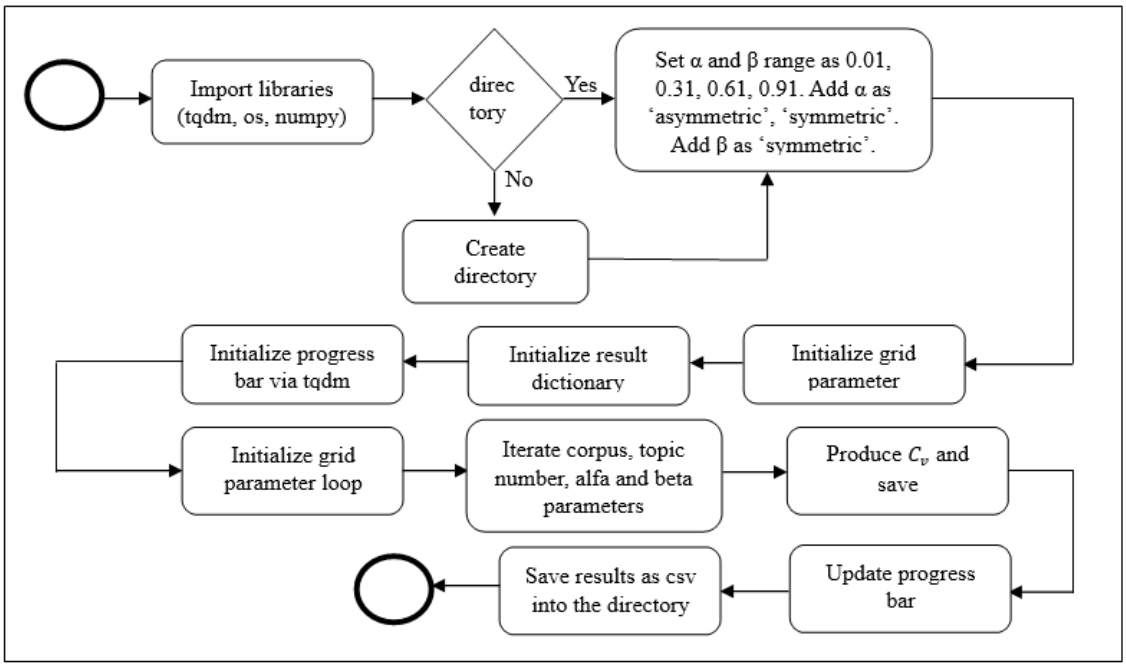


Figure 3.3. Flowchart of hyperparameter tuning (adapted from Kapadia, 2022)

3.6. Feature Extraction with TF-IDF and Bag of Words

Several tools exist for feature extraction in NLP, which aims to extract original data for proceeding by machine learning methods. The methods vary from topic modeling to vector-based models, such as a bag of words and TF-IDF (George, 2022; Semary et al., 2024). In this thesis, TF-IDF and Bag of Words are used for feature extraction.

Bag of Words is an NLP attribute for translating a series of documents into numbers representing text as a vector of frequencies (George, 2022). It checks for the existence of words and, if they exist, takes a snapshot of the count of words with a special arrangement. It is straightforward to apply but assumes that all words are equal. Its magnitude equals the count of occurrences of a word in the document. It ignores the word's order or concept. It is used in the Gensim package.

TF-IDF represents the statistical frequency of words in the document and takes the importance of the words (Latha, 2017). Term frequency means the frequency of the words in the document. Equation 3.1 represents TF-IDF equation closed form. In document *d*, a maximum number of times *t* occurs with *tf* of term, term frequency *t* in document *d* is given by rationing the frequency of occurrence of specific term with maximum occurrence in the corpus. Inverse document frequency *idf*, where the number of documents in term i occurs, n_i is the number of documents in total number of *N* documents, the term will be in a randomly picked document with probability $\frac{n_i}{N}$. Sci-kit learn package includes TF-IDF.

$$tf\ -idf\ (t,d,N) = tf(t,d).idf\ (t,N) \tag{3.1}$$

3.7. Latent Dirichlet Algorithm (LDA)

LDA method is used in topic modeling of maritime accident reports. The purpose of LDA can be categorized into two segments. Firstly, each document includes words that are separated with some topics; then, each topic is devoted to high contingency keywords (The School of Informatics at the University of Edinburgh, 2017).

Secondly, words related to the topic are chosen, and the proportions of sub-words related to topic words are calculated. Then, the words are allocated to the related topics with a proportion.

In terminology, corpora refer to places where documents are collected, and documents are the texts where the words are collected. LDA is a generative method that supports document management and benefits from natural language processing techniques for information retrieval (Blei et al., 2003; Momtazi & Naumann, 2013; Quatrini et al., 2021). Each document is observed from various combinations of independent topics. The benefit of this model is the time saver model, which concludes topics from data and allocates each document a probability of belonging to each topic. It highlights latent sides and detects tendencies over the topic quickly.

LDA acts like the probability of distributions over words or topics, and it has a probabilistic manner. Its terms are:

- A *word* serves as the fundamental element of discrete data, defined as an item drawn from a vocabulary indexed in a set from *1* to V. vth word included in V is given as an exponential equation as the *vth* power of w equals one, and the *uth* power of w equals zero, which means *u* and *v* cannot be equal.
- A *document* is a set of words denoted by *w* and each index is represented with *N* in a sequence of words, and so w_n is the *nth w*ord in sequence.
- A *corpus* is a compilation of documents. The number of words *w* is represented with *M* documents in a set of D documents.

LDA assumes that each word in the Documents *D* such that *N* is distributed with $Poisson(\xi)$, and θ is topic distribution, is distributed with Dirichlet $Dir(\alpha)$ (Blei et al., 2003). Each sequence of words (w_n) are created from various topics (z_n), and these topics (z_n) distributed with multinominal θ over words *w*. To reduce the complexity of the model, it is supposed that *k* refers to the dimensionality of Dirichlet and topic variable z is known and stable. The probability of words parametrized with a k x V matrix, ß is word distribution, is distributed with Dirichlet $Dir(\alpha)$ where $ß_{ij}$ equals to the conditional probability of *j*th exponent of *w* equals to 1 given *i*th exponent of z equals to 1, which means a stable amount is predicted. Supposing $Poisson(\xi)$ does not have a crucial impact on the following analyses, substantial size of the text length allocations is used

when demanded. Moreover, N is uncorrelated with the variables of θ and z. It is an auxiliary variable, and its arbitrary nature is neglected. Marginal allocation of text is seen in Eq. 3.2 (Blei et al., 2003).

$$p(w|\alpha,\beta) = \int p(\theta|\alpha)\left(\prod_{n=1}^{N}\sum_{z_n} p(z_n|\theta)p(w_n|z_n,\beta)\right) d\theta \quad (3.2)$$

At the end, Eq. 3.3 is adapted to the occurrence of individual documents, and probability of corpus will be as in Eq.3.3 (Blei et al., 2003):

$$p(D|\alpha,\beta) = \prod_{d=1}^{M}\int p(\theta_d|\alpha)\left(\prod_{n=1}^{N_d}\sum_{z_{dn}} p(z_n|\theta_d)p(w_{dn}|z_{dn},\beta)\right)d\theta_d \quad (3.3)$$

Lastly, the graphical model of LDA is represented as in Figure 3.4.

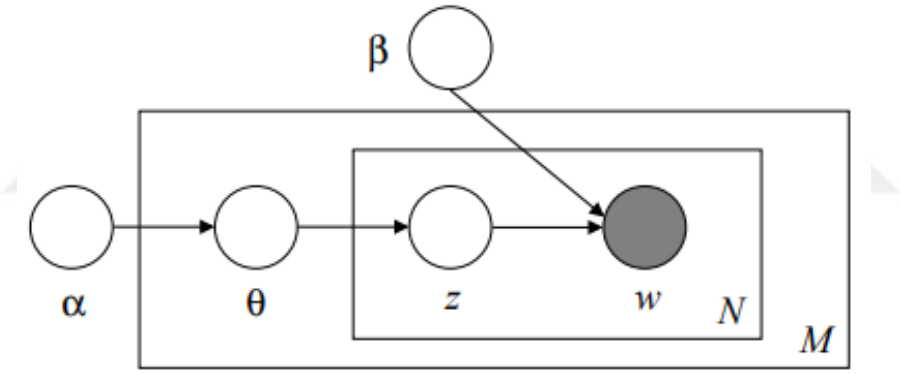


Figure 3.4. LDA Graphic Model (Blei et al, 2003)

3.8. N-Grams

As represented in Figure 3.4, the simplest LDA assigns likelihoods to sentences and sequences of words n-gram. A unigram model shows a single word sequence in N. Bi-gram models present a dual sequence of words, and trigram models illustrate three sequences of words. N-gram models predict the possibility of an n-gram placed in the previous words and allocate the likelihoods to all word sequences (Blei et al., 2003). The possibility of a text is the combination of unigrams as seen in Eq. 3.4.

$$p(w) = \sum_{z} p(z) \sum_{n=1}^{N} p(w_n|z) \quad (3.4)$$

The graphical representation of combination of unigrams is illustrated in Figure 3.5.

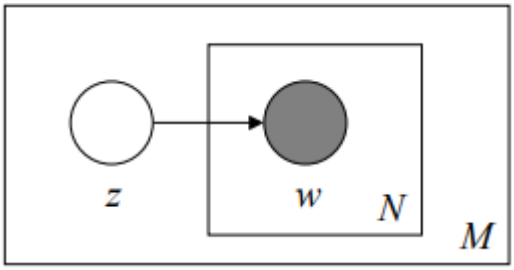


Figure 3.5. Graphic Model for combination of unigrams (Blei et al, 2003).

4. APPLICATION

This section explains seven sub-sections. First section indicates data extraction and text preprocessing processes. Then hyperparameter tuning results are identified. Next, feature extraction application outputs are given. 4.4. explains LDA Scikit-Learn outputs and 4.5 identifies Gensim outputs. Next section identifies unigram, bigram, and trigram results. Lastly, interpretation of the results with experts are covered.

4.1. Data Extraction and Text Preprocessing

A preprocessing stage is applied to the CSV file that includes text data on 242 accident reports between 2013 and 2023. Figure 4.1 represents preprocessing outputs step by step. In the first step, tokenization is applied, and each document in the corpora is separated as tokens, as seen in the figure. With the tokenization, white spaces are also converted to tokens, and extra commas are added. After the lower casing process, scr - single character (letter) and whitespace removal are performed. Successively, punctuation, number, and stop word elimination processes are applied.

	Report_Name	Summary	tokenized	lower	scr	punc	nums	stopwords_removed	special
0	Accident on the stern ramp of the ro-ro freight ferry Seatruck Progress with loss of 1 life	On 15 May 2019, the third officer was struck and fatally injured by a \nfreight vehicle semi-tra...	[On, 15, May, 2019, ,, the, third, officer, was, struck, and, fatally, injured, by, a, freight, ...	[on, 15, may, 2019, ,, the, third, officer, was, struck, and, fatally, injured, by, a, freight, ...	[on, 15, may, 2019, , the, third, officer, was, struck, and, fatally, injured, by, , freight, ve...	[on, 15, may, 2019, the, third, officer, was, struck, and, fatally, injured, by, freight, vehicl...	[on, may, the, third, officer, was, struck, and, fatally, injured, by, freight, vehicle, semi-tr...	[officer, struck, fatally, injured, freight, vehicle, semi-trailer, stern, semi-trailer, ashore,...	[officer, struck, fatally, injured, freight, vehicle, semitrailer, stern, semitrailer, ashore, p...
1	Capsize and full inversion of self□righting keelboat RS Venture Connect sail number 307 with los...	On 12 June 2019, Blackwell Sailing's self-righting RS Venture Connect \nkeelboat sail number 307...	[On, 12, June, 2019, ,, Blackwell, Sailing, ', s, self-righting, RS, Venture, Connect, keelboat,...	[on, 12, june, 2019, ,, blackwell, sailing, ', s, self-righting, rs, venture, connect, keelboat,...	[on, 12, june, 2019, , blackwell, sailing, , , self-righting, rs, venture, connect, keelboat, sa...	[on, 12, june, 2019, blackwell, sailing, self-righting, rs, venture, connect, keelboat, sail, nu...	[on, june, blackwell, sailing, self-righting, rs, venture, connect, keelboat, sail, number, rsvc...	[sailing, self-righting, connect, keelboat, sail, capsized, inverted, sailing, boat, crewed, exp...	[sailing, selfrighting, connect, keelboat, sail, capsized, inverted, sailing, boat, crewed, expe...

Figure 4.1. Output of text preprocessing processes step-by-step

4.2. Hyperparameter Tuning and Validation

Figure 4.2 illustrates ß values and corresponding coherence scores. As it is seen from the figure, 0.91 is the optimal ß value since the highest coherence indicates the best performance. The figure also proves that values belonging to different corpus sets have similar values.

Despite the ß values' consistent behavior, α values have a slightly fluctuated line than ß values. Therefore, the corpus range is selected as 100% and ß values are set to 0.91 to determine the correct α values. A table of the first 20 values of these parameters according to this filter is given in Appendix B. Also, Figure 4.3 presents a graph of alpha values and coherence scores. According to Appendix A and Figure 4.3, α values are asymmetric and have the highest coherence score at 0.91 level.

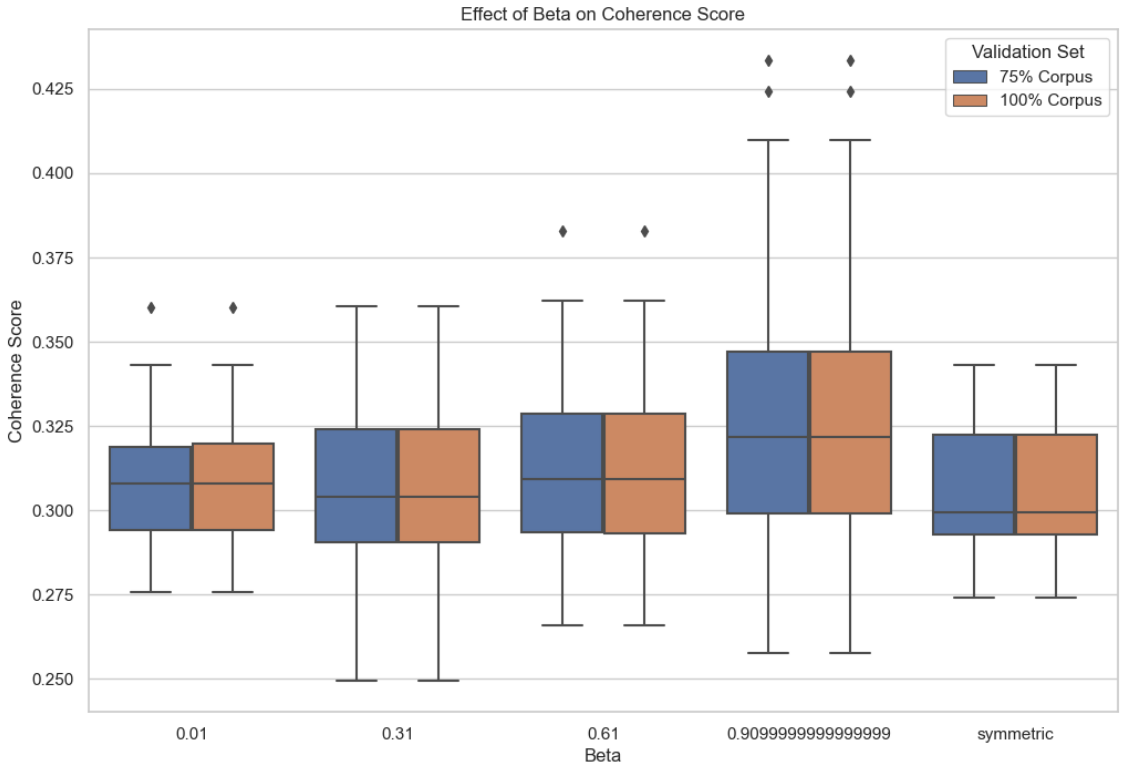


Figure 4.2. Beta – Coherence Scores

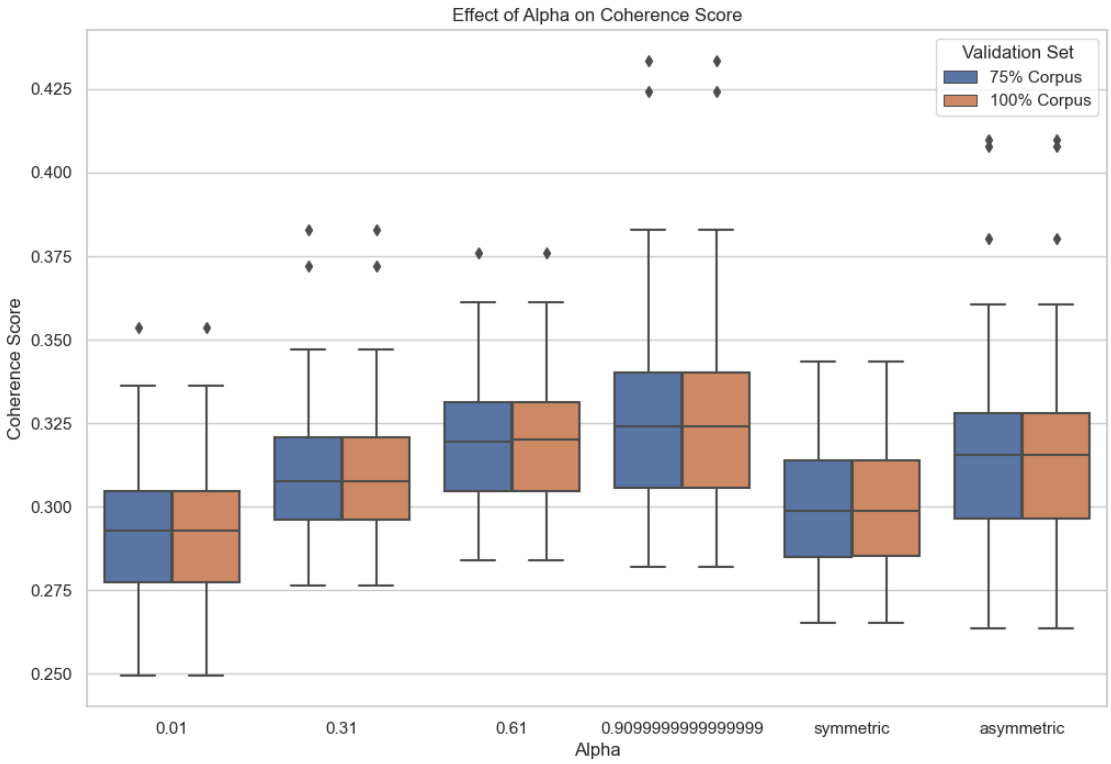


Figure 4.3. Alpha - Coherence Scores

Corresponding optimal values are determined from Appendix B, and are listed in the following (coherence scores below 0.40 are neglected):

- ß = 0.91, α = 0.91, and k = 9 for coherence = 0.43 ,
- ß = 0.91, α = 0.91, and k = 10 for coherence = 0.42,
- ß = 0.91, α = asymmetric, and k = 8 for coherence = 0.41,
- ß = 0.91, α = asymmetric, and k = 7 for coherence = 0.41.

LDA model is applied separately for all of the above optimal parameter value combinations. Resulting topic-word and topic-document classifications are given in Appendix C (for k = 10, 8, and 7). The optimal topic number is chosen as nine because it has the best coherence score. For k=9, topic-word and topic-document classification is presented in Section 4.4.

Figure 4.4 illustrates the coherence scores versus number of topics. The figure shows that the coherence range of topic numbers is narrow for both corpus sets.

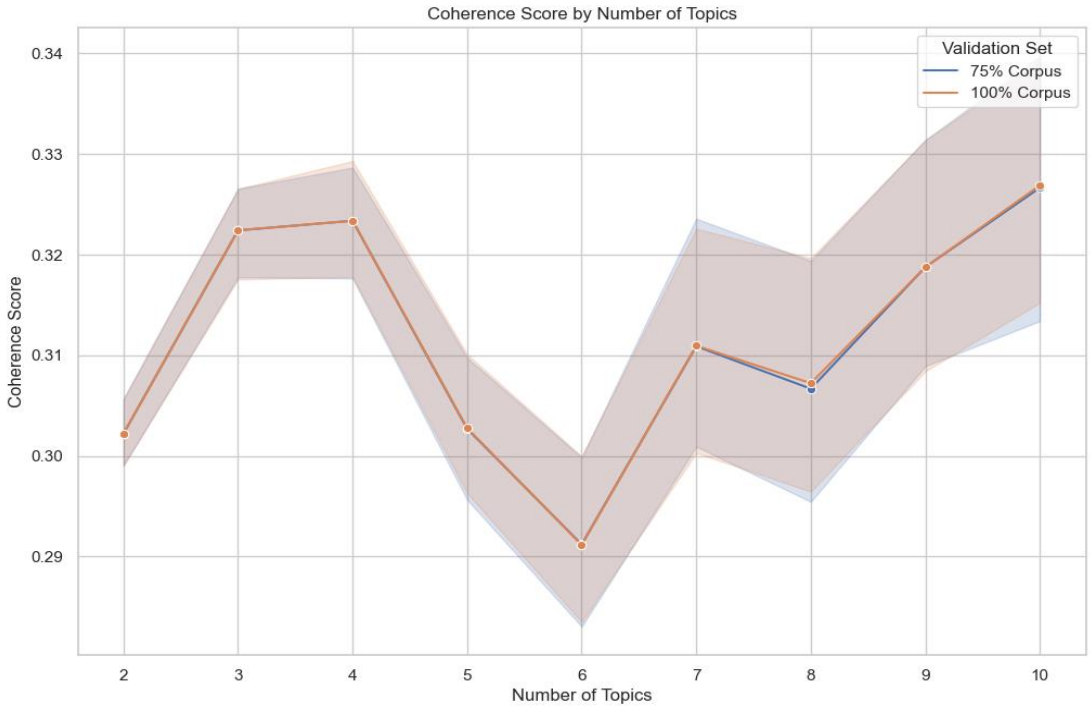


Figure 4.4. Coherence Score by Number of Topics

4.3. Feature Extraction with TF-IDF and Bag of Words

TF—IDF with scikit-learn provides 1066 distinct terms. Table 4.1 represents the most frequently occurring 20 words with their probabilities.

Table 4.1. Top 20 most occurring words with TF-IDF extraction

01. crew (0.058)	11. port (0.030)
02. skipper (0.043)	12. vessel (0.029)
03. boat (0.037)	13. deck (0.028)
04. cargo (0.033)	14. passenger (0.028)
05. safety (0.033)	15. bridge (0.027)
06. fishing (0.032)	16. overboard (0.025)
07. engine (0.031)	17. officer (0.024)
08. ferry (0.031)	18. collision (0.023)
09. board (0.030)	19. deckhand (0.022)
10. crewman (0.030)	20. pilot (0.022)

21 random states are configured to extract word cloud output, as represented in Figure 4.5.

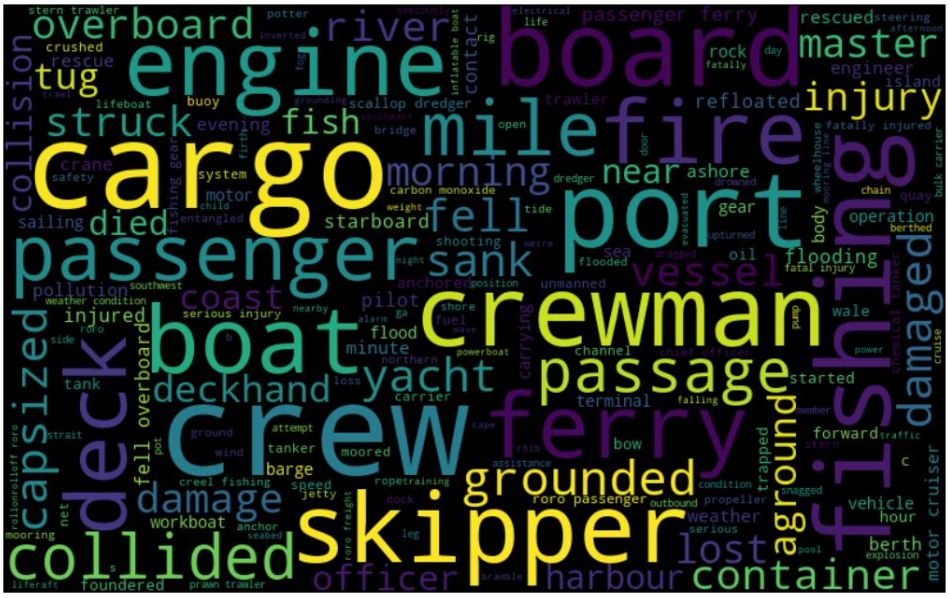


Figure 4.5. Word Cloud

4.4. LDA Sci-kit Learn Outputs

When the topic number is chosen as 9 and number of words are chosen as 15, outputs of the LDA are classified as presented in Table 4.2. Additionally, Table 4.3 briefly categorizes the document and the number of topic allocation outputs generated using the TF-IDF method. The most-seen topic is Topic 4, and the least-seen topics are Topic 2 and 5. The frequency of Topics 0, 2, 3, 5 and 7 are low. Topics 1, 4, 6 and 8 include a high number of reports. The salient words in Topic 0 are “ferry”, “engine”, “guardrail”, and “deck”. Topic 1 includes “collision”, “grounded”, “damaged”, “lookout” and other terms. Topic 4 categorizes fishing and sinking events. These reports also include words such as “wearing”, “capsized”, “safety”, “life raft” which are attributed to sinking accidents. Topic 5 categorizes “ECDIS”, “extinguishing”, “loss”, and “life” words related to death, fire and locational sourced accidents. Topic 6 categorizes “emergency”, “flooding”, “drill”, and “alcohol” sourced accidents. Topic 8 categorizes fire and gas related events and includes terms such as “carbon monoxide, poisoning, exhaust, alarm, gas, etc.”. The noteworthy words in other topics are as follows; Topic 2: “hose, crane, officer, worker, experience, etc.”, Topic 3: “GPS, propulsion, keel, helm, frame, berth, etc.”, Topic 7: “moor, tug, rope, cord, etc.”

Table 4.2. Topics with Sci-kit Learn LDA outputs for topic number 9

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8
fuel	bridge	rope	vehicle	skipper	training	deckhand	engine	carbon
ferry	officer	hose	yacht	crew	ecdis	crew	mooring	monoxide
davit	ferry	crane	gps	fishing	electrical	board	tug	engine
engine	anchor	officer	keel	crewman	rower	safety	line	exhaust
passenger	vessel	amphibious	sailing	boat	hose	emergency	forward	container
gybe	watch	worker	helm	overboard	life	door	injured	cabin
furnace	master	experience	propulsion	wearing	semitrailer	engine	buoy	bilge
buoy	cargo	barrier	berth	personal	display	boat	rope	heater
keel	pilot	shorebased	quayside	sank	pump	chain	team	boat
thermal	collision	quayside	cape	deck	instructor	flooding	tower	passenger
driver	port	deck	carriage	yacht	loss	drill	pool	cargo
guardrail	grounded	rigid	tractor	capsized	rowing	alcohol	trapped	poisoning
deck	passage	fall	frame	safety	extinguishing	scallop	passenger	alarm
vehicle	lookout	vehicle	ec	operation	unsafe	winch	keel	pump
lock	damaged	cargo	truck	liferaft	power	management	cord	gas

Table 4.3. Topics and Document categorization Sci-kit Learn LDA outputs

Topic	Count of Topic
0	6
1	57
2	5
3	7
4	88
5	5
6	37
7	11
8	26
Grand Total	**242**

4.5. LDA Gensim Outputs

With Gensim, a bag of words representation is found, as in Table 4.4. This figure shows only a small representation. The bag of words is too large; therefore, this table shows only a small representation.

Table 4.4. Bag of Words generated with Gensim

0	1	2	3	4	5	6	7
(0, 1)	(1, 1)	(2, 1)	(3, 1)	(4, 1)	(5, 2)	(6, 3)	(7, 1)
(49, 1)	(50, 1)	(51, 1)	(52, 11)	(53, 1)	(54, 1)	(55, 1)	(56, 3)
(53, 2)	(56, 4)	(68, 1)	(75, 1)	(89, 2)	(96, 1)	(97, 1)	(98, 3)
(5, 1)	(17, 1)	(20, 1)	(53, 1)	(56, 3)	(65, 1)	(97, 2)	(99, 3)
(0, 1)	(1, 1)	(35, 1)	(52, 1)	(53, 2)	(58, 2)	(79, 1)	(99, 3)
(52, 2)	(53, 3)	(138, 2)	(183, 1)	(192, 1)	(203, 1)	(205, 1)	(206, 4)
(1, 1)	(5, 1)	(36, 2)	(52, 6)	(75, 1)	(80, 1)	(198, 1)	(206, 9)
(5, 1)	(42, 1)	(148, 1)	(250, 1)	(251, 3)	(252, 1)	(253, 1)	(254, 2)
(0, 1)	(15, 1)	(20, 1)	(25, 1)	(29, 2)	(80, 1)	(97, 2)	(140, 2)
(30, 1)	(52, 9)	(56, 1)	(75, 1)	(97, 1)	(122, 1)	(123, 1)	(172, 2)
(6, 1)	(14, 1)	(31, 1)	(35, 1)	(36, 1)	(56, 1)	(58, 1)	(80, 1)
(29, 1)	(56, 1)	(106, 3)	(127, 1)	(140, 1)	(251, 1)	(315, 1)	(324, 4)
(5, 1)	(14, 1)	(18, 1)	(35, 1)	(36, 1)	(52, 2)	(80, 1)	(97, 1)

Preprocessed tokens are transferred to the dictionary, which has 2159 unique tokens are represented briefly in Table 4.5. A document term matrix is created, which does not include words seen in less than one and no more than 0.8. Also, it removes the most common and rare words, and each document are converted into a bag of words. The bag of words document for the first document is found as: [[(0, 1), (1, 1), (2, 1), (3, 1), (4, 1), (5, 2), (6, 3), (7, 1), ….] and it gives the highest probability for document 7 with 0.98766845 probability.

Table 4.5. Brief representation of 2159 unique tokens

accident
ashore
barrier
berthed
blocking
board
cannabis
communication
considered
contributory
control
death
deck
distraction
fatally
ferry
freight
guidance
injured
landbased
maritime
medium
met
mobile
...

Gensim pyLDAvis outputs give the arbitrary categorization for nine topics and 30 words. Topic 1 is the most prevalent one, and its representation is given in Figure 4.6. It is possible to see that the model categorized as Topic 1 includes “wearing,” “emergency,” “personal,” “flotation,” “alarm”, “device,” which refers to equipment-based accidents.

Topic 2 consists “engine”, “navigation”, “speed”, “fell”, and “alcohol”, mostly covers technical-based accidents except “fell”, and “alcohol”. Topic 3 categorizes the terms related with technic -environmental conditions such as "oil", "fire, "air", "gas", "fuel", "temperature", "heater", "explosion", "furnace", "explosion", "styrene", "monomer", "weather", and "boiler". Topic 4 has terms such as "fire", "engine", "emergency" and "lookout" are related to dominantly technical causes. Topic 5 classifies the terms "carbon", "monoxide", "gas", "exhaust", "poisoning", "electronic", and "speed" covers mostly technical terms related causes.

Topic 6 includes terms such as "wave", "wind", "inverted", "capsize", "propeller" which are dominantly related to environmental causes in mishaps. Topic 7 consists "wind", "fire", "speed", and "loss" which refers to environmental-based accidents. Topic 8 classifies "vhf", "propulsion", "collapse", "collision", and "speed" which is associated with technical causes.

Topic 9 categorizes "engine", "emergency", "unsupervised", "injury", "drowning", "steering", "cargo", "ferry" which is related to technical-based causes. Also, topics (4,5,6,7,8,9) have similar contexts, which means they could include the exact words somewhere, and topics (1,2,3) have distances from each other, but only topics (3) have similarities with topics (4,5,6,7,8,9). Additionally, the estimated term frequency within the selected topics (6,7,8,9) is low compared to topics (1,2,3,4,5). On the other hand, these frequencies have the highest rate for topics (1,2). Topics (3,4,5) have moderate frequency. In Appendix D, related graphs with rest of the topics are given.

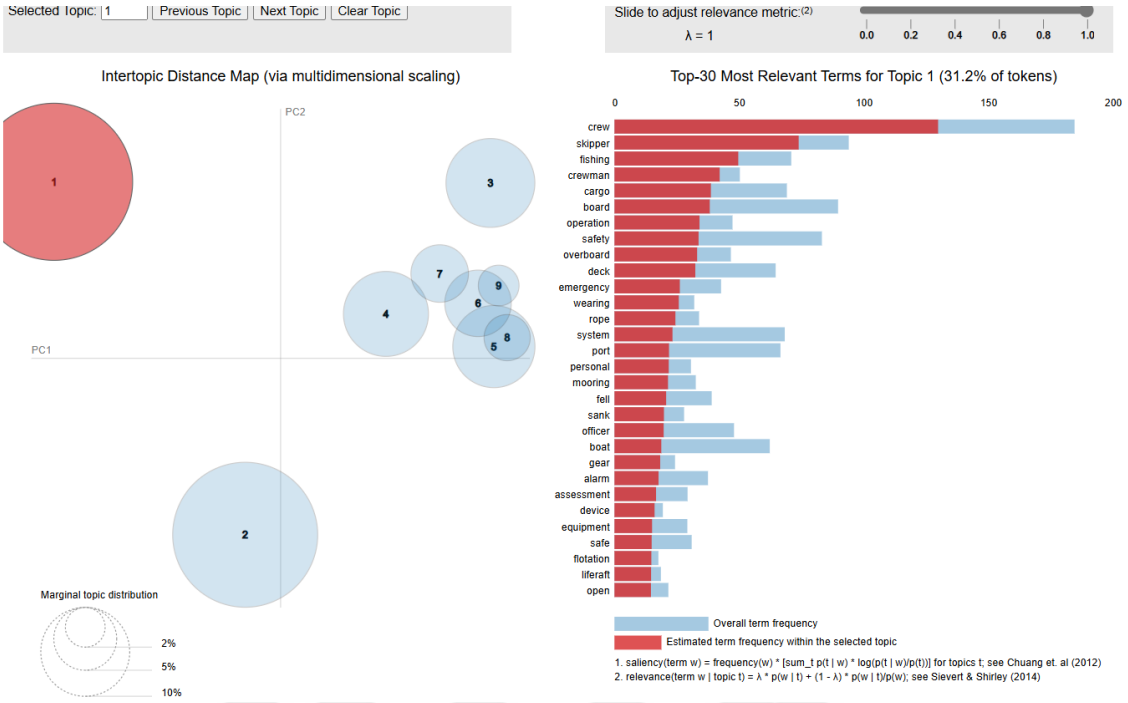


Figure 4.6. pyLDAvis output for topic1

4.6. N-Gram Models

Unigram, bigram, and trigram word representations are given in Figures 4.7, 4.8, and 4.9. Unigram words give the words related to vessels and their specifications. Bigram words have more meaningful results, which give an idea that “carbon monoxide” refers to poisons, “chief officer” and “officer watch” are related to failures, “personal flotation,” “flotation device,” and “wearing personal” highlight situations whether people have a personal flotation device or not. Trigram modeling outputs support the assumptions formed from bigram model results.

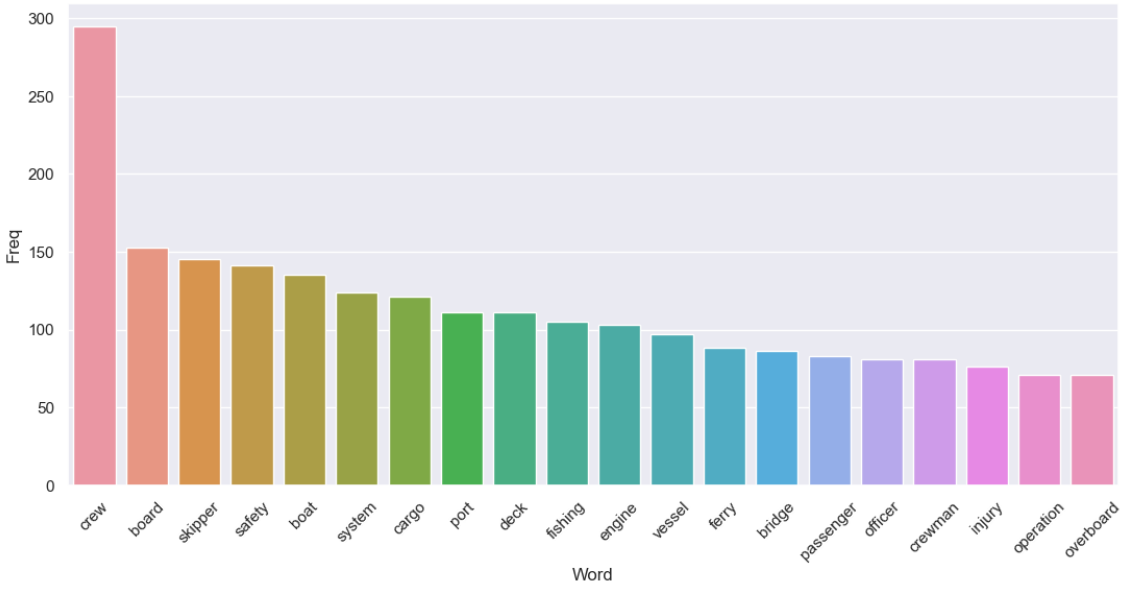


Figure 4.7. Unigram Words

As it is seen from Figure 4.8, the most frequent bigram words are carbon monoxide, safety management, chief officer and management system.

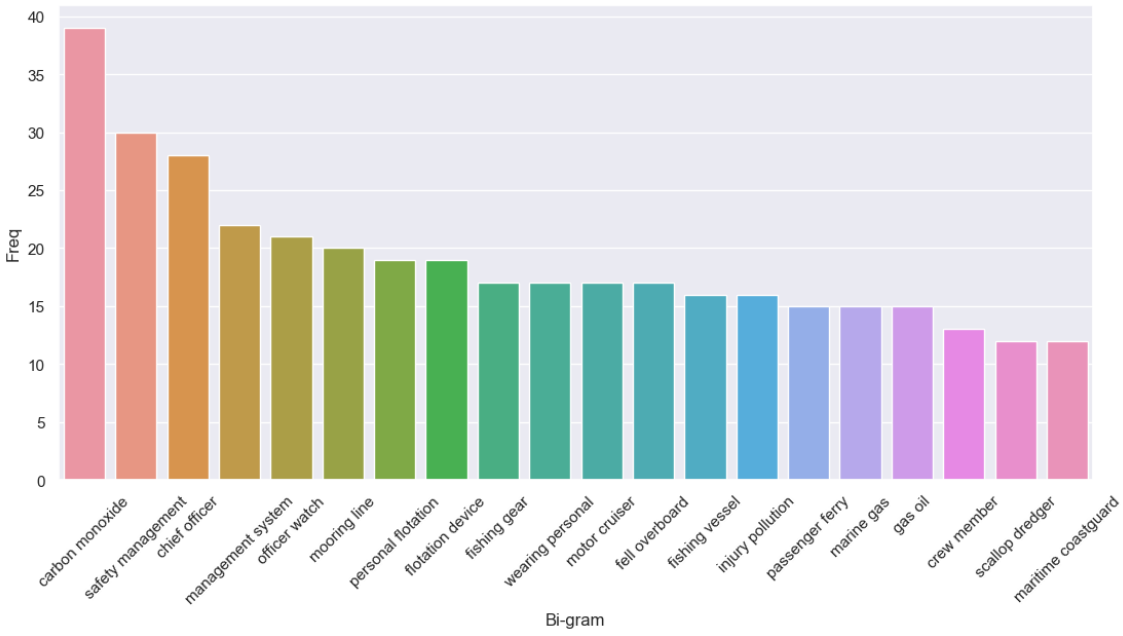


Figure 4.8. Bigram Words

As it is seen from Figure 4.9, the most frequent trigram words are safety management system, personal flotation device, marine gas oil, wearing personal flotation, and carbon monoxide poisoning.

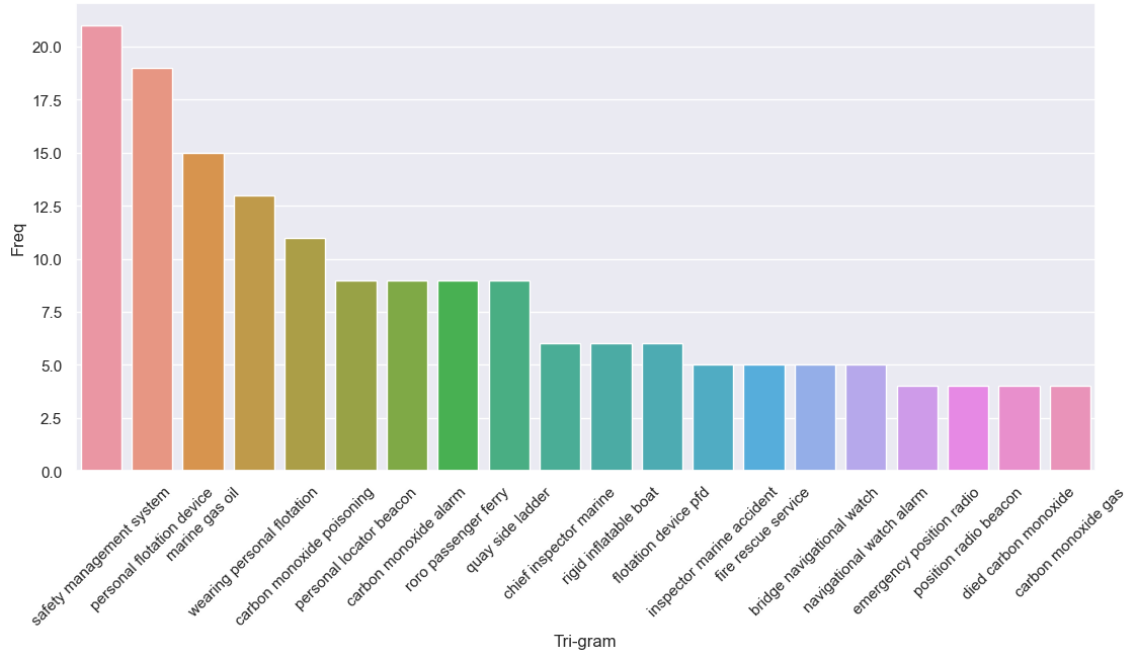


Figure 4.9. Trigram Words

4.7. Interpretation of the Results with Experts

The results of the analysis are interpreted with the help of academic experts who have expertise both in maritime industry and risk management fields. The extracted topic groups and frequency of words are evaluated by considering the HFACs framework. The results provide important insights for classifying accident reasons and preconditions for unsafe acts. Topics 0 and 1 include words related to personal causes such as “gybe” and “lookout”. This can be attributed to the preconditions for unsafe acts of the HFACS framework. Topic 2 and 5 have the words “experience”, “fall”, and “training.” Topic 4 has “wearing” and “personal” words, which could be referred to a personal flotation device. These words are found in the bigram and trigram results that correlate with this opinion. Topic 6 categorizes unsafe behavior as a personal factor, “alcohol.” Topic 8 also categorizes the technical factor “exhaust” and the related terms carbon monoxide poisoning, which are categorized under preconditions for unsafe acts.

The results clearly reveal that the dominant topics are mainly related with technical issues and environmental factors. This emphasized the importance of the physical and technological environment factors which are represented under the precondition for unsafe acts of the HFACS model.

The content of extracted topics is more spread than expected. This may be due to the fact that only summaries of reports were used in the analysis. However, the results give important insight on where to focus the risk management efforts in the maritime industry.

5. CONCLUSION AND DISCUSSION

The maritime safety market is not just growing, it's rapidly evolving. The number of maritime risk and safety management studies is on a steady rise, both in the maritime industry and the academic sphere. This thesis contributes to this crucial research field by providing a computerized content analysis for maritime accident reports. LDA, which is a probabilistic topic modeling approach, is used to extract the major causes of accidents. 242 UK maritime accident reports between 2013 and 2023 years are used in the analysis.

The topic number is identified as 9 after topic-coherence analysis. All documents are categorized under nine topics. Documents are distributed differently under the topics. Some of the topics have high documents (Topics 4, 1, 7, and 8), others have less (topics 5, 3, 0, and 7). Results include environmental (thermal), technical (extinguish, GPS, ECDIS, exhaust), and personal (gybe, alcohol, lookout, experience, fall) reasons. Most of the documents are classified under Topic 4 and Topic 1. Topic 0 classifies mishaps related to the ferry. Topic 1 includes collision, and grounding mishaps. Topic 2 includes falls and worker and experience-related accidents. Topic 3 classifies mainly yachts, sailing-based vessels, and GPS-based accidents. Topic 4 categorizes fishing and sinking events, and wearing (or not) personal floatation devices. Topic 5 includes life loss, extinguishment (probably fire), and ECDIS-related accidents. Topic 6 categorizes scallops, winch, alcohol, and emergency-based accidents. Topic 7 includes tugboat, trapping, moor, tug, rope and cord related events and injury-based accidents. Topic 8 classifies fire, carbon monoxide poisoning, heater, and gas related accidents.

One of the aims of this thesis is to explore the results by considering the HFACS framework and identify the areas of HFACS that caused the accidents. However, the results of LDA and TF-IDF show that technical and environmental factors are dominated in the results. Some of the topics include “experience”, “fall”, “training”, “alcohol” words. The extracted model mainly includes the HFACS’ environmental factors category under precondition for unsafe acts. LDA is an unsupervised method; therefore, the results are not aligned with the HFACS-based reasons. The proposed topic model can only show the main indicators for preconditions for unsafe acts. Therefore, the results are assessed by searching the existence of HFACS components in the results.

In summary, the proposed model demonstrates promising results in analyzing maritime accident reports. The model's remarkable success resides in semi-automatically analyzing and categorizing huge amounts of texts within minutes. Otherwise, it won’t be possible to read, memorize and categorize such a huge amount of document set which includes thousands of different words. Therefore, topic modeling is a very promising research area for all sectors.

The limitation of this study is that the data range is small and non-scaled (vessel type, accident type, accident zone, and investigation report). Due to the data's unstructured and complex nature, a potential future research direction may be scalarizing data based on vessel type, report type, and event type before performing topic modeling analysis. Additionally, summaries of the reports are used in the analysis. Another research direction may be performing the analysis with the full sections of the reports. However, it will increase the processing time.

For future work, Latent Semantic Analysis and long-short-term memory (LSTM) applications can be implemented, and the results may be compared. Semantic Analysis can present semantic relations. Also, it may provide more consistent and meaningful results for HFACS-based cause extraction. Different topic models can also be applied, and the evaluation of topic model criteria can be expanded variously, as in the literature, including perplexity, topic diversity, JS score, and so on. Instead of utilizing unsupervised topic models, supervised topic model application can be another option to improve this quality of the solution.

REFERENCES

Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. Information Systems, 112, 102131.

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.

Akyuz, E. (2017). A marine accident analysing model to evaluate potential operational causes in cargo ships. Safety science, 92, 17-25.

Akyuz, E., Celik, M., & Cebi, S. (2016). A phase of comprehensive research to determine marine-specific EPC values in human error assessment and reduction technique. Safety science, 87, 63-75.

Alemayehu, E., & Fang, Y. (2024). Supervised probabilistic latent semantic analysis with applications to controversy analysis of legislative bills. Intelligent Data Analysis, (Preprint), 1-23.

Alves, V., Schwanninger, C., Barbosa, L., Rashid, A., Sawyer, P., Rayson, P., ... & Rummler, A. (2008, September). An exploratory study of information retrieval techniques in domain analysis. In 2008 12th International Software Product Line Conference (pp. 67-76). IEEE

Amado, A., Cortez, P., Rita, P., & Moro, S. (2018). Research trends on Big Data in Marketing: A text mining and topic modeling based literature analysis. European Research on Management and Business Economics, 24(1), 1-7.

Antão, P., Sun, S., Teixeira, A. P., & Guedes Soares, C. (2023). Quantitative assessment of ship collision risk influencing factors from worldwide accident and fleet data. Reliability Engineering and System Safety, 234. https://doi.org/10.1016/j.ress.2023.109166

Asmussen, C. B., & Møller, C. (2019). Smart literature review: a practical topic modelling approach to exploratory literature review. Journal of Big Data, 6(1), 1-18.

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, 11–24. https://doi.org/10.1016/j.tranpol.2020.12.013

Bâra, A., & Oprea, S. V. (2024). The Impact of Academic Publications over the Last Decade on Historical Bitcoin Prices Using Generative Models. Journal of Theoretical and Applied Electronic Commerce Research, 19(1), 538-560.

Baydogan, C., & Alatas, B. (2019, November). Detection of customer satisfaction on unbalanced and multi-class data using machine learning algorithms. In 2019 1st international informatics and software engineering conference (UBMYK) (pp. 1-5). IEEE.

Becker, J., & Kuropka, D. (2003, July). Topic-based vector space model. In Proceedings of the 6th international conference on business information systems (pp. 7-12).

Benchimol, J., Kazinnik, S., & Saadon, Y. (2022). Text mining methodologies with R: An application to central bank texts. Machine Learning with Applications, 8, 100286.

Blei, D. M., & Lafferty, J. D. (2006b). Dynamic topic models. In Proceedings of the 23rd international conference on Machine learning (pp. 113-120).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 993-1022.

Blei, D., & Lafferty, J. (2006a). Correlated topic models. Advances in neural information processing systems, 18, 147.

Boyd-Graber, J., Hu, Y., & Mimno, D. (2017). Applications of topic models. Foundations and Trends® in Information Retrieval, 11(2-3), 143-296.

Çakır, E., Fışkın, R., & Sevgili, C. (2021). Investigation of tugboat accidents severity: An application of association rule mining algorithms. Reliability Engineering and System Safety, 209. https://doi.org/10.1016/j.ress.2021.107470

Cao, Y., Wang, X., Wang, Y., Fan, S., Wang, H., Yang, Z., … Shi, R. (2023). Analysis of factors affecting the severity of marine accidents using a data-driven Bayesian network. Ocean Engineering, 269. https://doi.org/10.1016/j.oceaneng.2022.113563

Changhai, H., & Shenping, H. (2019). Factors correlation mining on maritime accidents database using association rule learning algorithm. Cluster Computing, 22, 4551–4559. https://doi.org/10.1007/s10586-018-2089-z

Chen, Y., Liu, Z., Zhou, H., Zheng, J., & Wang, L. (2022). Social-material aspect of navigation technology: using structural topic models to identify the causes of ship accidents (1973–2018). The Journal of Navigation, 75(1), 35-56.

Chowdhary, K. R. (2020). Fundamentals of artificial intelligence (pp. 603-49). New Delhi:: Springer India.

Daud, S., Ullah, M., Rehman, A., Saba, T., Damaševičius, R., & Sattar, A. (2023). Topic Classification of Online News Articles Using Optimized Machine Learning Models. Computers, 12(1). https://doi.org/10.3390/computers12010016.

Debortoli, S., Müller, O., Junglas, I., Brocke, J. (2016). Text Mining For Information Systems Researchers: An Annotated Topic Modeling Tutorial. Communications of the Association for Information Systems. DOI:10.17705/1cais.03907

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391–407.

Dominguez-Péry, C., Tassabehji, R., Corset, F., & Chreim, Z. (2023). A holistic view of maritime navigation accidents and risk indicators: examining IMO reports from 2011 to 2021. Journal of Shipping and Trade, 8(1). https://doi.org/10.1186/s41072-023-00135-y

Du, P., Zeng, Z., Shen, Y., & Liu, S. (2023). Maritime risk assessment using a non-linear spatial multi-criteria decision method: A case study in the Bohai Sea and Yellow Sea, China. Ocean Engineering, 288. https://doi.org/10.1016/j.oceaneng.2023.115994

Eliopoulou, E., Alissafaki, A., & Papanikolaou, A. (2023). Statistical Analysis of Accidents and Review of Safety Level of Passenger Ships. Journal of Marine Science and Engineering, 11(2). https://doi.org/10.3390/jmse11020410

Fan, S., Blanco-Davis, E., Fairclough, S., Zhang, J., Yan, X., Wang, J., & Yang, Z. (2023). Incorporation of seafarer psychological factors into maritime safety assessment. Ocean and Coastal Management, 237. https://doi.org/10.1016/j.ocecoaman.2023.106515.

Fu, S., Yu, Y., Chen, J., Xi, Y., & Zhang, M. (2022). A framework for quantitative analysis of the causation of grounding accidents in arctic shipping. Reliability Engineering & System Safety, 226, 108706.

Fu, S., Zhang, Y., Zhang, M., Han, B., & Wu, Z. (2023). An object-oriented Bayesian network model for the quantitative risk assessment of navigational accidents in ice-covered Arctic waters. Reliability Engineering and System Safety, 238. https://doi.org/10.1016/j.ress.2023.109459

Gan, J., & Qi, Y. (2021). Selection of the optimal number of topics for LDA topic model—Taking patent policy analysis as an example. Entropy, 23(10). https://doi.org/10.3390/e23101301.

George, A. (2022). Python text mining. New Delhi, India: BPB Publications

Gupta, R. K., Agarwalla, R., Naik, B. H., Evuri, J. R., Thapa, A., & Singh, T. D. (2022). Prediction of Research Trends using LDA based Topic Modeling. Global Transitions Proceedings.

Harrando, I., Lisena, P., & Troncy, R. (2021, September). Apples to apples: A systematic evaluation of topic models. In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021) (pp. 483-493).

HFACS. HFACS Framework, https://www.hfacs.com/hfacs-framework.html (Access Date: 26.06.2024)

HS, C., & Shenoy, M. K. (2020). Advanced text documents information retrieval system for search services. Cogent Engineering, 7(1), 1856467.

Huynh-The, T., Banos, O., Le, B. V., Bui, D. M., Yoon, Y., & Lee, S. (2015). Traffic behavior recognition using the pachinko allocation model. Sensors, 15(7), 16040-16059

Hwang, T., & Youn, I. H. (2022). Latent-Cause Extraction Model in Maritime Collision Accidents Using Text Analytics on Korean Maritime Accident Verdicts. Applied Sciences (Switzerland), 12(2). https://doi.org/10.3390/app12020914

IMO (2024). Accessed from: https://www.imo.org/en/About/Conventions/Pages/ListOfConventions.aspx (Access date: 20.01.2024)

Jackson, P., & Moulinier, I. (2007). Natural language processing for online applications: Text retrieval, extraction and categorization (Vol. 5). John Benjamins Publishing.

Kandel, R., & Baroud, H. (2024). A data-driven risk assessment of Arctic maritime incidents: Using machine learning to predict incident types and identify risk factors. Reliability Engineering & System Safety, 243, 109779.

Kapadia, S. (2022, December 24). Evaluate topic models: Latent Dirichlet allocation (LDA). Medium. https://towardsdatascience.com/evaluate-topic-model-in-python-latent-dirichlet-allocation-lda-7d57484bb5d0 (Access Date: 10.06.2020)

Karami, A., Lundy, M., Webb, F., & Dwivedi, Y. K. (2020). Twitter and research: A systematic literature review through text mining. IEEE Access, 8, 67698-67717.

Kasyk, L., Wolnowska, A. E., Pleskacz, K., & Kapuściński, T. (2023). The Analysis of Social and Situational Systems as Components of Human Errors Resulting in Navigational Accidents. Applied Sciences, 13(11), 6780.

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: State of the art, current trends and challenges. Multimedia tools and applications, 82(3), 3713-3744.

Kim, G., & Lim, S. (2022). Development of an Interpretable Maritime Accident Prediction System Using Machine Learning Techniques. IEEE Access, 10, 41313–41329. https://doi.org/10.1109/ACCESS.2022.3168302

Lamirel, J. C., Lareau, F., & Malaterre, C. (2024). CFMf topic-model: comparison with LDA and Top2Vec. Scientometrics, 1-19.

Lan, H., & Ma, X. (2024). Risk Evolution Analysis of Seafarers' Unsafe Acts in Maritime Accidents Based on Directed Weighted CN. Applied Sciences, 14(6), 2595.

Lan, H., Ma, X., Qiao, W., & Deng, W. (2023). Determining the critical risk factors for predicting the severity of ship collision accidents using a data-driven approach. Reliability Engineering and System Safety, 230. https://doi.org/10.1016/j.ress.2022.108934

Latha, K. (2017). Experiment and evaluation in information retrieval models. London, England: CRC Press

Lee, K. N., Lee, H., Kim, J. H., Kim, Y., & Lee, S. H. (2023). Comparing Social Media and News Articles on Climate Change: Different Viewpoints Revealed. KSII Transactions on Internet & Information Systems, 17(11).

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In Proceedings of the 23rd international conference on Machine learning (pp. 577-584).

Li, W., & McCallum, A. (2008). Pachinko allocation: Scalable mixture models of topic correlations. J. of Machine Learning Research. Submitted.

Liao, S., Weng, J., Zhang, Z., Li, Z., & Li, F. (2023). Probabilistic Modeling of Maritime Accident Scenarios Leveraging Bayesian Network Techniques. Journal of Marine Science and Engineering, 11(8). https://doi.org/10.3390/jmse11081513

Liu, J., Chen, H., Li, D., Wang, Z., & Wang, J. (2024). Quantitative risk assessment of a spill fire caused by the continuously leaked fuel in a sealed ship engine room. Ocean Engineering, 303, 117664.

Liu, X., Yuan, H., Xiao, C., Wang, Y., & Yu, Q. (2022). Hybrid-driven vessel trajectory prediction based on uncertainty fusion. Ocean Engineering, 248. https://doi.org/10.1016/j.oceaneng.2022.110836

Ma, L., Ma, X., Wang, T., Zhao, Y., & Lan, H. (2024). A data-driven approach to determine the distinct contribution of human factors to different types of maritime accidents. Ocean Engineering, 295. https://doi.org/10.1016/j.oceaneng.2024.116874

MAIB (2022). Marine Accident Investigation Branch Annual Report 2022. Marine Accident Recommendations and Statistics. https://assets.publishing.service.gov.uk/media/64be8e9b1e10bf000e17cd46/MAIBAnnualReport2022.pdf (Accessed on: 20.05.2024)

Mcauliffe, J., & Blei, D. (2007). Supervised topic models. Advances in neural information processing systems, 20.

Meesad, P. (2021). Thai fake news detection based on information retrieval, natural language processing and machine learning. SN Computer Science, 2(6), 425.

Mendonça, M., & Figueira, Á. (2024). Topic Extraction: BERTopic's Insight into the 117th Congress's Twitterverse. Informatics, 11(1). https://doi.org/10.3390/informatics11010008.

Momtazi, S., & Naumann, F. (2013). Topic modeling for expert finding using latent Dirichlet allocation. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 3(5), 346–353. https://doi.org/10.1002/widm.1102

NBC News (2012). Accessed from: https://www.nbcnews.com/sciencemain/10-causes-titanic-tragedy-620220 (Access Date: 15.05.2024)

Özaydın, E., Fışkın, R., Uğurlu, Ö., & Wang, J. (2022). A hybrid model for marine accident analysis based on Bayesian Network (BN) and Association Rule Mining (ARM). Ocean Engineering, 247. https://doi.org/10.1016/j.oceaneng.2022.110705

ÖZKANLISOY, Ö., & AKKARTAL, E. (2021). The Effect of Suez Canal Blockage on Supply Chains. Dokuz Eylül Üniversitesi Denizcilik Fakültesi Dergisi, 14(1), 51-79.

Pan, J., Wang, Y., Wang, T., & Xu, M. (2024). Study on Assessment of Collision Probability between Ship and Bridge Based on Automatic Identify System Data. Journal of Marine Science and Engineering, 12(3). https://doi.org/10.3390/jmse12030452

Panichella, A. (2021). A systematic comparison of search-based approaches for LDA hyperparameter tuning. Information and Software Technology, 130, 106411.

Pilatis, A. N., Pagonis, D. N., Serris, M., Peppa, S., & Kaltsas, G. (2024). A Statistical Analysis of Ship Accidents (1990–2020) Focusing on Collision, Grounding, Hull Failure, and Resulting Hull Damage. Journal of Marine Science and Engineering, 12(1), 122.

Pramanik, P., & Jana, R. K. (2023, November 30). Identifying research trends of machine learning in business: a topic modeling approach. Measuring Business Excellence. Emerald Publishing. https://doi.org/10.1108/MBE-07-2021-0094

Puisa, R., Lin, L., Bolbot, V., & Vassalos, D. (2018). Unravelling causal factors of maritime incidents and accidents. Safety science, 110, 124-141.

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015, July). Short and sparse text topic modeling via self-aggregation. In 24th International Joint Conference on Artificial Intelligence, IJCAI 2015 (pp. 2270-2276). AAAI Press/International Joint Conferences on Artificial Intelligence.

Quatrini, E., Colabianchi, S., Costantino, F., & Tronci, M. (2022, January 1). Clustering Application for Condition-Based Maintenance in Time-Varying Processes: A Review Using Latent Dirichlet Allocation. Applied Sciences (Switzerland). MDPI. https://doi.org/10.3390/app12020814

Rawson, A., & Brito, M. (2023). A survey of the opportunities and challenges of supervised machine learning in maritime risk analysis. Transport Reviews, 43(1), 108–130. https://doi.org/10.1080/01441647.2022.2036864

Roberts, M. E., Stewart, B. M., Tingley, D., & Airoldi, E. M. (2013). The structural topic model and applied social science. In Advances in neural information processing systems workshop on topic models: computation, application, and evaluation (Vol. 4, No. 1, pp. 1-20).

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In WSDM 2015 - Proceedings of the 8th ACM International Conference on Web Search and Data Mining (pp. 399–408). Association for Computing Machinery. https://doi.org/10.1145/2684822.2685324

Roshdi, A., & Roohparvar, A. (2015). Information retrieval techniques and applications. International Journal of Computer Networks and Communications Security, 3(9), 373-377.

Semary, N. A., Ahmed, W., Amin, K., Pławiak, P., & Hammad, M. (2024). Enhancing machine learning-based sentiment analysis through feature extraction techniques. PLoS ONE, 19(2 February). https://doi.org/10.1371/journal.pone.0294968

Shahbazi, Z., & Byun, Y. C. (2021). Fake media detection based on natural language processing and blockchain approaches. IEEE Access, 9, 128442-128453.

Shappell, S. A., & Wiegmann, D. A. (2000). The human factors analysis and classification system--HFACS.

Shi, J., Liu, Z., Feng, Y., Wang, X., Zhu, H., Yang, Z., ... & Wang, H. (2024). Evolutionary model and risk analysis of ship collision accidents based on complex networks and DEMATEL. Ocean Engineering, 305, 117965.

Shi, X., Zhuang, H., & Xu, D. (2021). Structured survey of human factor-related maritime accident research. Ocean Engineering, 237, 109561.

Shin, S. H., Kwon, O. K., Ruan, X., Chhetri, P., Lee, P. T. W., & Shahparvari, S. (2018). Analyzing sustainability literature in maritime studies with text mining. Sustainability, 10(10), 3522.

Subhashini, R., & Senthil Kumar, V. J. (2011). A framework for efficient information retrieval using NLP techniques. In Computer Networks and Information Technologies: Second International Conference on Advances in Communication, Network, and Computing, CNC 2011, Bangalore, India, March 10-11, 2011. Proceedings 2 (pp. 391-393). Springer Berlin Heidelberg.

Swain, A. D. (1990). Human reliability analysis: Need, status, trends and limitations. Reliability Engineering & System Safety, 29(3), 301–313. doi:10.1016/0951-8320(90)90013-d

Syed, S., & Spruit, M. (2018). Exploring Symmetrical and Asymmetrical Dirichlet Priors for Latent Dirichlet Allocation. International Journal of Semantic Computing, 12(3), 399–423. https://doi.org/10.1142/S1793351X18400184

Terragni, S., Candelieri, A., & Fersini, E. (2023). The role of hyper-parameters in relational topic models: Prediction capabilities vs topic quality. Information Sciences, 632, 252–268. https://doi.org/10.1016/j.ins.2023.02.076

The HFACS framework. HFACS, Inc | The HFACS Framework. (n.d.). https://www.hfacs.com/hfacs-framework.html (Access Date: 26.06.2024)

The School of Informatics at the University of Edinburgh. (2017, January 27). Prof. David Blei - Probabilistic Topic Models and User Behavior []. Retrieved from https://www.youtube.com/watch?v=FkckgwMHP2s (Accessed in: 29.05.2024)

Thompson, L., & Mimno, D. (2020). Topic modeling with contextualized word representation clusters. arXiv preprint arXiv:2010.12626.

Ung, S. T. (2018). Human error assessment of oil tanker grounding. Safety science, 104, 16-28.

Vayansky, I., & Kumar, S. A. P. (2020). A review of topic modeling methods. Information Systems, 101582. doi:10.1016/j.is.2020.101582.

Wallach, H. M., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In Advances in Neural Information Processing Systems 22 - Proceedings of the 2009 Conference (pp. 1973–1981). Neural Information Processing Systems.

Wan, S., Yang, X., Chen, X., Qu, Z., An, C., Zhang, B., ... & Bi, H. (2022). Emerging marine pollution from container ship accidents: Risk characteristics, response strategies, and regulation advancements. Journal of Cleaner Production, 134266.

Wang, C., Blei, D., & Heckerman, D. (2012). Continuous time dynamic topic models. arXiv preprint arXiv:1206.3298.

Wang, L., Huang, R., Shi, W., & Zhang, C. (2021). Domino effect in marine accidents: Evidence from temporal association rules. Transport Policy, 103, 236–244. https://doi.org/10.1016/j.tranpol.2021.02.006

Wang, X., & McCallum, A. (2006). Topics over time: a non-markov continuous-time model of topical trends. In Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 424-433).

Wang, Y., & Fu, S. (2022). Framework for Process Analysis of Maritime Accidents Caused by the Unsafe Acts of Seafarers: A Case Study of Ship Collision. Journal of Marine Science and Engineering, 10(11). https://doi.org/10.3390/jmse10111793

Wu, B., Yan, X., Wang, Y., & Soares, C. G. (2017). An evidential reasoning-based CREAM to human reliability analysis in maritime accident process. Risk analysis, 37(10), 1936-1957.

Yan, K., Wang, Y., Jia, L., Wang, W., Liu, S., & Geng, Y. (2023). A content-aware corpus-based model for analysis of marine accidents. Accident Analysis and Prevention, 184. https://doi.org/10.1016/j.aap.2023.106991

Yang, X., Zhi, J., Zhang, W., Xu, S., & Meng, X. (2023). A Novel Data-Driven Prediction Framework for Ship Navigation Accidents in the Arctic Region. Journal of Marine Science and Engineering, 11(12). https://doi.org/10.3390/jmse11122300

Yıldırım, U., Başar, E., & Uğurlu, Ö. (2019). Assessment of collisions and grounding accidents with human factors analysis and classification system (HFACS) and statistical methods. Safety Science, 119, 412-425.

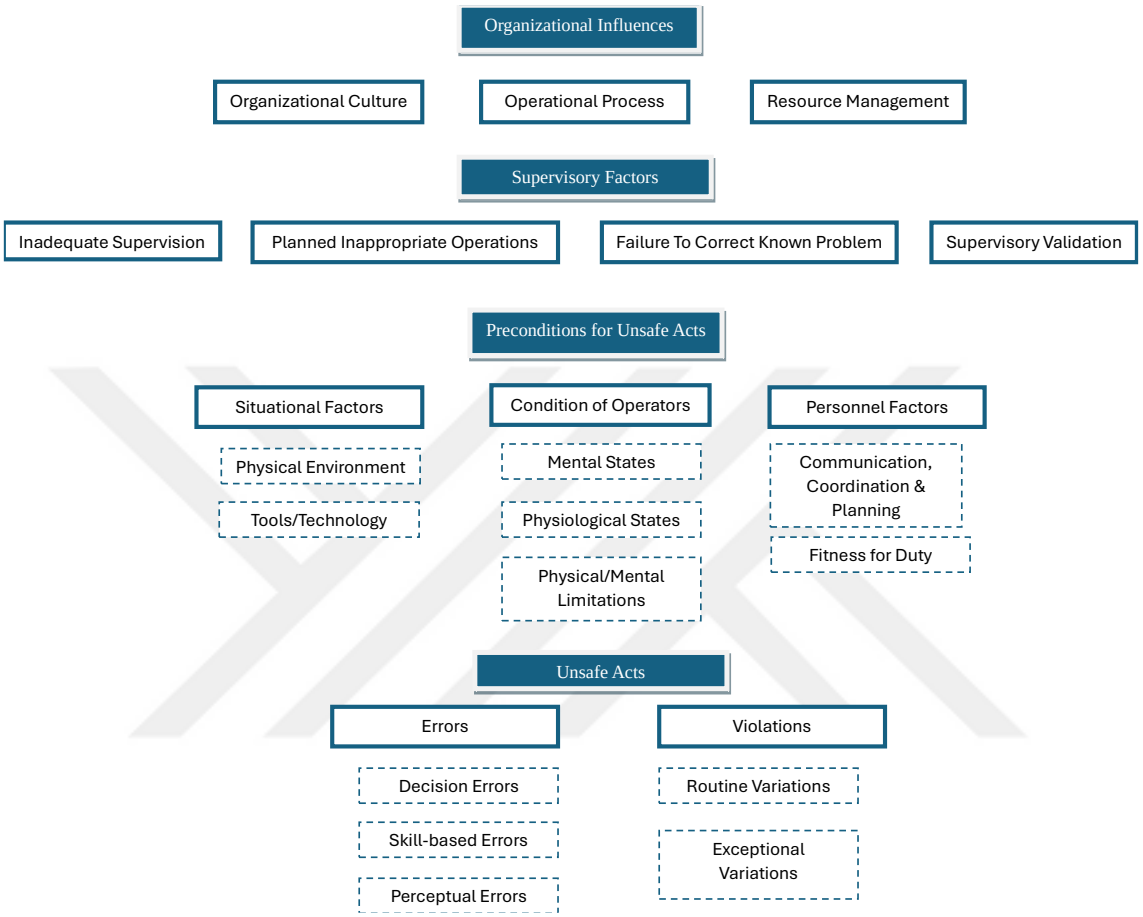
Yu, J., Wu, Z., Liu, W., & Zhao, W. (2023). Identifying the Causes of Ship Collisions Accident Using Text Mining and Bayesian Networks. Elektronika Ir Elektrotechnika, 29(6), 58–67. https://doi.org/10.5755/j02.eie.35630

Zhang, J., Gao, W., & Jia, Y. (2023). WES-BTM: A Short Text-Based Topic Clustering Model. Symmetry, 15(10), 1889.

Zhang, J., Jin, M., Wan, C., Dong, Z., & Wu, X. (2024). A Bayesian network-based model for risk modeling and scenario deduction of collision accidents of inland intelligent ships. Reliability Engineering and System Safety, 243. https://doi.org/10.1016/j.ress.2023.109816

Zhou, K., Xing, W., Wang, J., Li, H., & Yang, Z. (2024). A data-driven risk model for maritime casualty analysis: A global perspective. Reliability Engineering and System Safety, 244. https://doi.org/10.1016/j.ress.2023.109925

APPENDIX A. HFACS Framework (Source: https://www.hfacs.com/hfacs-framework.html)



APPENDIX B. Coherence scores of k, α, and β = 0.91 values for 100% corpus values

Validation_Set	Topics	Alpha	Beta	Coherence
100% Corpus	9	0.91	0.91	0.433376
100% Corpus	10	0.91	0.91	0.424221
100% Corpus	8	asymmetric	0.91	0.40973
100% Corpus	7	asymmetric	0.91	0.407934
100% Corpus	7	0.91	0.91	0.382919
100% Corpus	9	asymmetric	0.91	0.380131
100% Corpus	10	0.61	0.91	0.376072
100% Corpus	10	0.31	0.91	0.371956
100% Corpus	5	0.61	0.91	0.353757
100% Corpus	10	0.01	0.91	0.353687
100% Corpus	6	0.91	0.91	0.352533
100% Corpus	7	0.61	0.91	0.352274
100% Corpus	5	0.91	0.91	0.347647
100% Corpus	9	0.31	0.91	0.347202
100% Corpus	5	asymmetric	0.91	0.34686
100% Corpus	10	asymmetric	0.91	0.344567
100% Corpus	3	symmetric	0.91	0.343442
100% Corpus	4	asymmetric	0.91	0.340698
100% Corpus	3	0.91	0.91	0.333601
100% Corpus	6	0.31	0.91	0.333085

APPENDIX C. Topic-word and topic-document classification for k = 7, k =8, and k=10.

Topic-word clasification for k =7

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
chief	hose	keel	speedboat	watch	grounding	crew
carbon	carbon	yacht	frame	aground	tidal	skipper
yacht	gas	channel	manoeuvred	distracted	stevedore	boat
monoxide	monoxide	pilot	aground	grounding	scrabster	fishing
buoy	bramble	manager	jacketed	cargo	compartment	safety
modification	fuel	chain	fitting	air	semitrailer	crewman
exhaust	truck	crushed	ballast	oow	breakwater	port
moored	tanker	sailing	broadcast	electronic	ecdis	board
expedition	chain	chief	tank	ferry	shoal	deck
traffic	boat	arrangement	closefitting	ecdis	astern	passenger
test	marine	stack	receive	heater	chemical	vessel
cockpit	engine	cargo	thorn	engine	judgment	ferry
poisoning	container	climbed	liferafts	tank	electronic	overboard
consciousness	temperature	victim	contributing	thruster	skill	engine
canopy	oil	marina	ecdis	bridge	fault	cargo

Topic-document clasification for k =7

Row Labels	Count of Document
0	6
1	8
2	4
4	15
5	2
6	207
Grand Total	**242**

Topic-word clasification for k =8

Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7
monoxide	anchor	pilot	keel	ecdis	creel	crew	tug
carbon	cargo	port	yacht	yacht	air	skipper	officer
rescue	ferry	traffic	helm	stevedore	heater	fishing	container
exhaust	officer	bridge	sailing	crane	plb	crewman	mooring
pool	container	channel	disabled	aground	oil	boat	bridge
collision	bridge	ladder	inverted	tank	furnace	safety	watch
cargo	watch	chief	boat	channel	vent	engine	line
passenger	tug	tanker	receive	watchkeeper	pipe	overboard	passenger
hose	manager	harbour	knocked	cargo	staff	board	speed
gas	wind	breakwater	instructor	passage	rower	deck	cargo
vessel	bow	grounding	survivor	carrier	rope	port	pilot
deck	chain	power	strap	tidal	exhaust	deckhand	master
cruiser	chief	vt	inversion	aid	singlehanded	wearing	chief
boat	speed	officer	smoke	master	explosion	emergency	inflatable
line	bramble	expedition	ignition	distracted	secured	sank	test

Topic-document clasification for k =8

Row Labels	**Count of Document**
0	30
1	26
2	14
3	1
4	15
5	4
6	133
7	19
Grand Total	**242**

Topic-word clasification for k =10

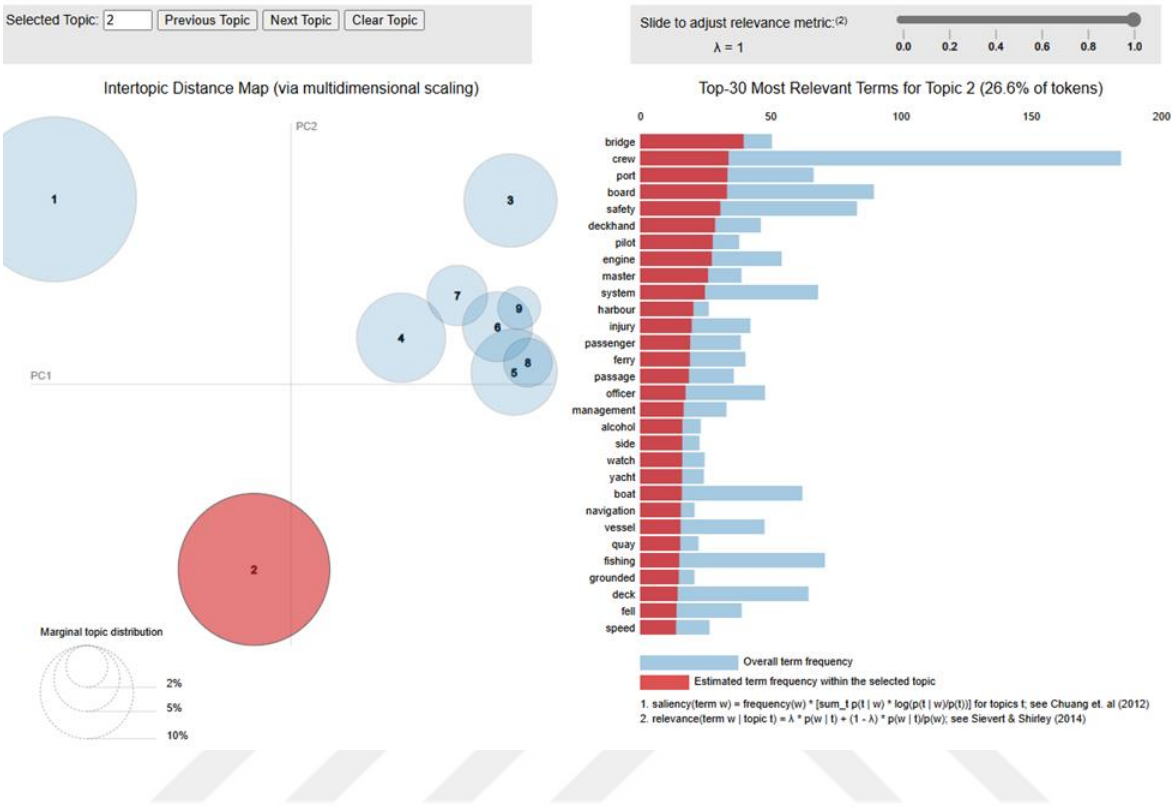
Topic 0	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9
carbon	ferry	cargo	keel	bridge	skipper	rope	fuel	crew	load
monoxide	steering	officer	boat	channel	fishing	deckhand	ai	skipper	isle
exhaust	net	chief	environmental	collision	watch	mooring	vhf	crewman	container
engine	engine	operation	sailing	ecdis	passage	crew	yacht	boat	tidal
yacht	chain	crewman	hose	damage	officer	fishing	ferry	safety	loaded
gas	passenger	oil	inverted	grounding	personal	deck	jetty	overboard	anchoring
poisoning	damaged	crane	knowledge	ferry	anchor	vessel	thruster	tug	truck
boat	collided	stevedore	disabled	vessel	bridge	passenger	container	emergency	condition
crane	engineer	pilot	record	engine	crew	board	davit	rescue	jumped
ferry	maintenance	mooring	broadcast	navigation	device	modification	tanker	flooding	boatman
compartment	bridge	test	liferafts	radar	boat	pool	tower	board	view
buoy	alert	safe	started	grounded	vessel	winch	ladder	equipment	violent
manufacturer	speed	response	day	pilot	aground	skipper	propeller	passenger	hazard
cruiser	action	deck	steering	visibility	flotation	line	oil	engine	northwest
traffic	shock	vehicle	distress	port	wearing	crewman	passenger	port	firth

Topic-document clasification for k =10

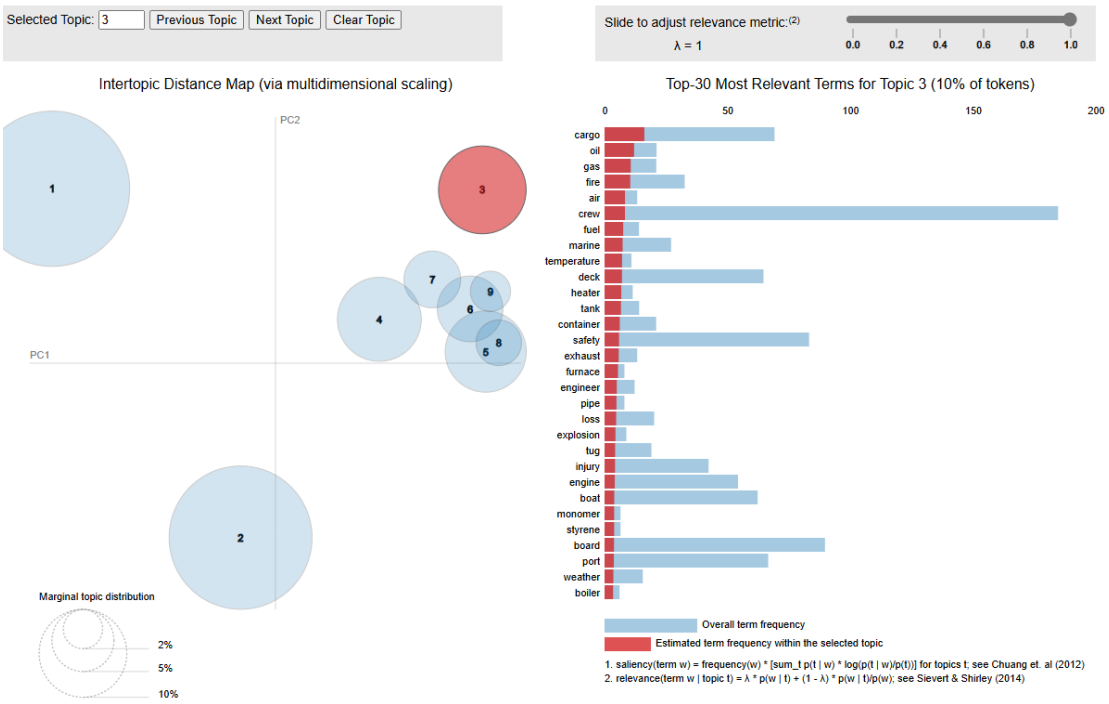
Row Labels	Count of Document
0	17
1	10
2	28
3	3
4	15
5	46
6	30
7	10
8	82
9	1
Grand Total	**242**

APPENDIX D. pyLDAvis graphs for topics from 2 to 9

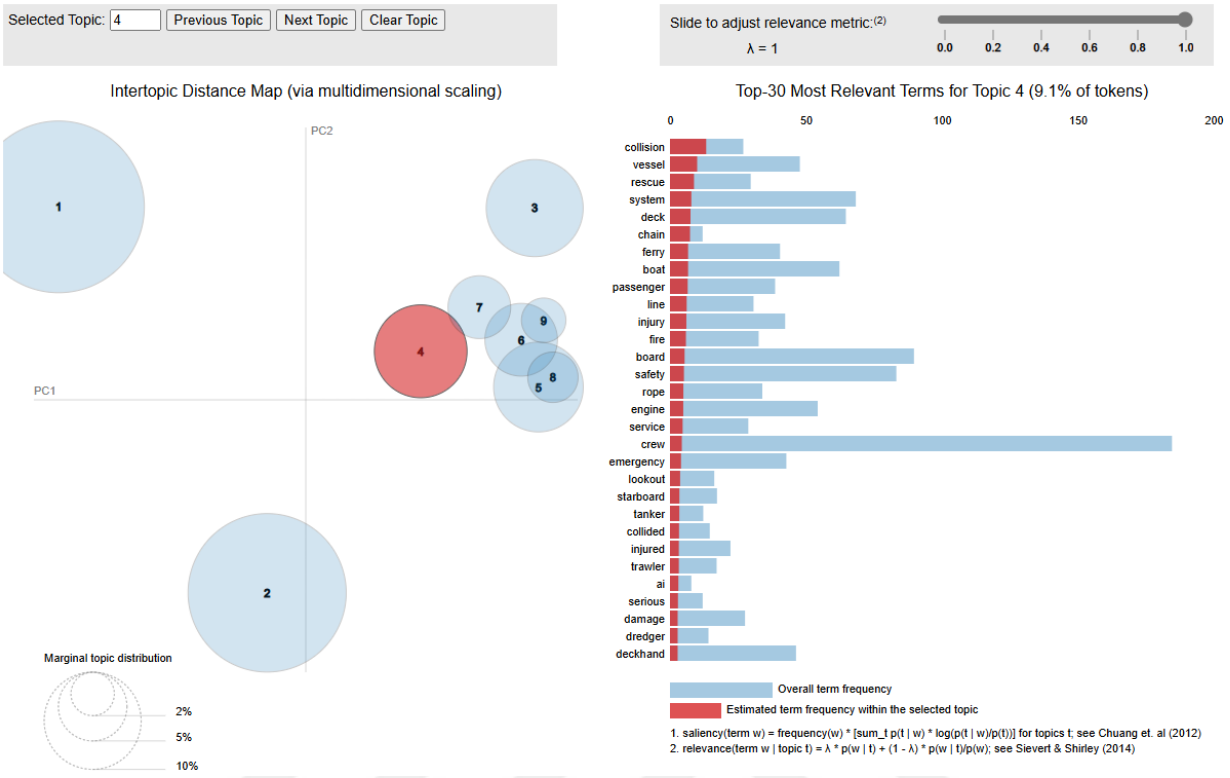
Topic number for 2.



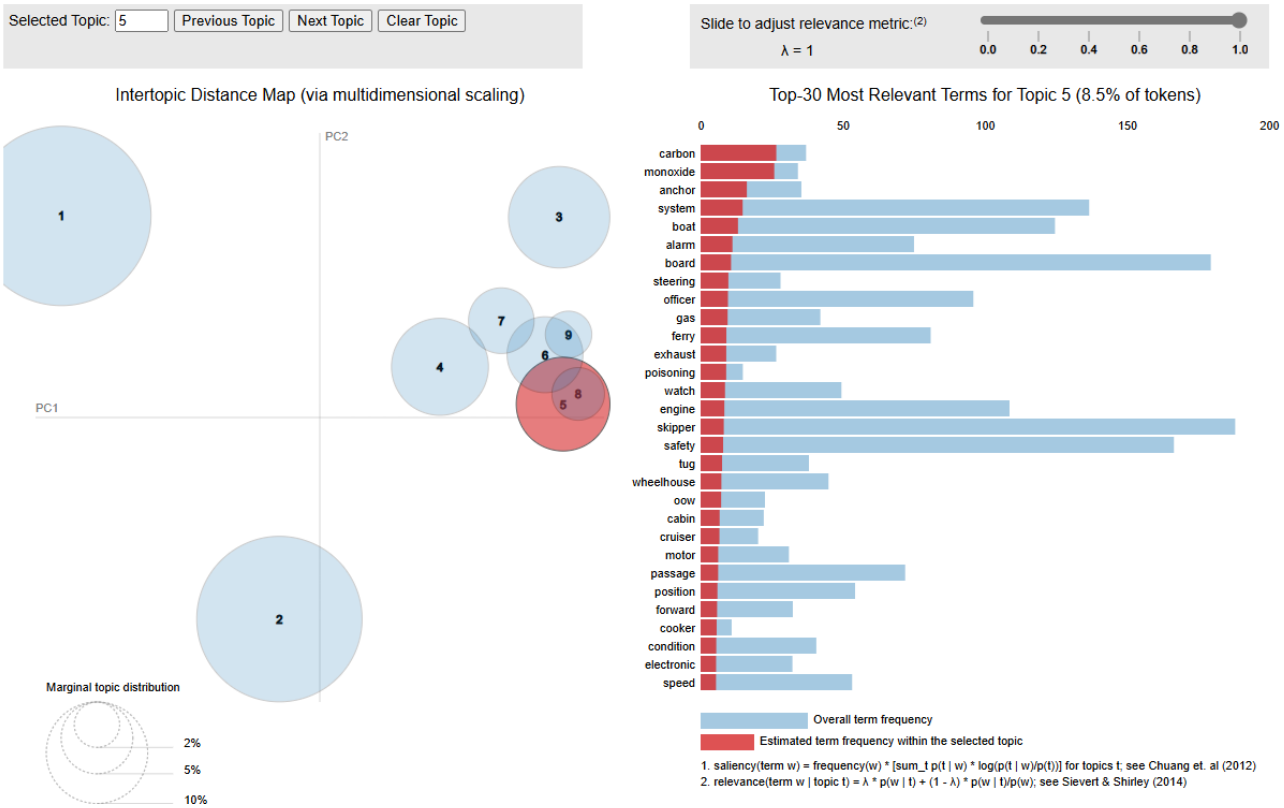
Topic number for 3.



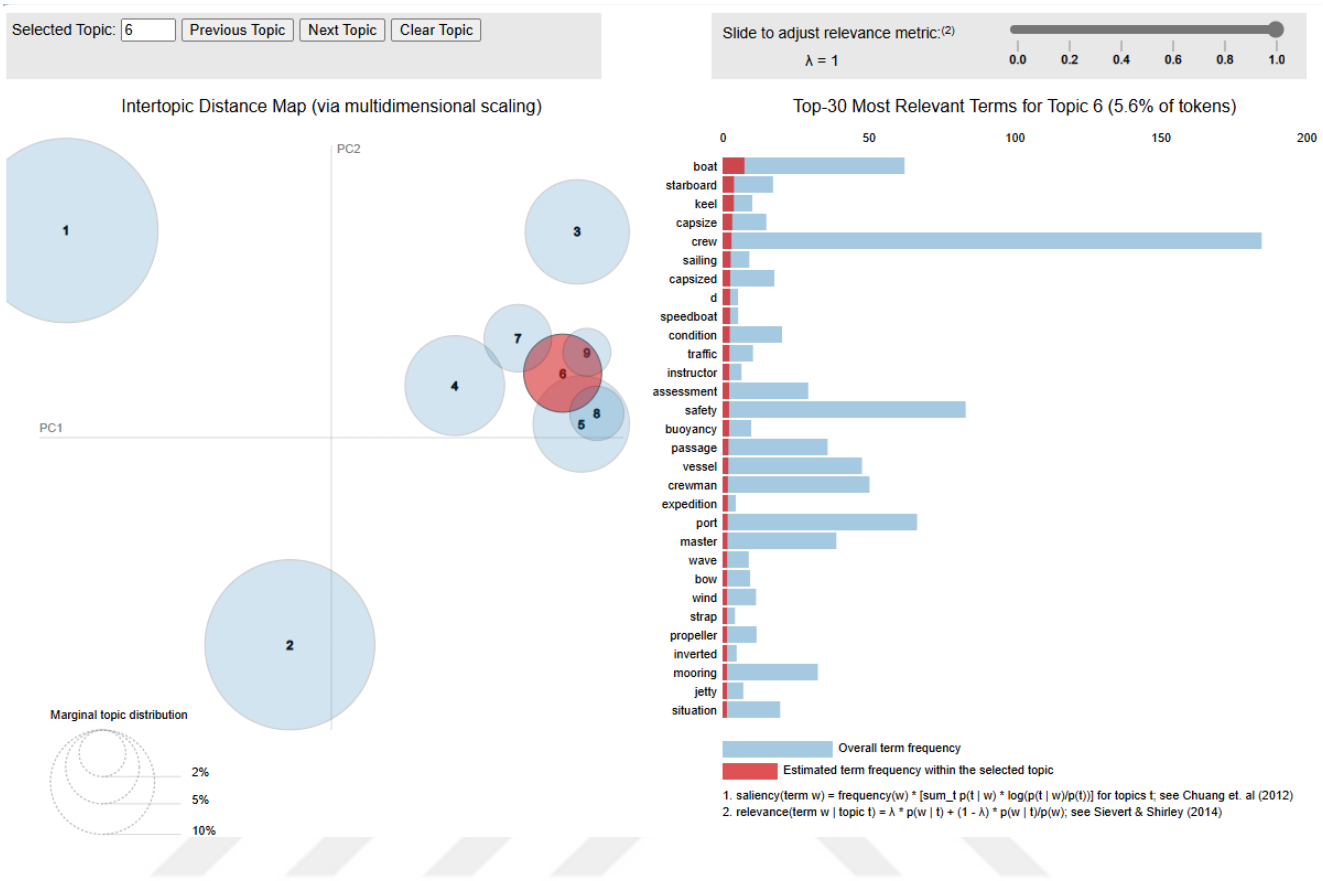
Topic number for 4.



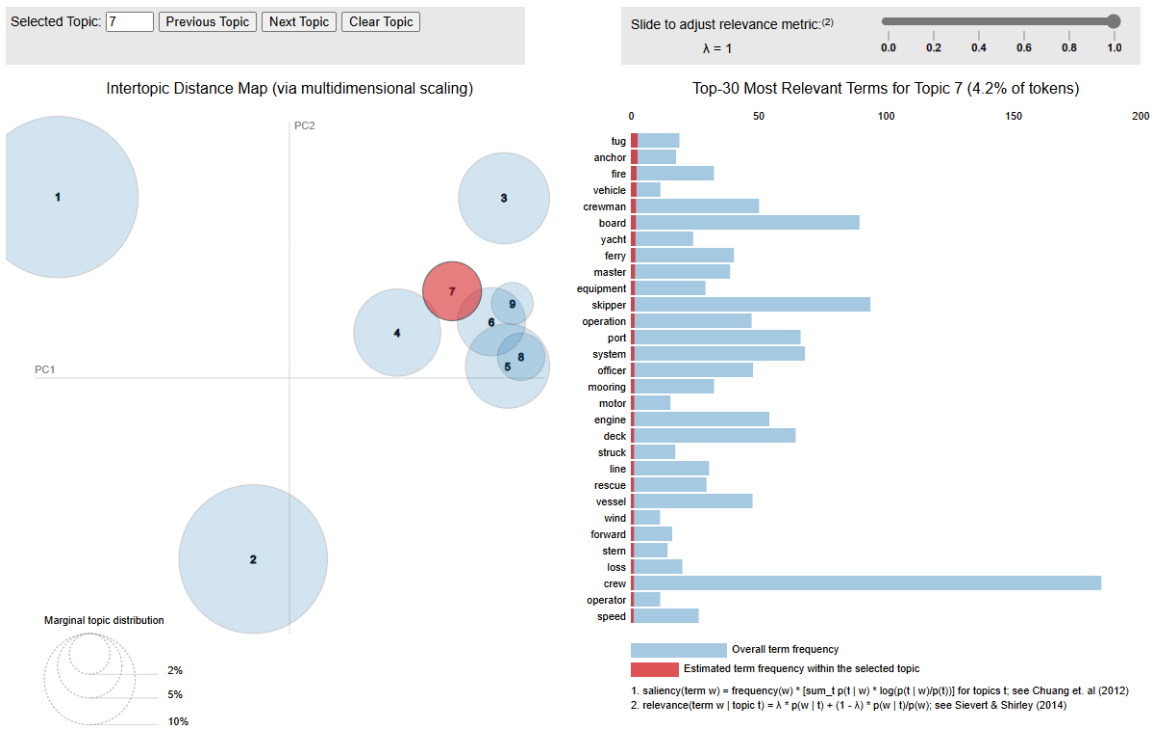
Topic number for 5.



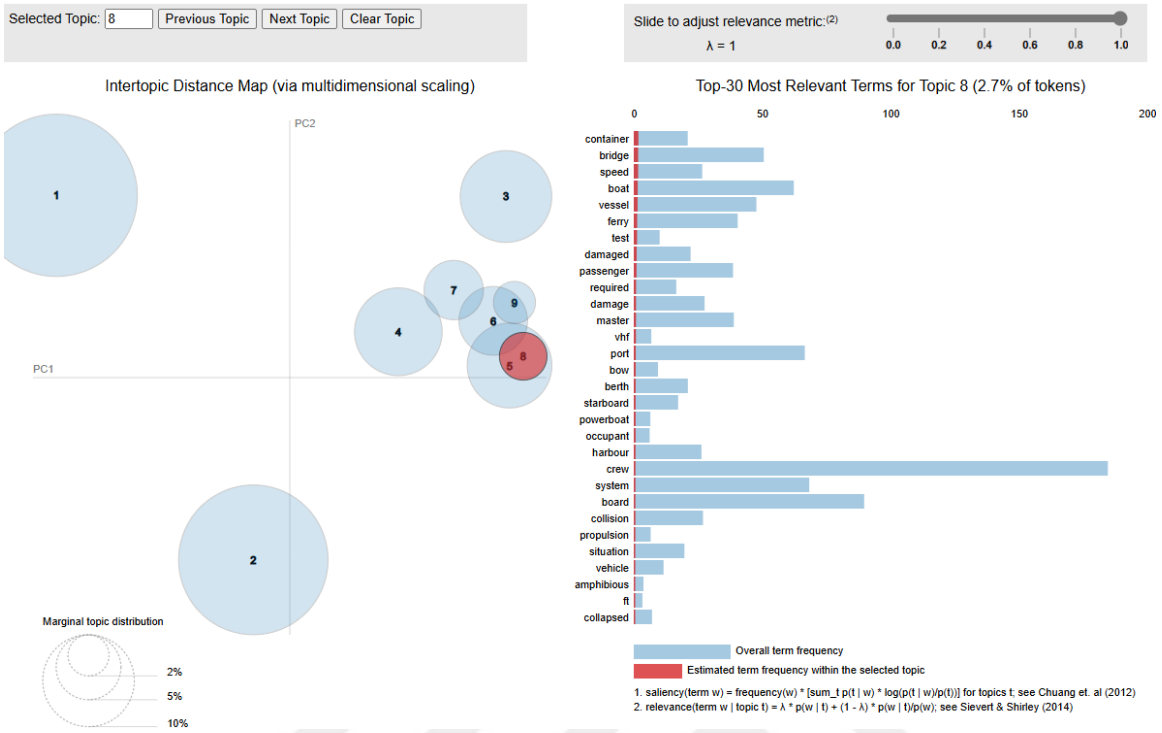
Topic number for 6.



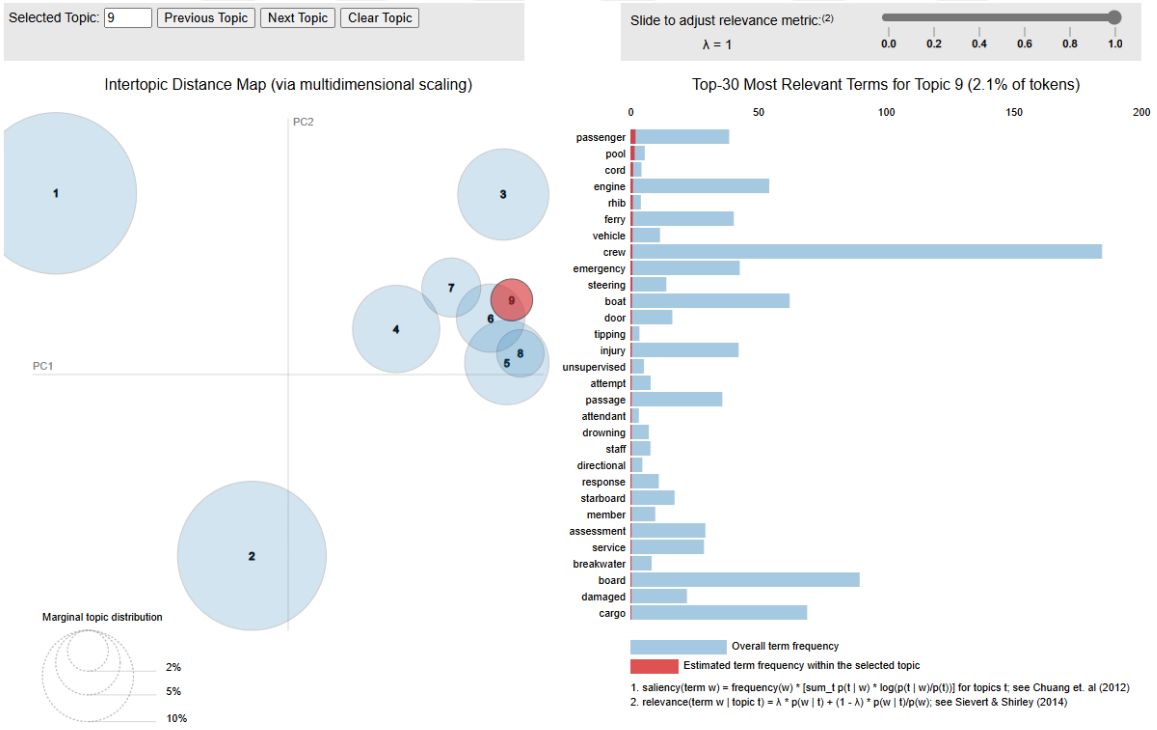
Topic number for 7.



Topic number for 8.



Topic number for 9.



BIOGRAPHY

Ceren Kesmen holds Bachelors degree in Industrial Engineering. She was graduated from the faculty with honor degree. Her main research interests are Data Mining Techniques, Natural Language Processing, and Machine Learning.