International Journal of Information Management Data Insights 1 (2021) 100005

Contents lists available at ScienceDirect

International Journal of Information Management Data Insights

journal homepage: www.elsevier.com/locate/jjimei

Text mining with sentiment analysis on seafarers' medical documents

Nalini Chintalapudi [a], Gopi Battineni [a,*], Marzio Di Canio [b], Getu Gamo Sagaro [a], Francesco Amenta [a,b]

[a] Telemedicine and Tele pharmacy Center, School of Medicinal and Health Products Sciences, University of Camerino, 62032 Camerino, Italy
[b] Research Department, Centro Internazionale Radio Medico (C.I.R.M.), 00144 Rome, Italy

ARTICLE INFO

Keywords:
Text mining
Seafarers
Sentiment analysis
Word clouds
Machine learning

ABSTRACT

Digital health systems contain large amounts of patient records, doctor notes, and prescriptions in text format. This information summarized over the electronic clinical information will lead to an improved quality of healthcare, the possibility of fewer medical errors, and low costs. Besides, seafarers are more vulnerable to have accidents, and prone to health hazards because of work culture, climatic changes, and personal habits. Therefore, text mining implementation in seafarers' medical documents can generate better knowledge of medical issues that often happened onboard. Medical records are collected from digital health systems of Centro Internazionale Radio Medico (C.I.R.M.) which is an Italian Telemedical Maritime Assistance System (TMAS). Three years (2018–2020) patient data have been used for analysis. Adoption of both lexicon and Naïve Bayes' algorithms was done to perform sentimental analysis and experiments were conducted over R statistical tool. Visualization of symptomatic information was done through word clouds and 96% of the correlation between medical problems and diagnosis outcome has been achieved. We validate the sentiment analysis with more than 80% accuracy and precision.

1. Introduction

Sailing on ships is one of the risky occupations with its perks towards personal health and safety measures, and seafarers are highly vulnerable to have accidents and different diseases because of work culture, climatic changes, personal habits, etc. (Bal, Arslan & Tavacioglu, 2015; Nittari et al., 2019). On other side, seafarers may largely experience fatal and serious injuries. The main causes behind this are dangerous work practices, neglecting ship rules, and regulations (Çakir, 2019).

There could be possible differences in accident rates between merchant ships from different nationalities. It is proven by Ádám, Rasmussen, Pedersen and Jepsen (2014) work that specifies that western European seafarers (Danish) were had an overall injured rate of 17.5 per 100,000 person-days, and this is significantly higher than Eastern European, Southeast Asian, and Indian seaman. When a seafarer is ill or has an injury , the ship in charge contact the Telemedical Maritime Assistance Service (TMAS) center for immediate help (Westlund, Attvall, Nilsson & Jensen, 2016). These centers can be able to assist them through telemedicine , and patient records are carefully stored in digital health systems. In this study, we have collected seafarer medical documents from the Italian TMAS center called Centro Internazionale Radio Medico (C.I.R.M), which is offering medical assistance to seafarers for 85 years (Mahdi & Amenta, 2016).

The applications of text mining (TM) have continuously evolved in modern times. In the healthcare industry, there are several studies on artificial intelligence (AI) techniques including machine learning (ML), drug classification, and predictive analytics (Battineni, Sagaro, Chinatalapudi & Amenta, 2020). However, because of medical database evolution, TM is getting high in demand to understand patient opinion on provided medical services (Friedman, Rindflesch & Corn, 2013). In the healthcare sector, text analytic operations are very often used to understand patient satisfaction and maintain streamlined operations (Kim & Delen, 2018).

Besides, opinion mining or sentiment analysis can be able to manage the interpretation of subjective statements and user responses (Rathore, Kar & Ilavarasan, 2017). In healthcare, sentiment analysis is popular because of its advantages during the assessment of medical records and makes it easy for doctors in decision making. When the patient got a sickness, the doctor diagnoses the patient's condition based on the symptomatic data and stores it in digital health systems called electronic health records (EHR).

In EHR, the doctor describes his opinion or observations that intends to understand patient feedback (Denecke & Deng, 2015). In such conditions, opinion mining helps to understand the patient attitude regarding to the principal contextual polarity of health records. This is true especially when doctors and patients express their opinion about health services and issues through online platforms like social media, blogs, and websites. Many studies (Khadjeh Nassirtoussi, Aghabozorgi, Wah & Ngo,

* Corresponding author.
E-mail address: gopi.battineni@unicam.it (G. Battineni).

https://doi.org/10.1016/j.jjimei.2020.100005
Received 30 October 2020; Received in revised form 9 December 2020; Accepted 10 December 2020

2014; Zhang, Chen & Liu, 2015) attempted to use opinion mining in different sectors like healthcare, business, banking, and others but there is no study in the maritime domain. In this work we have implemented sentiment analysis with TM to evaluate symptomatic information of seafarer's pathologies since life at sea makes it complicated in the absence of health professionals onboard.

In this paper, we randomly selected more than 3000 seafarer three years (2018–20) medical documents from C.I.R.M, and we have conducted experiments over three different corpora related to diagnostics of medical problems. Each corpus indicates the severity of an individual medical problem. We incorporated text mining methods in medical document analysis and a review of common pathologies that occurred onboard. In particular, the following research questions have been explored in this analysis.

RQ1: What are the frequent medical problems that occurred onboard?
RQ2: How text mining approaches are incorporated in the seafarer's health records?
RQ3: How seafarers' express problems associated with diseases?
RQ4: Is there any association between different symptom clusters knowledge extraction through health data mining?

The rest of the article was framed as follows. Section 2 provides the research background, including the main causes of health issues happening onboard. Section 3 covers the methods part including data extraction, interpretation, and experimental framework. Section 4 presents the interpretation of results and research outcomes. Section 5 provides a discussion on findings and existing literature. Finally, Section 6 provides study conclusions and the future scope of the present research.

2. Research background

Over the years, seafarers' health has to get great attention to international platforms, probably because of their significant contributions in world trades (Zhang en & Zhao, 2017). The international maritime organization (IMO) estimates that over 90% of global trade was done through ships (Zaman, Pazouki, Norman, Younessi & Coleman, 2017). Because of ship movements, there is a possibility of always having musculoskeletal strain, continuous noise, and vibration (Nittari et al., 2019). Despite this, being a seafarer getting injuries onboard is common as seafaring is a high-risk occupation when compared to others. As of this, onboard safety is one of the major concerns for both seafarers and ship owners.

Seafaring is the most hazardous occupation in terms of personal health and safety measures of seafarers. Since seafarers are exposed to continuous working hours on the sea, it makes them have different health problems. O Occupational injuries usually happen in small vessels, and may be often serious (Zytoon & Basahel, 2017). The special work challenges associated with the marine industry and ship activities were implied to risk involvement in accidents at work. These challenges might involve the behavioral safety of employees, workplace conditions, job type, onboard system management, and safety measures (Fabiano, Currò & Pastorino, 2004). Human conditions were playing an important role in the cause of occupational injuries. Moving from one-to-many places onboard can cause severe accidents, and the highest number of accidents was registered on deck.

When compared with onshore workers , factors such as rapid climate changes, a higher degree of air humidity, rain, wind, and intense solar radiation can cause mental and physical problems for seafarers (Wadsworth, Allen, Wellens, McNamara & Smith, 2006). Length in working hours, socio-psychological factors, and others may also influence the seafarers' health. Moreover, infectious and non-communicable disease burden and quality of services provided onboard may represent other problems in seafarers (Oldenburg, Baur & Schlaich, 2010).

Depression in sailors commonly associated their low mind-set with confinement from family, manager requests, inconvenience resting, and contract length (Mellbye & Carter, 2017). Decreasing seafarer's confinement from family is an inherently big challenge (though they had a video conferencing facility with family contacts), while other work elements might be limited with proper work-natural intercessions in a joint effort by ship owners.

Better social support and supervision while onboard are significant predictors of seafarers' mental health status rather than someone who had a history of mental disorder and poor education. At the same time, keeping physically active is also mandatory. For instance, providing guidance for suitable exercises or conducting sports activities on board could explore the release of happy hormones such as endorphin or serotonin (Battineni, Di Canio, Chintalapudi, Amenta and Nittari, 2019). Consequently, this will promote a sense of happiness and health. In the research of seafarers' happiness index, they addressed the direct connection between fitness and mental health (Kim & Jang, 2018). Therefore, it is always recommended to have an insight knowledge of factors that profoundly affect the behavioral and emotional well-being of seafarers.

On the other hand, prediction of patient emotion using TM methods completely depending on text data and could provide useful information based on EHR's of patients. Many studies highlighted the incorporation of TM on healthcare datasets. In Kukafka et al., authors developed a common language and framework for an International Classification of Functioning (ICF), health, and disability with advancing of TM techniques (Kukafka, Bales, Burkhardt & Friedman, 2006). It is also reported that with the application of natural language processing methods, it is easy to identify adverse trails related to central venous catheters (Penz, Wilcox & Hurdle, 2007). Similarly, Stusser and Dickey (2013) concluded TM and data mining in medical records can largely help to improve medical care and reduce costs. In this study, we aimed to recognize key pathologies that naturally occurred among sailors through TM approaches.

3. Methods

In this section, we summarize the random collection of medical documents from the C.I.R.M repository and an exploration of TM with sentiment techniques. The analysis was conducted by using R statistics 1.2.5 version including relevant packages like tm, ggplot2, and word cloud that are available on CRAN mirror. We collected seafarer's health documents from the C.I.R.M repository and preprocessed them for simulation purposes. Fig. 1 explains the experimental framework of TM with opinion mining approaches.

3.1. Experimental framework

The random selection of three-year patient documents was extracted. Each document contains textual information of the seafarer's medical problem and symptomatic data. We extracted 3112 documents available in the last three years (2018–2020). The documents are prepared in a CSV file format to ensure the importing of all the documents into the R platform.

3.2. Text preparation

Text preparation or preprocessing is a method, which transfers human text into a machine-readable format for further analysis. Text preprocessing contains many steps like text normalization, stemming, and tokenization.

3.2.1. Text normalization

In TM, it is necessary to conduct model training or algorithm with high amounts of data (Uysal & Gunal, 2014). For doing it, assembling frequent words into text corpus is commonly used in model benchmarking. This was done by the 'tm' package to do corpus building and is followed by removing of stop words (cleansing). Thereafter, we built term to document possibly as a special tokenizer. Each language has its

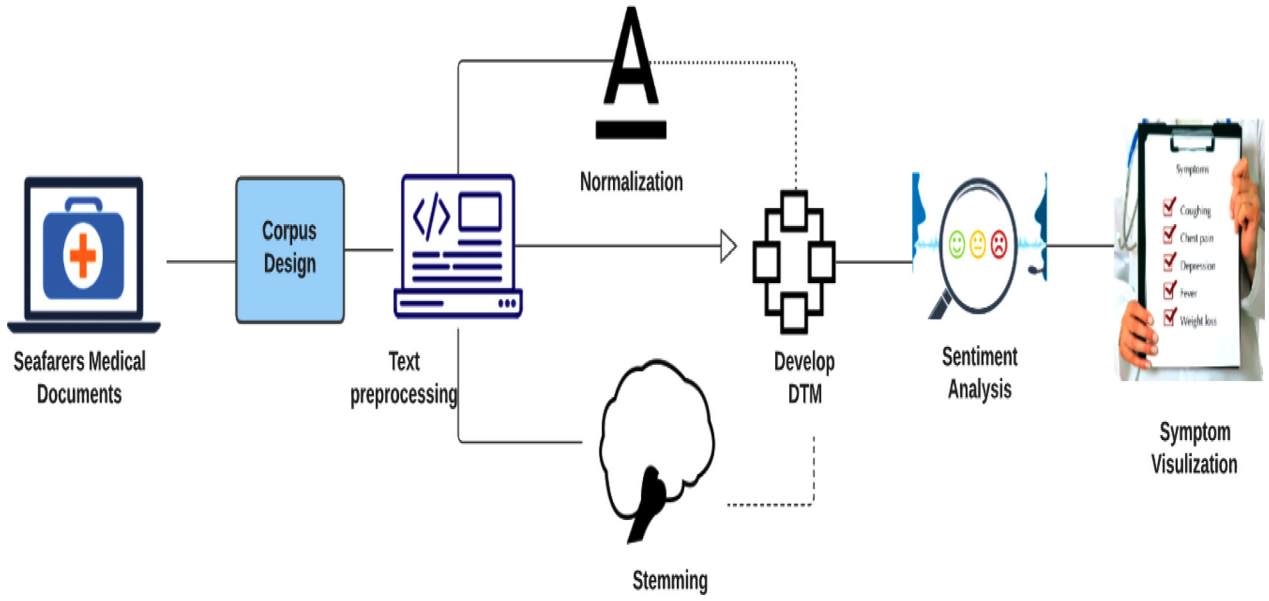

Fig. 1. An experimental framework.

Table 1
Sample medical abstracts in C.I.R.M repository.

Year	Case Number	Medical abstract
#2018	# 237	# Day 7/4, A Lituan Tankist, Slipping from The Ladder, While Going into The Deck, Falling With The Right Side Of The Thigh / Seat (Bringing A Significant Ecchymosis On That Side) And Rolling Was Throwing The Head Not In The Way Violent, Bringing Exercises On The Skin And A Small Ecchimosis In One Eye. He Did Not Report the Fact, Because, Other Than The Dams, He Feeled Well, In His Say. Yesterday, 8/4, I Received What Happened, Because I Had Headache And As Sensations Of Pressure / Pulsations. Mass Of Bags Of Ice Water. Some Heads Sometimes Slows Down And Sometimes And Stronger.
#2018	# 1352	# For Approximately A Week Has Present Abscess In The Inguinal Region Sin, Which Has Increased In Volume In The Last Days And Is Dolent To The Touch. Vital Parameters In The Standard.
#2019	#126	# Commander Informed That The Patient Disassembling From The Guard Shift Eaten A Sandwich Immediately Accusing After Acute Epigastric Pain. Patient Declares To Be The Carrier Of Iatale Hernia. A Buscopan Supposition Has Been Administered With Improvement Of Pain Symptomatology.
#2019	#2786	# A Favorable Opinion Is Required To Transfer Patient Sorentini Giulio, 58 Years, With Tracheo-Esophageal Fistula, From Crotone To Genoa, By Plane. The Patient Paz Would Be Accompanied By The Doctor.
#2019	#1985	# Captain Communicates Symptoms Of Maritime On Board, Lament Of Ocular Inflammation For More Than Twenty (20) Days. We Are Asked By A Medical Council
#2020	#36	# Gastro enteric Symptoms With Fever In Resolution After Drinking Water Not Certainly Drinkable
#2020	#101	# 29-Year Maritime Affected By About A Week Not Better Specified High Airway Infective Syndrome (Refer Cool), Dry Cough And Thermal Rise. 37.1 C
#2020	#1173	#30 Years Present Abdominal Skin Rash That Does Not Improve with Anti-Fungal Pomata. We Have No Images

individual corpus. In this study, collected patient records were available in the English language and further converted into data frames to build a corpus and mine similar symptom words. Table 1 presents sample patient comments that we received from the patient end. These medical abstracts or comments are aimed to diagnose patient problems but with full of noisy text, special characteristics, symbols, numbers, and buzzwords. All these were removed through normalization process.

3.2.2. Stemming

The follow-up to normalization, stemming is employed in cutting the beginning or end word by taking into account common prefixes and suffixes, which can be found in an inflected word (Kostoff, Toothman, Eberhart & Humenik, 2001). In this type of medical text, a number of inflected words, for example 'suffering', 'suffered' are identified through stemming and represent with the common word 'suffer'.

3.2.3. Tokenization

After normalization steps, a simple feature selection or symptom tokenization was done wit with the symptom words. Then, every single token can represent a symptom sparse matrix using the term frequency-inverse document frequency (t_f - id_f) described in Shafiei et al. (2007). We consider corpus C with N medical documents (d_j) where $j = 1,2, —, N\text{-}1, N$, and symptoms tokenized as term t. The t_f - id_f-weighing scheme considers the relative importance of individual symptoms in the given documents and assigns to term (t_k) as a weight in document d_j defined in Eq. (1).

$$t_f - id_f\left(tk_{,}d_j\right) = t_f\left(t_{k,}d_j\right) * id_f\left(t_k\right) \quad (1)$$

Where t_f ($t_{k,}$ d_j) represents symptom frequency (i.e., the number of times symptom occurrence in medical documents), id_f (t_k) represents inverse medical document frequency, and d_f (t_k) presents the number of documents containing symptom.

3.3. Develop DTM or TDM

A document term matrix (DTM) is a mathematical matrix explaining the symptom frequency in medical records. In DTM, rows represent collected medical documents and columns correspond to terms (symptoms). In contrast, rows in the term-document matrix (TDM) represent terms (symptoms) and columns represent documents. Table 2 summarizes outcome of the TDM simulation.

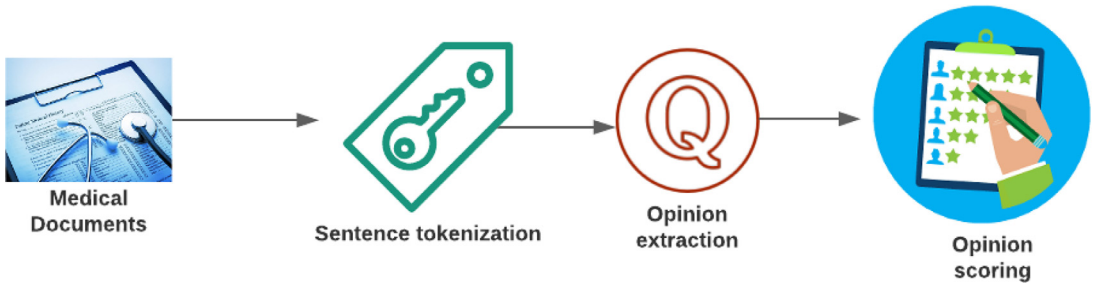

Fig. 2. The experimental approach of lexicon sentiment classification.

Table 2
TDM outcomes.

<<Term Document Matrix (terms: 2777, documents: 3112)>>	
Non-/sparse entries	1937/16,043
Sparsity	99%
Maximal term length	15
Weighting	term frequency (t_f)

Table 3
High associated symptom words correlated to diagnosis outcome.

Outcome diagnosis	Extracted terms (limit correlation 96%)			
Swelling	stiffness	Tenderness	redness	rash
Association	0.99	0.98	1	1
pain	swelling	Redskin	facture	bleeding
Association	1	0.99	1	1
abdominal	Swelling	Nausea	Fever	vomit
Association	1	1	0.96	0.98

3.4. Sentiment analysis

Sentiment classification, polarity measurement, and clustering of the entire corpus can be handled in this technique. We employed both lexicon and machine learning-based sentiment classification.

3.4.1. Lexicon sentiment classification

In this stage, it handles symptom classification and corpus clustering methods. We used the lexicon sentiment analysis (LSA) technique that defines individual terms search in the document and do weight calculation. LSA classifier is a sentiment scoring function that highlights all the symptoms in the corpus, no need for labeled data, and easy decision making through scoring functionality (Lan, Zhang, Lu & Wu, 2016). Fig. 2 explains the experimental approach of TM with sentiment analysis in real-time medical practice.

All the corpus words are compared with lexicon words, and the overall corpus sentiment score will be the difference between positive and negative assigned words. Thus, the polarity score of each diagnostic comment in the corpus is defined as

Sentiment score = $\sum_{j=1}^{n} Ps - \sum_{k=1}^{n} Ns$; Ps and N_s denote positive and negative signs
If sentiment score>0, the overall diagnosis outcome has positive
If sentiment score<0, the overall diagnosis outcome has negative
If sentiment score=0, overall diagnosis outcome has neutral

3.4.2. ML-based sentiment classification

In this study, sentiment classification was done through the Naïve Bayes classifier. It is a simple probabilistic classifier that works on Bayes theorem (Yaacob, Nasir, Yaacob & Sobri, 2019) and remains popular for text categorization, document evolution based on a characteristic of term frequency (t_f). It is one of the most popular among ML algorithms as it requires a small amount of trained data to estimate parameters required for classification. The basic idea behind Naïve Bayes' algorithm is to estimate category probabilities given in a text document through combined categorical probabilities and syntax is expressed below.

consider training dataset X for classes positive and negative
calculate the prior probability
For class positive = $\frac{\text{number of positive terms}}{\text{Total terms}}$
And
For class negative = $\frac{\text{number of negative terms}}{\text{Total terms}}$
calculate the total number of word frequencies (n_i) for both classes A(n_a) and B(n_b)
calculate the conditional probability of keyword occurrence
P (w_1/class positive) = word count/ n_i (positive)
P (w_1/class negative) = word count/ n_i (negative).....
P (w_{n-1}/class positive) = word count/ n_i (positive)
P (w_n/class positive) = word count/ n_i (negative)
Perform uniform distribution to avoid the zero-frequency problem
New document Z is classifies based on probability of both positive and negative groups P(Z/W)
$P(\frac{\text{Positive}}{W}) = P(\text{positive}) * P(\frac{\text{Word1}}{\text{Class Positive}}) * P(\frac{\text{Word2}}{\text{Class Positive}}) \ldots\ldots\ldots * P(\frac{\text{Wordn}}{\text{Class Positive}})$
$P(\frac{\text{Negative}}{W}) = P(\text{negative}) * P(\frac{\text{Word1}}{\text{Class negaive}}) * P(\frac{\text{Word2}}{\text{Class negaative}}) \ldots\ldots\ldots * P(\frac{\text{Wordn}}{\text{Class negative}})$
The class with the highest probability is the one new document Z has assigned

4. Results

4.1. Text mining of patient documents

A random collection of patient documents with seafarer's medical problems was done. Over the total corpus, we initiate tokenization, stop words and white space removal, and lower-case conversion to extract only individual symptomatic data sets. Thereafter, symptoms were represented by the collection of diagnostic word model with (t_f - id_f) weighing method to produce a sparse matrix that limits the outcome of at least three characters of diagnosis length. The sparse matrix removed 99% of sparse terms (Refer Table 2) and Fig. 3 presents the term frequency of symptom words (i.e., each word appears at least ten times; $w \geq 10$) over three-year patient documents. The highest number of symptom occurrences was generated with library package ggplot2, which also determines the relationship of individual terms of a specific diagnosis.

As mentioned, association functionality largely depends on the correlation of given terms and outcome diagnosis. Table 3 presents the association functions of swelling, pain, and abdominal problems are having the highest correlation of at least 96%. Correlation value limits between 0 (lower band) to 1 (higher band).

4.2. Word clouds

Text mining methods allow highlighting the high-frequency terms in documents or text paragraphs. World cloud or text cloud is a visual

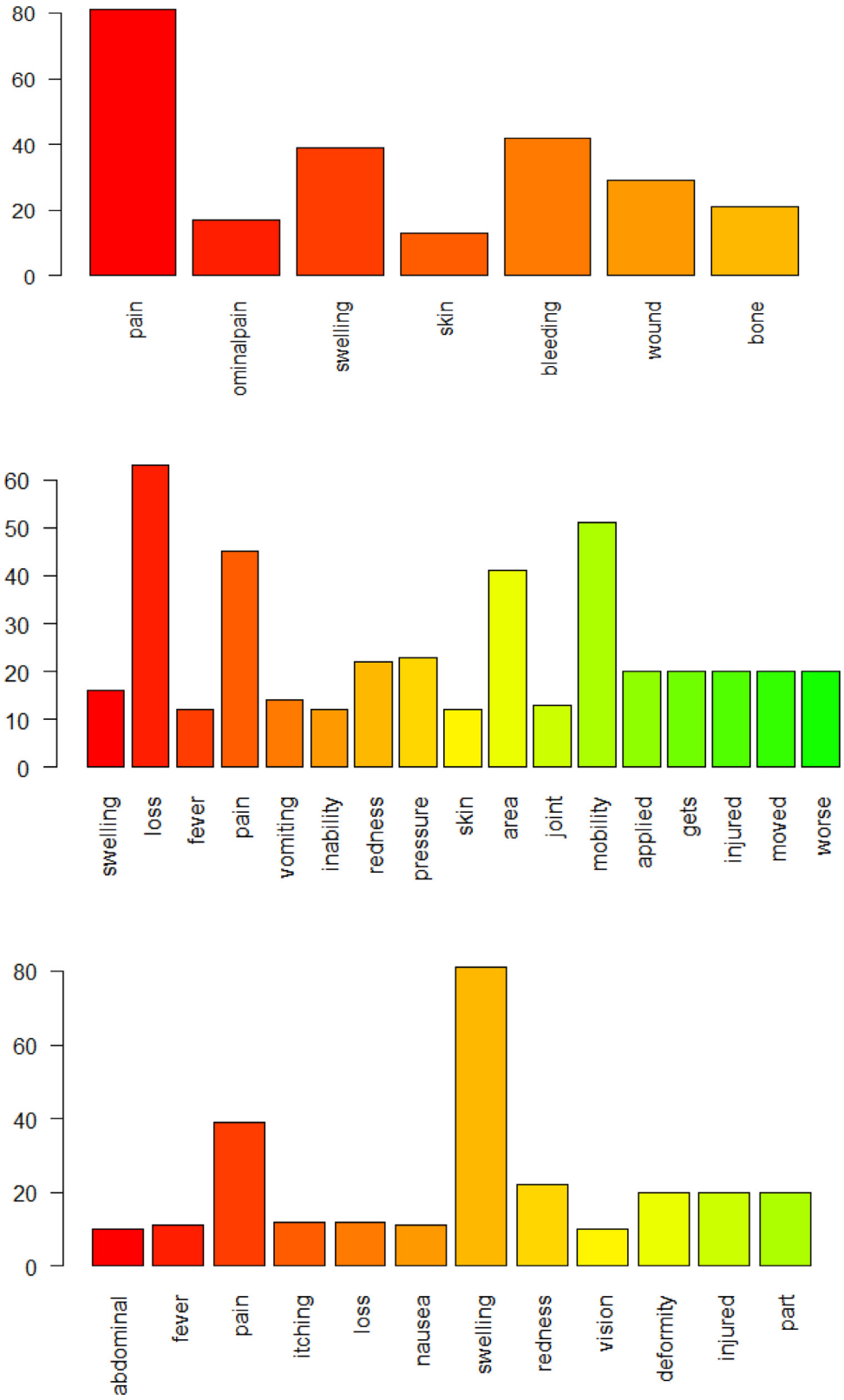

Fig. 3. Bar plot diagrams of highest frequency terms in three document groups (2018: top, 2019: middle, 2020: bottom).

representation of text data. These are visually engaging than manual data presentation. The three-year symptomatic (terms) word clouds of medical records are depicted in Fig. 4.

It is clear from the word cloud outcome, that most alignments of medical records are related to pain, swelling, bleeding, injuries, wounds, skin diseases, fever, abdominal pain, etc. In general, no seafarer is not likely to hurt himself or want to have injuries while working on merchant ships (Hystad, Nielsen & Eid, 2017). It is most common to have accidents during workplaces as a result of human errors, even if people follow preventive measures.

Fig. 4. Commonality word clouds of frequently occurred health problem terms (2018: left, 2019: middle, 2020: right).

Table 4
Summary of collected patient symptoms among retrieved documents.

Symptom group	Total number of documents Retrieved	Total number of patients symptoms
2018	1002	937
2019	1136	951
2020	974	889

Table 5
Sentiment scoring distribution of individual document groups (scores between −1 to 1 range considered as neutral opinion).

Score	Negative	Neutral	Positive	Row Total
2018 group	226	236	453	915
2019 group	161	399	413	973
2020 group	136	404	349	889
Total	523	1039	1215	2777

4.3. Sentiment analysis outcome

At first, we retrieved the patient documents, and then symptomatic extraction was done. Later, patient feedback is considered to conduct opinion-mining for defining positive, negative, and neutral sentiments. Thereafter, symptom frequency analysis was evaluated to identify popular disease terms and create their preassembled lexicon. Table 4 summarizes collected patient documents and symptoms. The preassembled lexicon was merged with medical library terms for sentiment classification. Lexicon scoring was performed over three sentiment groups (Fig. 5). The distribution of these scores can be observed in Table 5.

Overall sentiment of total three-year groups (2018, 2019, and 2020) has been positively achieved, with sentiment ratios (-ve: neutral: +ve) of 1:1:2, 1:2:2.5, and 1:3:2.5 respectively. All these outcomes explain patient comments about disease explanation are frequently matched positively with physician diagnosis.

In machine learning sentiment analysis, accuracy, precision, and recall were used to estimate the performance of opinion mining. Accuracy is defined as overall true outcomes of certain group model sentiments, and precision is the ratio of true positive sentiments to total sentiments of a particular group (Battineni, Sagaro, Nalini, Amenta & Tayebati, 2019). The performance metrics of the Naïve Bayes classifier on three document groups in a five-round experimental setup are indicated in Table 6.

A similar study of opinion mining to predict drug satisfaction levels among patients who experienced the effect of the drug was performed (Gopalakrishnan & Ramaswamy, 2017). They applied neural networks (NN), and support vectors (SVM) to illustrate the performance of two different drug groups. Radial basis neural networks produce better performance than SVM with an average precision of 88.6% for drug 1 and 89.8% for drug 2. In this study, we achieved over 80% of mean accuracy and precision that is well accepted in the medical domain.

5. Discussion

This study presents the importance of TM approaches in extracting clinical symptomatic information of seafarers, and patient condition and feedback were assessed through sentimental analysis. Because of the large medical data available either online or in clinical practices, it is difficult or a long-lasting process through conventional statistical methods to figure out accurate health data. On the other hand, online blogs or social media sites are not intended to disclose personal information due to low access, and privacy issues (Battineni et al., 2020). Hence, text mining methods are used to handle the clinical data and explore healthcare topics. TM is also helpful to rise the medical issues and patient opinion, and it is a special domain in exploratory analysis of text (Grover & Kar, 2017).

5.1. Contributions to literature

In medical communities, disease-related symptoms, health-related topics, and medical issues are mandatory for healthcare centers, physicians, and patients. The data availability thorough different platforms can change the study behavior of Information Systems (IS). This is high-

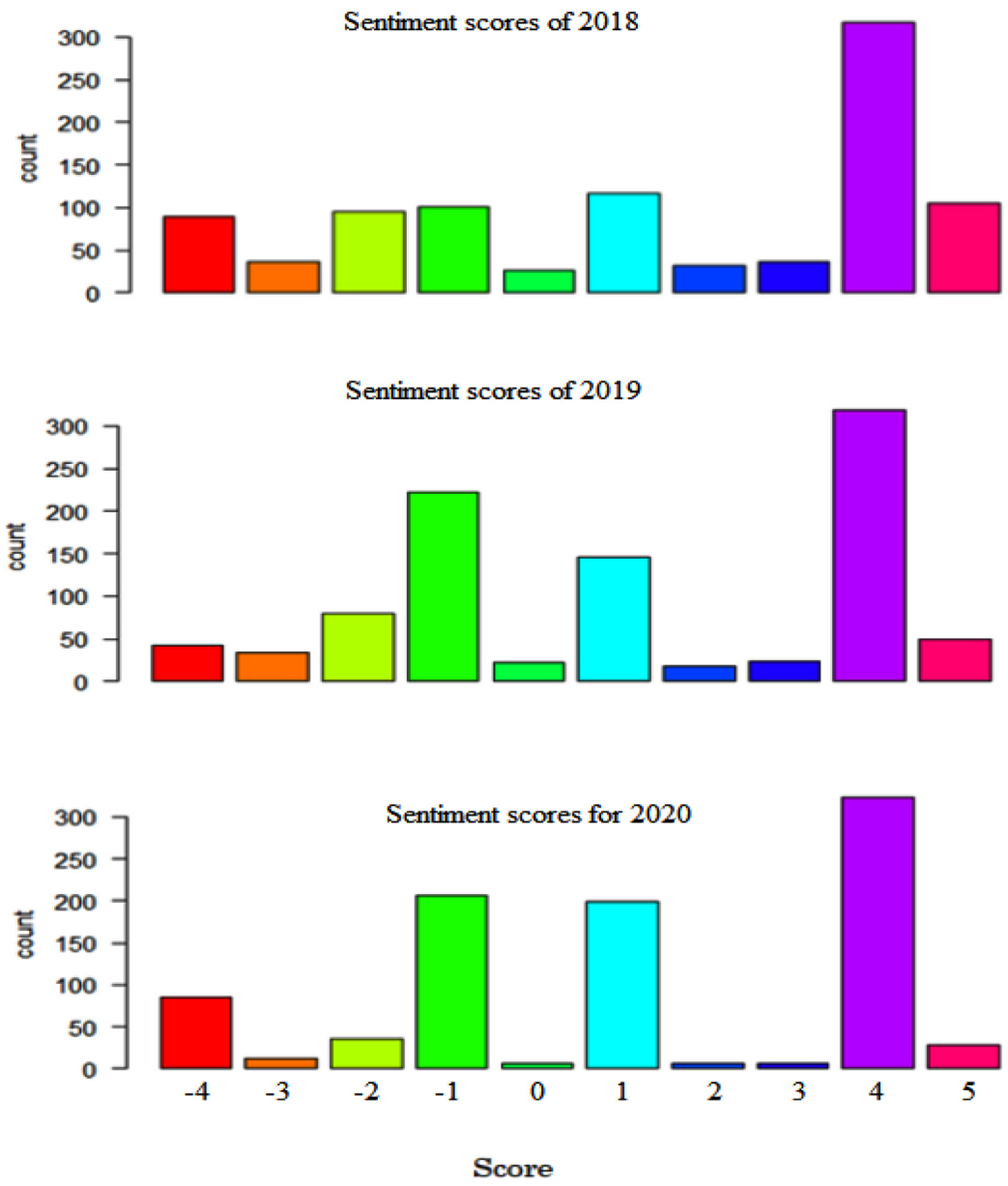

Fig. 5. Comparison of three symptom sentiment score groups [scale ranging from −5 (highly negative) to +5 (highly positive)].

Table 6
Accuracy comparison over entire corpus test patient group comments.

Experiment Number	Trained Patient comments	Negative Accuracy (%)	Precision (%)	Recall (%)	Neutral Accuracy (%)	Precision (%)	Recall (%)	Positive Accuracy (%)	Precision (%)	Recall (%)
1	10	51.2	52.3	50.6	53.7	52.9	53.1	50.4	51.2	53.2
2	20	58.1	61.8	62.4	60.2	63.6	62.4	57.3	59.8	58.3
3	30	64.8	63.1	64.6	67.4	69.1	68.7	63.1	60.6	62.6
4	50	74.9	76.2	75.2	76.6	73.2	74.3	73.3	76.1	71.8
5	100	83.6	80.3	81.6	85.4	87.8	86.8	81.7	82.4	83.4

lighted in Kar and Dwivedi (2020), described that text mining, sentimental analysis, image analytics and network science describe the insights of data science. The huge number of social media sites like Twitter or Facebook includes sentiment-based comments. To analyze this, opinion mining involves the TM to classify or evaluate the text source of sentimental content. Preliminary studies have highlighted the medical sentiments of clinical trials and online blog sources, also social media texts suggesting that they can capture the patient opinion in medical settings (Denecke & Deng, 2015). In an online medical society or social media networks, and healthcare organizations creates data, and they need patient's opinion for their sites. During the Arab spring event, TM with sentimental analysis can be a powerful predictive tool and successfully applies to extract social media events for sentiment classification (Akaichi, Dhouioui & Lopez-Huertas, 2013). In medical domain, sentiments with natural language models can classify patient comments on hospital experience.

Several studies were identified with the inclusion of sentimental analysis to find fraudulent information among health tweets. Some au-

thors have developed a method that makes it easy to predict and analyze health issues over social sites, and results highlight the correlation accuracy between HIV ailments is 98% (Mittal, Iqbaldeep, Pandey, Verma & Goyal, 2018). A standard model in the analysis of medical user opinion depends on the information available in social media was developed (Yang, Lee & Kuo, 2016). Detection of multiple forms of medical sentiments is useful because of the possible interference with a patient medical condition, medication, and treatment. Another study has suggested that the level of SA may depend on ontology's in diabetes (Salas-Zárate et al., 2017), authors adopted corpus from Twitter and labeled user option as positive, negative, and neutral, and achieved 82% of precision.

In this paper, we achieved a 96% of correlation between symptoms and diagnosis outcome with an imbalanced dataset. The sentiment analysis was validated with Naïve Bayes and produced over 80% of accuracy, precision, and recall. In our opinion, this study is an initiative for better understanding of seafarers' problems and allow other researchers to investigate this topic in maritime medicine.

5.2. Implications to practice

Seafarers are exposed to an environment of 24/7 stay duration onboard, which addresses some facts that can affect mental health. Some of these factors can be controlled over time, but others cannot be treated (Thomas, Sampson & Zhao, 2003). Therefore, understanding disease trends are relevant in obtaining positive implications. Although, safety has to be improved at workplaces, since seafarers undergo more occupational injuries, when compared with ashore workers (Fabiano et al., 2004). It is important to follow some preventive methods to overcome the accident rate onboard. Many researchers were tried to investigate the reasons behind occupational injuries at merchant ships and identify risk factors. In the research of Hansen, Nielsen & Frydenberg (2002), 209 among 1993 accidental cases were reported with permanent disability (of 5% or more), age was considered as a major risk factor of permanent disability. Besides, seafarers are away from home and had little contact with their friends and family (Hystad en & Eid, 2016). Cultural diversity between individual workers can make it more challenging to build good relationships, and consequently, it makes them feel more isolated (Håvold, 2007).

Textual analytics strategies in medical care are generally used to process medical text contents. In the past, medical texts were targeted on patient condition and diagnosis description. Contemporarily, user-generated clinical textual content either online blogs or social media contents is explored with opinion mining (Feinerer, Hornik & Meyer, 2008). TM techniques to this seafarer's medical text enables knowledge discovery that focus among marine community works. There are some reasons behind the involvement of doctors and patients in online communities, considering the fact of many doctors are not available to explain in-depth knowledge of patient symptoms and causes. Therefore, online communities like Twitter are highly effective for doctors and patients to have healthy decisions and benefit from medical treatment from information management and peers of social media sites.

6. Conclusion

We extracted three-year patient records to understand patient views and experiences through TM and sentimental analysis. This paper presents the intervention of TM on real-time medical records of seafarers and evaluated the performance of sentiment classification in terms lexicon scoring, and machine learning models. It provided a great advantage to healthcare centers like C.I.R.M for better understanding and visualization of seafarer's problems, monitor health records, and assess patient feedback. Accidents and gastrointestinal pathologies are commonly reported by seafarers, and it can be largely projecting through symptom aliments in word clouds.

Author contributions

Conceptualization, G.B. and N.C.; methodology, G.B.; formal analysis, N.C.; investigation and experiments, N.C.; resources, G.B. and G.G.S.; data curation, G.B. and N.C.; writing—original draft preparation, G.B.; writing—review and editing, G.B. and F.A.; supervision, F.A.; project administration, F.A.; funding acquisition, F.A.

Declaration of Competing Interest

The authors declare no conflict of interest.

Funding

This work is partly supported by a grant of the ITF Trust (No. 1276/2018) for the epidemiological analysis and data mining operations.

References

Ádám, B., Rasmussen, H. B., Pedersen, R. N. F., & Jepsen, en J. R. (2014). Occupational accidents in the Danish merchant fleet and the nationality of seafarers. *Journal of Occupational Medicine and Toxicology*. 10.1186/s12995-014-0035-4.

Akaichi, J., Dhouioui, Z., & en Lopez-Huertas, M.J. (2013). Perez, "Text mining facebook status updates for sentiment classification"„ doi:10.1109/ICSTCC.2013.6689032.

Bal, E., Arslan, O., & Tavacioglu, L. (2015). Prioritization of the causal factors of fatigue in seafarers and measurement of fatigue with the application of the lactate test. *Safety Science*. 10.1016/j.ssci.2014.08.003.

Battineni, G., et al. (2020a). Factors affecting the quality and reliability of online health information. *Digial Health, 6*, bll 1–11. 10.1177/2055207620948996.

Battineni, G., Di Canio, M., Chintalapudi, N., Amenta, F., & Nittari, G. (2019a). Development of physical training smartphone application to maintain fitness levels in seafarers. *International Maritime Health*. 10.5603/IMH.2019.0028.

Battineni, G., Sagaro, G. G., Chinatalapudi, N., & Amenta, F. (2020b). Applications of machine learning predictive models in the chronic disease diagnosis. *Journal of Personalized Medicine*. 10.3390/jpm10020021.

Battineni, G., Sagaro, G.G., .Nalini, C., Amenta, F., & en Tayebati, S.K. (.2019)."Comparative machine-learning approach: A follow-up study on type 2 diabetes predictions by cross-validation methods", *Machines*, doi:10.3390/machines7040074

Çakir, E. (2019). Fatal and serious injuries on board merchant cargo ships. *International Maritime Health, 70*(2), bll 113–118. 10.5603/IMH.2019.0018.

Denecke, K., & Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. *Artificial Intelligence in Medicine*. 10.1016/j.artmed.2015.03.006.

Fabiano, B., Currò, F., & Pastorino, R. (2004). A study of the relationship between occupational injuries and firm size and type in the Italian industry. *Safety Science*. 10.1016/j.ssci.2003.09.003.

Feinerer, I., Hornik, K., & Meyer, D. (2008). Text mining infrastructure in R. *Journal of Statistical Software*. 10.18637/jss.v025.i05.

Friedman, C., Rindflesch, T. C., & Corn, M. (2013). Natural language processing: State of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine. *The Journal of Biomedical Informatics*. 10.1016/j.jbi.2013.06.004.

Gopalakrishnan, V., & Ramaswamy, C. (2017). Patient opinion mining to analyze drugs satisfaction using supervised learning. *The Journal of Applied Research and Technology*. 10.1016/j.jart.2017.02.005.

Grover, P., & Kar, A. K. (2017). Big data analytics: A review on theoretical contributions and tools used in literature. *The Global Journal of Flexible Systems Management*. 10.1007/s40171-017-0159-3.

Hansen, H. L., Nielsen, D., & Frydenberg, M. (2002). Occupational accidents aboard merchant ships. *Occupational and Environmental Medicine*. 10.1136/oem.59.2.85.

Håvold, J.I. (.2007)."National cultures and safety orientation: A study of seafarers working for Norwegian shipping companies", *Work Stress*, doi:10.1080/02678370701424594.

Hystad en, S.W., .& Eid, J. (2016)."Sleep and fatigue among seafarers: The role of environmental stressors, duration at sea and psychological capital", *Safety and Health at Work*, doi:10.1016/j.shaw.2016.05.006.

Hystad, S. W., Nielsen, M. B., & Eid, J. (2017). The impact of sleep quality, fatigue and safety climate on the perceptions of accident risk among seafarers. *European Review of Applied Psychology*. 10.1016/j.erap.2017.08.003.

Kar, A. K., & Dwivedi, Y. K. (2020). Theory building with big data-driven research – Moving away from the 'What' towards the 'Why'. *The International Journal of Information Management*. 10.1016/j.ijinfomgt.2020.102205.

Khadjeh Nassirtoussi, A., Aghabozorgi, S., Wah, T.Ying, & en Ngo, D.C.L. (2014)."Text mining for market prediction: A systematic review", *Expert Systems with Applications*., doi:10.1016/j.eswa.2014.06.009.

Kim, J. H., & Jang, S. N. (2018). Seafarers' quality of life: Organizational culture, self-efficacy, and perceived fatigue. *International Journal of Environmental Research and Public Health, 15*(10). 10.3390/ijerph15102150.

Kim, Y. M., & Delen, D. (2018). Medical informatics research trend analysis: A text mining approach. *Health Informatics Journal*. 10.1177/1460458216678443.

Kostoff, R. N., Toothman, D. R., Eberhart, H. J., & Humenik, J. A. (2001). Text mining using database tomography and bibliometrics: A review. *Technological Forecasting and Social Change*. 10.1016/S0040-1625(01)00133-0.

Kukafka, R., Bales, M. E., Burkhardt, A., & Friedman, C. (2006). Human and automated coding of rehabilitation discharge summaries according to the international classification of functioning, disability, and health. *The Journal of the American Medical Informatics Association*. 10.1197/jamia.M2107.

Lan, M., Zhang, Z., Lu, Y., & en Wu, J. (2016).“Three convolutional neural network-based models for learning sentiment word vectors towards sentiment analysis”„ doi:10.1109/IJCNN.2016.7727604.

Mahdi, S. S., & Amenta, F. (2016). Eighty years of CIRM. A journey of commitment and dedication in providing maritime medical assistance. *International Maritime Health*. 10.5603/IMH.2016.0036.

Mellbye, A., & Carter, T. (2017). Seafarers' depression and suicide”. *International Maritime Health*. 10.5603/IMH.2017.0020.

Mittal, M., Iqbaldeep, K., Pandey, S. C., Verma, A., & Goyal, L. M. (2018). Opinion mining for the tweets in healthcare sector using fuzzy association rule. *EAI Endorsed Transactions on Pervasive Health and Technology*. 10.4108/eai.13-7-2018.159861.

Nittari, G., et al. (2019). Design and evolution of the Seafarer's health passport for supporting (tele)-medical assistance to seafarers. *International Maritime Health*. 10.5603/IMH.2019.0024.

Oldenburg, M., Baur, X., & Schlaich, C. (2010). Occupational risks and challenges of seafaring. *Journal of Occupational Health*. 10.1539/joh.K10004.

Penz, J. F. E., Wilcox, A. B., & Hurdle, J. F. (2007). Automated identification of adverse events related to central venous catheters. *The Journal of Biomedical Informatics*. 10.1016/j.jbi.2006.06.003.

Rathore, A.K., .Kar, A.K., .& en Ilavarasan, P.V. (.2017).“Social media analytics: Literature review and directions for future research”, *Decision Analysis*., doi:10.1287/deca.2017.0355

Salas-Zárate, M. D. P., Medina-Moreira, J., Lagos-Ortiz, K., Luna-Aveiga, H., Rodríguez-García, M. Á., & Valencia-García, R. (2017). Sentiment analysis on tweets about diabetes: An aspect-level approach. *Computational and Mathematical Methods in Medicine*. 10.1155/2017/5140631.

Shafiei, M. et al.,(2007). “Document representation and dimension reduction for text clustering”„ doi:10.1109/ICDEW.2007.4401066.

Stusser, R. J., & Dickey, R. A. (2013). Quality and cost improvement of healthcare via complementary measurement and diagnosis of patient general health outcome using electronic health record data: Research rationale and design. *The Journal of Medical Systems*. 10.1007/s10916-013-9977-9.

Thomas, M., Sampson, H., & Zhao, M. (2003). Finding a balance: Companies, seafarers and family life. *Maritime Policy & Management*. 10.1080/0308830032000051630.

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. *Information Processing and Management*. 10.1016/j.ipm.2013.08.006.

Wadsworth, E. J. K., Allen, P. H., Wellens, B. T., McNamara, R. L., & Smith, A. P. (2006). Patterns of fatigue among seafarers during a tour of duty. *American Journal of Industrial Medicine*. 10.1002/ajim.20381.

Westlund, K., Attvall, S., Nilsson, R., & Jensen, O. C. (2016). Telemedical Maritime Assistance Service (TMAS) to Swedish merchant and passengers ships 1997–2012. *International Maritime Health*. 10.5603/IMH.2016.0006.

Yaacob, W. F. W., Nasir, S. A. M., Yaacob, W. F. W., & Sobri, N. M. (2019). Supervised data mining approach for predicting student performance. *Indonesian Journal of Electrical Engineering and Computer Science*. 10.11591/ijeecs.v16.i3.pp1584-1592.

Yang, F. C., Lee, A. J. T., & Kuo, S. C. (2016). Mining health social media with sentiment analysis. *The Journal of Medical Systems*. 10.1007/s10916-016-0604-4.

Zaman, I., Pazouki, K., Norman, R., Younessi, S., & en Coleman, S. (2017).“Challenges and opportunities of big data analytics for upcoming regulations and future transformation of the shipping industry”„ doi:10.1016/j.proeng.2017.08.182.

Zhang en, P., & Zhao, M. (2017).“Maritime health of Chinese seafarers”, *Marine Policy*„ doi:10.1016/j.marpol.2017.06.028.

Zhang, Y., Chen, M., & Liu, L. (2015). A review on text mining. In *Proceedings of the IEEE international conference on software engineering and service sciences, ICSESS: 2015* (pp. bll 681–685). Nov. 10.1109/ICSESS.2015.7339149.

Zytoon, M. A., & Basahel, A. M. (2017). Occupational safety and health conditions aboard small- and medium-size fishing vessels: Differences among age groups. *International Journal of Environmental Research and Public Health*. 10.3390/ijerph14030229.