bioengineering

MDPI

Article

LASSO Regression Modeling on Prediction of Medical Terms among Seafarers' Health Documents Using Tidy Text Mining

Nalini Chintalapudi [1,*], Ulrico Angeloni [2], Gopi Battineni [1], Marzio di Canio [1,3], Claudia Marotta [2], Giovanni Rezza [2], Getu Gamo Sagaro [1], Andrea Silenzi [2] and Francesco Amenta [1,3]

1 Clinical Research Centre, School of Medicinal and Health Products Sciences, University of Camerino, 62032 Camerino, Italy; gopi.battineni@unicam.it (G.B.); marzio.dicanio@unicam.it (M.d.C.); getugamo.sagaro@unicam.it (G.G.S.); francesco.amenta@unicam.it (F.A.)
2 General Directorate of Health Prevention, Ministry of Health, 00144 Rome, Italy; u.angeloni@sanita.it (U.A.); c.marotta@sanita.it (C.M.); g.rezza@sanita.it (G.R.); a.silenzi@sanita.it (A.S.)
3 Research Department, International Radio Medical Centre (C.I.R.M.), 00144 Rome, Italy
* Correspondence: nalini.chintalapudi@unicam.it; Tel.: +39-35-33776704

Abstract: Generally, seafarers face a higher risk of illnesses and accidents than land workers. In most cases, there are no medical professionals on board seagoing vessels, which makes disease diagnosis even more difficult. When this occurs, onshore doctors may be able to provide medical advice through telemedicine by receiving better symptomatic and clinical details in the health abstracts of seafarers. The adoption of text mining techniques can assist in extracting diagnostic information from clinical texts. We applied lexicon sentimental analysis to explore the automatic labeling of positive and negative healthcare terms to seafarers' text healthcare documents. This was due to the lack of experimental evaluations using computational techniques. In order to classify diseases and their associated symptoms, the LASSO regression algorithm is applied to analyze these text documents. A visualization of symptomatic data frequency for each disease can be achieved by analyzing TF-IDF values. The proposed approach allows for the classification of text documents with 93.8% accuracy by using a machine learning model called LASSO regression. It is possible to classify text documents effectively with tidy text mining libraries. In addition to delivering health assistance, this method can be used to classify diseases and establish health observatories. Knowledge developed in the present work will be applied to establish an Epidemiological Observatory of Seafarers' Pathologies and Injuries. This Observatory will be a collaborative initiative of the Italian Ministry of Health, University of Camerino, and International Radio Medical Centre (C.I.R.M.), the Italian TMAS.

Keywords: seafarers; text mining; lasso regression; disease mapping; correlations

Citation: Chintalapudi, N.; Angeloni, U.; Battineni, G.; di Canio, M.; Marotta, C.; Rezza, G.; Sagaro, G.G.; Silenzi, A.; Amenta, F. LASSO Regression Modeling on Prediction of Medical Terms among Seafarers' Health Documents Using Tidy Text Mining. *Bioengineering* **2022**, *9*, 124. https://doi.org/10.3390/bioengineering9030124

Academic Editor: Christoph Herwig

Received: 7 February 2022
Accepted: 16 March 2022
Published: 17 March 2022

1. Introduction

Seafarers are regularly on the move and working at sea. As a result of long-term voyages, many international seafarers are away from their friends and families for at least six months per year [1]. The shipping industry is popular and used by 88% of world trade. However, it has a higher mortality rate, an injury rate, and a disease rate than land-based workers [2]. There are many risks associated with working at sea, including climate changes, seawater, humidity, and sun exposure [3,4]. Seafarers' health and living conditions are affected by their working environment. Medical issues onboard are explained via telephone or internet by the captain in consultation with Telemedical Maritime Assistance Services (TMAS) doctors onshore. Health services are provided based on the severity of the case. Later, patient data are stored as digital health documents, which are created as text records [5,6].

Data analysts including those from different fields are often tasked with analyzing text or other unstructured data [7]. From the seafarers' telemedicine data, we can extract pathology information, symptoms information, patient condition info, prescriptions from

doctors, etc. We can use these data to make better medical decisions. For doing, we need advanced computational tools and skills to analyze unstructured text [8]. One way to learn more is to use word frequency analysis. Sentiment analysis allows one to extract mapping words and emotions from symptomatic words. As a result, we can develop a basic understanding of Text Mining (TM) approaches, which are widely applied to retrieve symptomatic data [9]. Medical pattern recognition software that converts texts to natural language using medical directories, algorithms, and information knowledge [10,11]. The use of TM techniques has proven to be beneficial in detecting depression symptoms with comprehensive diagnostic accuracy, according to studies in mental health [12]. According to [13], the integration of TM knowledge with neurological data can be used to detect neurological diseases and syndromes based on annual data.

Natural language processing (NLP) and machine learning (ML) techniques are used in sentiment analysis (also known as opinion mining) [14]. Medical sentimental analysis is used for the evaluation of medical records and automated decision support systems [15,16], as well as for assessing a patient's emotions and sentiments based on medical history [17]. Likewise, a hybrid text model has been designed to help healthcare workers diagnose diseases such as diabetes and dementia [18,19]. Social media and the web can be analyzed with TM to identify knowledge about diabetes diagnosis and treatment. In this study, patterns in the web and social media text were analyzed to discover previously overlooked diagnoses or conditions of diabetes via comparison with those from standard diabetes treatment and diagnosis [19].

The literature on TM primarily focuses on sentiment analyses, such as emotion recognition, handwritten/typed text analysis, and data visualization. The authors of [20] provided a list of all methodologies and approaches for performing sentiment analysis based on three categories: machine learning, dictionaries, and ontologies. The study provides a brief introduction to opinion mining issues such as data sparsity, binary classification, and polarity shift problems in distinct domains. Various analytical algorithms are discussed to extract the text from mixed documents with the typed and handwritten text [21]. The study examines the sentimental analysis of medical treatment in detail and shows new challenges and possibilities in the medical field [22].

A previously published study [9] was extended in the present paper to highlight the importance of tidy TM in presenting symptomatic words of diseases common to seafarers. To represent the patient's emotions or feedback, we applied sentimental analysis to medical abstract documents with disease names and symptoms. Seafarers' medical documents are not aligned with the TM documentation. By mapping tidy TM symptom words to common diseases occurring onboard, we attempted to circumvent this limitation. The present study is the first comprehensive analysis of seafarers' diseases that has been conducted by researchers who have an understanding of maritime medicine.

We can understand a subject's emotions and behavior when studying such medical documents with text since emotions play an active role in human behavior. A domain-specific corpus from the digital health database contains clinical documents, which is critical for decision-making [15]. Text mining algorithms are applied to text containing seafarers' medical documents in the analysis of text containing TM of telemedical documents. The least absolute shrinkage and selection operator (LASSO) regression models were applied for text classification as well as for defining word relevance through this variable selection.

2. Materials and Methods

Methods used for the analysis included document collection, pre-processing, and sentiment analysis. The frequency of disease terms was determined by analyzing sentiment in the Bing lexicon. By using these techniques, we estimated the most commonly used words among seafarers.

2.1. Data Collection

Medical text data of seafarers were examined from 2006 to 2021 and data was analyzed among 41,292 seafarers who got telemedical assistance through the International Radio Medical Centre (C.I.R.M.). The Centre establishes digital medical files for each case after it makes contact with the ship and updates them and this study analysed these files.

Data for the last 15 years (2006–2021) were extracted from 41,292 text documents containing patient information, a seafarer sending a message (Tx), and a doctor receiving a message (Rx). Messages TX include ship name and radio call sign, position, destination port, estimated time of arrival, course, speed, patient's age, nationality, qualification, vital signs like breathing, pulse, temperature, and blood pressure, symptoms of localized pain, medical history, and medicines available on board. The message Rx contains a doctor's questions, treatment, diagnosis, and diet and prevention instructions, as well as all the patient's treatment information.

Documents include text data such as symptomatic information, doctor prescriptions, treatment information, medication details, etc. C.I.R.M. physicians classified diagnoses according to the International Classification of Diseases (ICD)-10 (WHO, 2007). Health management, epidemiology, and clinical analysis rely on this standard. Table 1 provides an example of medical abstracts for treatments. For a smooth experimental setup in the R framework, all text data were prepared as a CSV document.

Table 1. Sample of medical abstracts and treatment of given diagnosis.

Year	Case Number	Diagnosis	Medical Abstract	Suggested Treatment
2006	88	Abdominalgia	Mild pain in the lower part of the stomach and temperature.	Discontinue aspirin. Keep patient bed rest in the most comfortable position. Apply an ice bag wrapped in a cotton cloth on the painful area if it relieves pain.
2008	17	Acute Gastritis	The patient said he has stomach pain; he has a history of hyperacidity.	Keep patient rest in a sitting position. Give buscopan one tablet every six hrs. give antacid every six hrs. Give omeprazole one tablet every twelve hrs. Light boiled food diet with a large intake of mineral water. Give news in twelve hrs.
2009	151	Allergic Reaction	The rash on a body appears in various places namely round an eye, bridge of the nouse, behind an ear, on a breast and a back, on a neck and hands.	Keep resting cotton loose-fitting clothes. continue ciprofloxacin milligram. Cetrizine or chlophenaramine. Boiled food diet with abundant water. Avoid all contact with cargo.
2013	597	Fever	Stomach pain with loose motions, mainly at night. Burning sensation during urination especially during evenings when the fever sets in.	Keep bed rest far from air draughts and extremes of temp. Apply ice bag wrapped in a cotton cloth on the head when temperature rises above 39 °C Continue Paracetamol, Ciprofloxacin, continue also Buscopan.
2014	1042	Haemorrhage	Patient with profuse blooding from yesterday at the gingival level (maybe the presence of abscess) and of the urinary tract. He has lost knowledge several times yesterday and today, already underway in fluid therapy.	Continue fluid therapy with Ranitidine fl inside the flexo, Tranexamic acid is not available onboard. Give as antibiotic Amoxicillin 1 g CPR if not allergic. Urgent disembarkation should be organized with a faster vehicle.
2016	197	Anxious-Depressive Syndrome	Please note that for the last two days the patient had been complaining of improper sleep. He reported that he was feeling a little depressed. He also reported that he does not feel capable of keeping navigational watches during hours of darkness as it gives him a feeling of loneliness.	Keep at rest in the bed or armchair as he prefers but, in any case, under continuous control by a friendly person. Remove from his cabin dangerous objects (knives, forks, glasses, razor blades, belts, shoelaces, dangerous drugs, gas lighters, anything through which he can injure himself or other people).
2020	746	Foreign Body	One of the people in the crew has swollen right eye. He got some foreign dust particles inside his eye, he rubbed his eye with his dirty hands, the eye started swelling and itching. We gave him an eyewash and suggested washing the eye regularly. Looks like due to rubbing the eye, he developed an eye infection. Kindly advise treatment we can give.	Keep rest not necessary in bed in a semi-dark room. Wash accurately's the eye with sterile saline solution or e Optrex or other eye leashes. Then when dry apply eye antibiotic ointment and cover with a sterile or light bandage.
2021	54	Odontalgia	Complain regarding the patient's tooth on the lower left molar. It was found out that the filling was been detached which causes pain.	Keep at rest. Apply inside the tooth cavity a small ball of cotton wool soaked in clove oil. Administer Paracetamol one 500 mg tablet every 6 h and Co-amoxiclav one gram tablet every 12 h. A light diet with easily chewable foods and a large intake of liquids.

2.2. Corpus Pre-Processing

Data cleaning ensures that user data is consistent, reliable, and accurate, and the text should be organized logically, especially for in-text data. We come across questions regarding punctuation, abbreviations, and contractions after reading the corpus. The removal of stop words and stem words, as well as the treatment of lower- and uppercase letters, is also needed. In tidy TM, clean_corpus is a function within the tidytext package that helps process the corpus [23]. With some default tools, like strings (for text cleaning), this package can convert upper case letters into lower case.

2.3. Tidy Text Mining and Packages

We can manage text simply and easily by using tidy data standards. The tidy data structure has the variables as columns, observations as rows, and each observational unit type as a table [24]. Thus, a tidy text format appears as a table with one token per row. For text analysis, tokens are semantically meaningful words, and tokenization is the process of dividing text data into tokens.

With tidy TM, the token is stored in every row, which can be a single word, a sentence, an n-gram, or a paragraph. The 'tidytext' package provides the functionality of tokenization with commonly encountered text units. In this package, there is no requirement that the user maintains a clean text format at all times. Using dplyr and other tidy tools, the text is processed, filtered, and imported, and then data is converted into a document-term matrix (DTM) for use in ML applications, and ggplot2 can then be used to visualize and interpret these models [25]. Figure 1 shows the flowchart representation with help of tidy data principles.

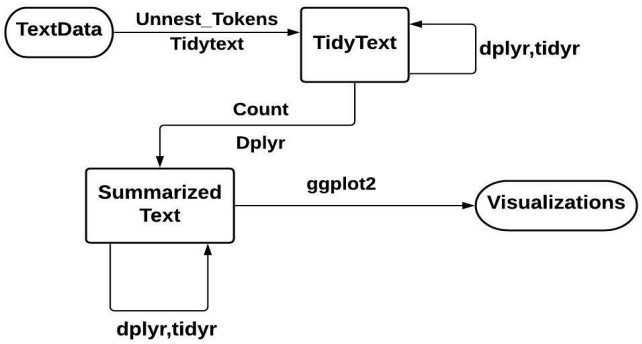


Figure 1. Flowchart representation of typical text analysis using principles of tidy data.

2.4. Sentimental Analysis

In order to make product sales more effective, a company manager might want to find out if the product reviews are positive or negative. It is possible to examine text sentiment using a word's combination and the sentimental content of the entire text by analyzing the sentiment of the singular words. With tidy TM, lexicon-based sentiment analysis is frequently used to calculate sentiment distributions based on lexicon alignments [26]. This semantic alignment can be negative, positive, or neutral. The lexicons dictionary can be formed either manually or automatically. Figure 2 illustrates the architecture of lexicon-based analysis.

A lexicon-based analysis determines the semantic orientation of the text by looking at adverbs and adjectives. This can be converted into a single score for the entire value in the final assessment. Three general lexicons can be found in tidy TM, namely AFINN, BING, and NRC. In the AFINN lexicon, sentimental scores range from -5 to 5, with negative scores for negative sentiment and positive scores for positive sentiment. As with the Bing lexicon, the NRC lexicon categorizes sentiments equally into yes/no categories as positive/negative.

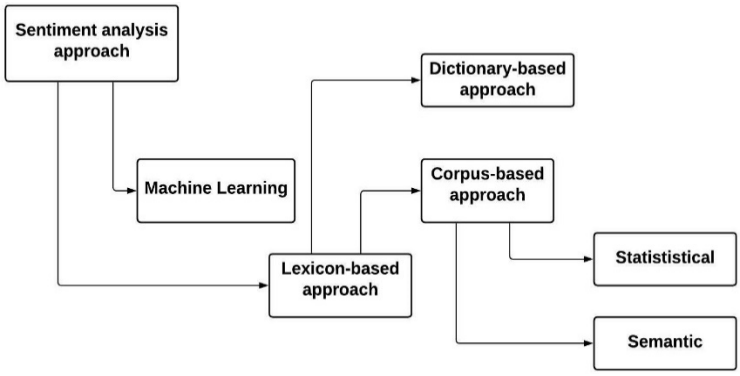


Figure 2. Lexicon-based sentimental analysis architecture for text documents.

2.5. Calculation of Word and Document Frequency (TF-IDF)

The term frequency-inverse document frequency (TF-IDF) indicates how relevant a medical term is to a particular telemedical document. A TF-IDF is a method for measuring the importance of a specific word in a document in comparison to the total number of documents. It is calculated as follows:

$$\mathrm{TF(t)} = \frac{\text{Number of times term t appears in a document}}{\text{Total number of terms in}} \quad (1)$$

where TF(t) presents how often a particular term appears in a document. Similarly, the inverse document frequency (IDF) is a metric value representing the information provided by a particular word. Mathematically it is presented as a fraction of logarithmically inverse documents that contain the words. It is mathematically denoted as

$$\mathrm{IDF}\ (\mathrm{t}) = \log\ \frac{\text{number of documents}}{\text{number of documents containing term}} \quad (2)$$

Simply, Equation (2) is logarithm of document number in corpus (nominator) divided by number of documents where particular term appears (denominator). It is likely to get words such as 'is', 'are', 'the', and 'an' in the calculation of most frequent words in the corpus. By removing buzz and stop words from the medical abstracts, we will get words like 'seafarers', 'accidents', and 'pathologies'. Thus, the TF-IDF metric measures the frequency of terms and weights them by how rarely they are used. The term frequency in medical documents refers to how frequently a particular term appears within an individual document. Seafarers' medical records are often referred to as 'frequency' in general. The terms 'seafarers', 'accidents', and 'pathologies' occurred very often in this work, and they are very frequently used in a given document, resulting in its low rating for TF-IDF.

2.6. Word Clouds

One way to visualize the high probability words in the text using TM packages is by using word clouds [26]. The word clouds can also be called text clouds and are created with the TM package (tm), and word cloud creator package (word cloud) [27], both of which are available in R for helping visualize words in text quickly. In the word cloud, the size represents the frequency of words. They may appear like a visualization of popular positive and negative words, but the size of words cannot be compared to the sentiment they convey.

2.7. LASSO Regression Model

LASSO takes advantage of shrinkage to accomplish linear regression. When a data value shrinks towards a central point, such as a mean, shrinkage occurs. As a result, it is well suited to models with high levels of multicollinearity. It also allows automated parts of model selection, such as parameter elimination and variable selection [28].

The LASSO regression model is being used more and more in medical diagnosis to predict disease outcomes and side effects. The model has been applied to brain modeling [29], biomarker selection [30], healthcare cost prediction [31], and early detection of cardiovascular diseases [32]. The classification of medical documents is widely used in healthcare, but few studies have been conducted on it. Our work demonstrates how LASSO is applied to text data using the principles of tidy data. This model extends supervised machine learning to text classification.

LASSO problems are quadratic programming problems that aim to minimize. In statistics, it was written as

$$\sum_{i=1}^{n}\left(yi-\sum_{j}x_{ij}\beta_j\right)2+\lambda\sum_{j=1}^{p}\left|\beta_j\right| \quad (3)$$

The above equation is the same as the minimization of the sum of squares with constraint $\Sigma\left|\beta_j\right| \leq s$. To interpret the model easily, some of β s can be shrunk to almost zero and results regression model to do easy interpretation. Here λ is a tuning parameter (i.e., amount of shrinkage). When $\lambda = 0$, no parameters are eliminated. When it increases, bias increases and when it decreases, variance increases.

2.8. Model Training and Evaluation

After the data were ready, they were divided into training and testing sets. The data are used both for building the model and evaluating its performance. LASSO regularization was applied for our logistic regression model with the glmnet package, and it can help to detect keywords in a prediction. We compare the LASSO model performance with two supervised models, namely Support Vector Machines (SVM) [33] and Random Forest (RF) [34]. We then validate the model using cross-validation (CV). As a resampling method for statistical analysis, this approach is known as rotation estimation. Therefore, in order to implement the CV technique, the data sample is segmented into different subsets. The analysis is performed on a subset called a training set. The results are then verified on the other subset called a testing set or a validation set [35]. By creating a data frame that can be displayed to each document in the dataset, the model's performance is evaluated by applying tidy data principles. Based on the percentage of correctly classified outcomes over the total outcomes, classification performance is calculated.

3. Results

A benefit of using tidy data is sentiment analysis, which can be performed as an inner join. The ability to perform sentiment analysis as an inner join is another practical example of using TM as tidy analysis, similar to removing stop words as an antijoin operation.

3.1. Sentimental Analysis

As part of our study, we examined how sentiments of each symptom varied across categorical diseases of ICD 10. We first determine the sentiment scores for each symptomatic word by using Bing lexicon and inner join functions. Figure 3 shows how sentimental scores (Y-axis) and the plotting of medical documents for certain diseases change and become more positive or negative over time (X-axis).

The data frame includes both a word and a sentiment, so we can easily determine the number of words that contribute to each statement. Figure 4a shows sentiment visualization in the form of word clouds, while Figure 4b shows the word distribution count for positive and negative sentiments.

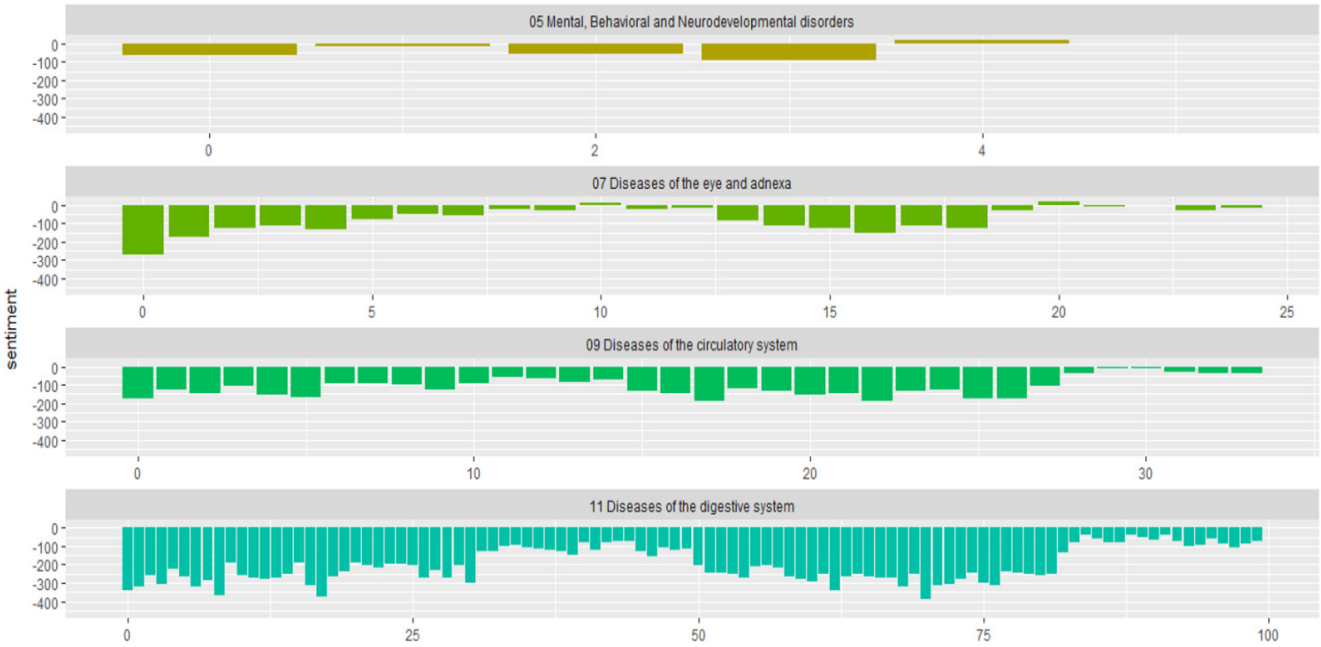


Figure 3. Lexicon based sentimental scores of the ICD 10 disease types (this is the plot of each disease sentiment changes towards more negative or positive over the times appearing in a dataset).

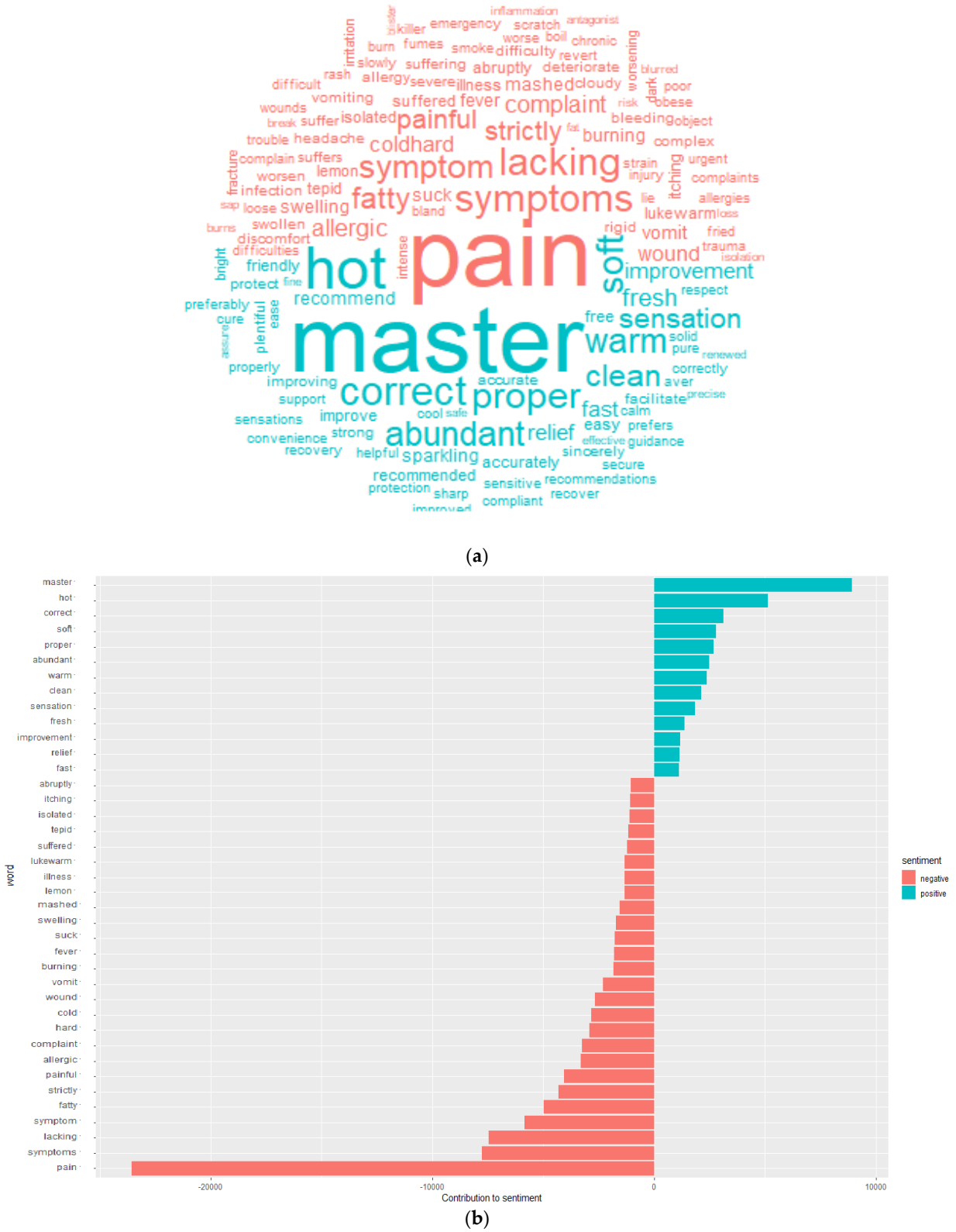
3.2. TF-IDF Calculation

By reducing the weight of commonly occurring words and boosting the weight of words that are less frequently used in the document corpus, TF-IDF identifies medical keywords associated with each disease category. We examined a large number of medical documents of seafarers, identifying symptoms of several disease categories. Using the ICD-10, we classified the documents into 22 groups. Most of the important words did not appear in all the categories. Figure 5 shows how disease documents categorize the major keywords. ICD code 05 (mental, behavioral and neurodevelopmental disorders) contains the keywords friendly, excited, dangerous, violent and depression with TF-IDF scores 0.001174, 0.001323, 0.001260, 0.001169 and 0.000754. Seafarers tend to suffer from anxiety and depression more often than onshore workers due to their long days away from their families [36]. As a result, tidy TM packages are able to identify a given disease's most common symptoms.

3.3. Bigrams and Correlations

Using Bigrams, we can also show how medical words relate to one another. Bigrams are visual representations of words arranged in a graph or network. In this study, we considered a graph with multiple nodes. Graphs were created using the igraph package, which has powerful manipulation and analysis functions. In Figure 6, we can see a relationship between different words in the medical records of seafarers. The diet nodes are connected to words including fatty, spicy, semiliquid, coffee, cigarettes, spices, etc. There are also triplets with similar meanings ('cloth', 'cotton', or 'woolen').

Additionally, correlation establishes a link between dependent and independent words. Having a negative correlation indicates a decrease in what we are measuring. This enabled us to determine which words are closely related to a particular medical term. In this experiment, we choose some popular words and find other words that are most related to them (Figure 7). There is 92.3% correlation between the word 'intestinal' and the word 'urinary'.



(**a**)

(**b**)

Figure 4. (**a**). Word cloud picturization of positive (green) and negative (red) sentimental words (most of the word alignments are associated with words pain, master, symptoms, hot, correct, lacking etc.). (**b**). Word count that contributes both negative and positive sentiments; the 'pain' word had the highest negative sentiment count (23,557) and the 'master' word has the highest positive sentiment count (8935).

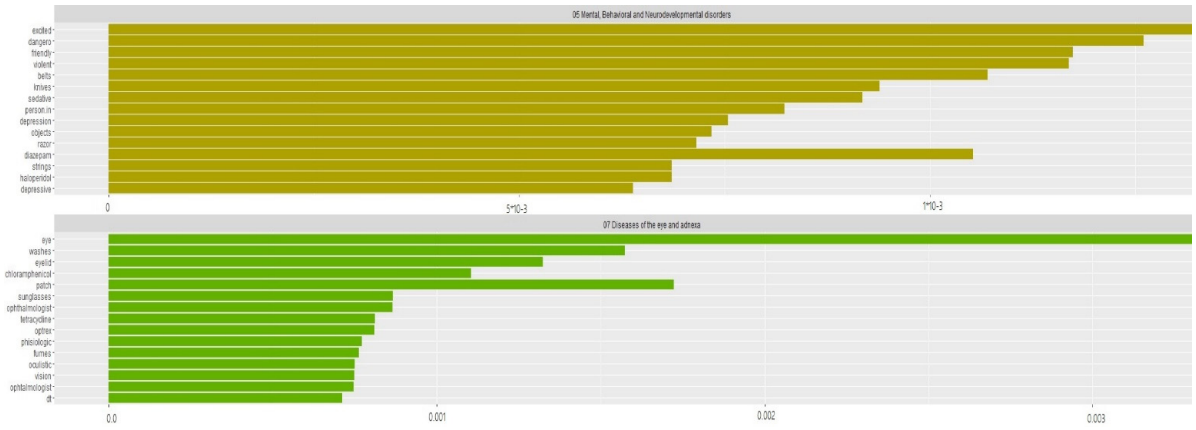


Figure 5. TF-IDF word count for mental health and eye diseases category; the highest frequency symptomatic words calculated by TF-IDF are vital to disease diagnosis. This outcome presents the proper distinguishment of keywords that are important to specific categorical documents within the collection in a group of documents.

Figure 6. Data visualization networks (Common bigrams that occurred in categorical disease documents).

```
  A tibble: 19,740 x 3
item1      item2      correlation
<chr>      <chr>            <dbl>
1 urinary    intestinal       0.923
2 intestinal urinary          0.923
3 functions  urinary          0.870
4 urinary    functions        0.870
5 pressure   blood            0.865
6 blood      pressure         0.865
7 dear       master           0.851
8 master     dear             0.851
9 functions  intestinal       0.830
10 intestinal functions       0.830
# ... with 19,730 more rows
```

Figure 7. Correlation table between the symptomatic words.

3.4. Text Classification with ML Modelling

From the large sample of medical documents, we selected two categorical documents, such as cardiovascular and digestive diseases. The document sample is 8803 and it is further divided into 70:30, where 70% of documents are for training and 30% are for test purposes. Three models were trained with an optimal parameter which was defined by CV validation. Each experiment was conducted with a 10-k value. The Receiver Operating Characteristics (ROC) curves can be used in medical diagnosis to test the model's ability to predict text in documents [37]. In particular, ROC curves are known for their ability to visualize binary classification. In Table 2, we compare the performance metrics in terms of accuracy, sensitivity, specificity, and ROC.

Table 2. Performance comparison of adopted models (k = 10).

Model	Accuracy (%)	Sensitivity (%)	Specificity (%)	ROC
SVM	64.2	68.3	45.3	0.597
RF	59.0	59.8	55.4	0.613
LASSO	93.8	97.9	80.6	0.976

The accuracy can be measured as the ratio between some of the true predicted documents and the total number of documents. Among the 2641 documents tested, the LASSO model correctly predicted 2477 documents, while 164 were incorrectly predicted, resulting in a 93.8% accuracy. The Figure 8 presents the ROC curve outcome for LASSO regression model where false positive rate (1-specificity) on *x*-axis and true positive rate (sensitivity) on *y*-axis. It is obvious that the ROC value of 0.976 shows a perfect classification of the categorical documents included. These results indicate that LASSO regression outperforms the other two classification models. Moreover, sensitivity is the percentage of correctly predicted outcomes, and this model scored 97.9%, a relatively high value for prediction problems.

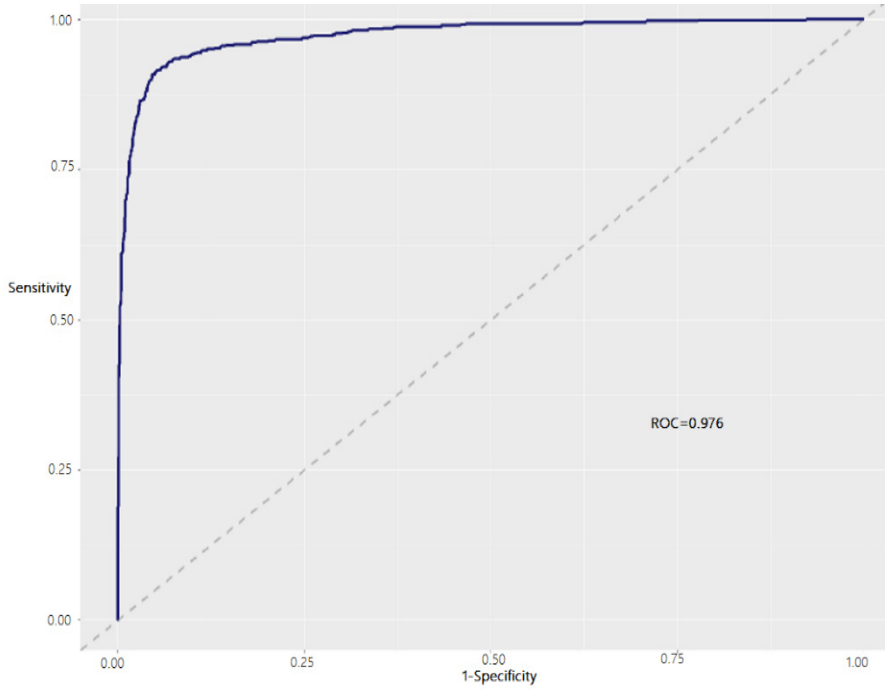


Figure 8. ROC curve for text classification using LASSO regularized regression.

4. Discussion

This paper presents an integrated analysis of medical documents of seafarers that incorporates TM and sentiment analysis. Due to sailors' distance from medical facilities, health care is difficult. When doctors onshore diagnose correctly, they can prescribe effective treatments [38–40]. Using electronic health records (EHRs) can help reorganize services, train staff on the benefits of working with different teams, and protect staff health, according to this study. A lack of assistance will put staff and seafarers at risk.

Seafarers often suffer from swelling, fatigue, general weakness, headaches, confusion, and other symptoms. Different studies have found a connection between symptoms, including hemoptysis, thrombocytopenia, and high C-reactive protein levels [41,42]. A hypoxemic patient can develop acute respiratory distress syndrome (ARDS), sepsis, and even multi-organ failure in a short amount of time [43]. A seafarer's disease type must be diagnosed quickly and treated according to his symptomatology.

Medical care professionals can utilize TM to collect and write information from the patient's EHRs to support decision-making and disease diagnosis [44]. Pattern-based mining from EHRs assists doctors, physicians, and laboratory experts by retrieving significant knowledge from the system [45]. Pletscher-Frankild and coworkers developed a TM software program that identifies human and disease genes in text and identifies disease-gene associations [46]. A study analyzed 100 posts from epilepsy patient forums to quantitatively analyze patient perspectives on treatments, which may aid medical experts in designing clinical decision making based on patient-derived information [47].

In recent reports, the importance of EHRs to the functioning of healthcare systems has been emphasized [48]. Seafarers, for example, have been provided with telematics services for years. Medical prescriptions are issued in a single, standardized format. Physicians and pharmacies are required to submit prescription data electronically. Online medical records are preferred by doctors when prescribing treatments. Text information is stored in these records. These types of data can be processed by healthcare experts, but they can only estimate which diseases they are dealing with.

TM has been used to predict hospital admissions based on emergency department initial medical records [49]. TM can provide valuable information and make it easier for bed management teams to make decisions. It is reported that TM also plays an influential role in the autonomous classification of hospital admissions. Moreover, it is reported that TM examines the performance of text classification from clinical data from hospital reports [50]. This describes how TM can identify hospitalized patients for the treatment of a given disease based on the information associated with patient admission. This study used a machine learning approach called LASSO regression to distinguish disease documents based on clinical terms. The application of statistical and epidemiological methods in medical research was done with TM. The authors in [51], review the last twenty-year reports to identify commonly encountered and emergent methods used to investigate medical research problems.

In medical research, TM was applied to statistical and epidemiological analyses. Meaney and colleagues reviewed reports from the last twenty years to identify commonly encountered and emerging research methods [51]. The TM has been used to predict hospital admissions using emergency department records [49]. Providing this information will help bed management teams make better decisions. TM is also utilized to classify hospital admissions as well as hospital reports [50]. The information is derived from the admission information for the patient. Through TM, patients can be identified for specific diseases.

In an emergency situation, TM knowledge is extremely valuable because it allows high-quality data to be generated in real time. It is imperative for patients and medical staff to be careful when providing information [52]. Therefore, a gap can develop between the data scientists, scholars, and medical professionals capable of producing the data. The importance of sharing data across care providers will also be affected by this gap. A tidy TM needs to be presented in these situations so that raw data can be visualized in a prescribed format and distributed evenly among specialists. Medical abstracts have been organized

neatly in this paper, and symptom maps for illnesses experienced onboard have been visualized. LASSO regression models are also used to validate the results. Remote doctors provide maritime telemedicine assistance, but these practices illustrate the limitations of using digital health records to produce quality data.

5. Conclusions

For the management of unstructured datasets, tidy-based TM has proven to be a comprehensive and efficient tool. It is relatively difficult to recognize treatments and relevant facts in medical documents written in languages other than English. By combining tidy TM packages and libraries with semantic manipulation, we developed a comprehensive approach to identifying onboard diseases. An ICD-10 symptom mapping was also undertaken. Symptom correlation plots, which measure how different health problems are linked together, were also presented. Using LASSO regression, this study successfully predicted text data among documents with 93.8% accuracy. These tidy TM libraries can effectively classify text documents in healthcare analysis projects. As well as delivering medical assistance, this approach may be used to develop health observatories and to classify diseases. We propose to apply the knowledge developed in this work to the Epidemiological Observatory of Seafarers Pathologies and Injuries, a collaborative initiative between the Ministry of Health, University of Camerino, and the International Radio Medical Center (C.I.R.M.).

Author Contributions: Planning of the study, N.C., U.A., G.B., M.d.C., C.M., G.R., G.G.S., A.S. and F.A.; Conceptualization, N.C., G.B., C.M., A.S. and F.A.; Methodology, N.C., G.B., M.d.C. and G.G.S.; Formal analysis, N.C. and G.B.; Investigation and experiments, N.C., G.B. and G.G.S.; Resources, N.C., U.A., G.B., M.d.C., C.M., G.R., G.G.S., A.S. and F.A.; Data curation, N.C.; Writing—original draft preparation, G.B.; Text review and editing, N.C., U.A., G.B., M.d.C., C.M., G.R., G.G.S., A.S. and F.A.; Supervision, U.A., G.R. and F.A.; Project administration, U.A., G.R. and F.A.; Funding acquisition, F.A. All authors have read and agreed to the published version of the manuscript.

Funding: The Italian Ministry of Health supported this study by grant No. J59J21011210001 in part of developing the Epidemiological Observatory of Seafarers Pathologies and Injuries.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data examined for the present study were collected and stored in a closed database by the Centro Internazionale Radio Medico (C.I.R.M.), the Italian Maritime Telemedical Assistance Service (TMAS) in the frame of health surveillance activities performed onboard ships. Data were extracted from the database by C.I.R.M. operators and anonymized before being used for research purposes. C.I.R.M. President as legal representative of the entity where medical data are kept has authorized access to authors for collecting data of this work. The programing code for experiments that are involved can be found in https://github.com/nalinichintalapudi/Tidymodels-for-medical-text-data.git.

Acknowledgments: We greatly acknowledge the support given by ITF Trust by grant No. 1276/2018 for epidemiological analysis and data mining operations.

Conflicts of Interest: No author has any conflict during the preparation and publication of the manuscript.

References

1. Abila, S.S.; Acejo, I.L. Mental health of Filipino seafarers and its implications for seafarers' education. *Int. Marit. Health* **2021**, *72*, 183–192. [CrossRef] [PubMed]
2. Guillot-Wright, S. The changing economic structure of the maritime industry and its adverse effects on seafarers' health care rights. *Int. Marit. Health* **2017**, *68*, 77–82. [CrossRef] [PubMed]
3. Caruso, G. Do seafarers have sunshine. In Proceedings of the 8th International Symposium on Maritime Health (ISMH) Book of Abstracts, Rijeka, Croatia, 8–13 May 2005.
4. Laraqui, O.; Manar, N.; Laraqui, S.; Ghailan, T.; Deschamps, F.; Hammouda, R.; Laraqui, C.E.H. Prevalence of skin diseases amongst Moroccan fishermen. *Int. Marit. Health* **2018**, *69*, 22–27. [CrossRef] [PubMed]
5. Mahdi, S.S.; Amenta, F. Eighty years of CIRM. A journey of commitment and dedication in providing maritime medical assistance. *Int. Marit. Health* **2016**, *67*, 187–195. [CrossRef] [PubMed]

6. Sagaro, G.; Battineni, G.; Di Canio, M.; Amenta, F. Self-Reported Modifiable Risk Factors of Cardiovascular Disease among Seafarers: A Cross-Sectional Study of Prevalence and Clustering. *J. Pers. Med.* **2021**, *11*, 512. [CrossRef]
7. Antons, D.; Grünwald, E.; Cichy, P.; Salge, T.O. The application of text mining methods in innovation research: Current state, evolution patterns, and development priorities. *R&D Manag.* **2020**, *50*, 329–351. [CrossRef]
8. Battineni, G.; Sagaro, G.G.; Chintalapudi, N.; Amenta, F. Conceptual Framework and Designing for a Seafarers' Health Observatory (SHO) Based on the Centro Internazionale Radio Medico (C.I.R.M.) Data Repository. *Sci. World J.* **2020**, *2020*, 8816517. [CrossRef] [PubMed]
9. Chintalapudi, N.; Battineni, G.; Di Canio, M.; Sagaro, G.G.; Amenta, F. Text mining with sentiment analysis on seafarers' medical documents. *Int. J. Inf. Manag. Data Insights* **2020**, *1*, 100005. [CrossRef]
10. Ribeiro, J.; Duarte, J.; Portela, F.; Santos, M. Automatically detect diagnostic patterns based on clinical notes through Text Mining. *Procedia Comput. Sci.* **2019**, *160*, 684–689. [CrossRef]
11. Grover, P.; Kar, A.K. Big Data Analytics: A Review on Theoretical Contributions and Tools Used in Literature. *Glob. J. Flex. Syst. Manag.* **2017**, *18*, 203–229. [CrossRef]
12. Wu, C.-S.; Kuo, C.-J.; Su, C.-H.; Wang, S.; Dai, H.-J. Using text mining to extract depressive symptoms and to validate the diagnosis of major depressive disorder from electronic health records. *J. Affect. Disord.* **2019**, *260*, 617–623. [CrossRef]
13. Karami, A.; Ghasemi, M.; Sen, S.; Moraes, M.F.; Shah, V. Exploring diseases and syndromes in neurology case reports from 1955 to 2017 with text mining. *Comput. Biol. Med.* **2019**, *109*, 322–332. [CrossRef] [PubMed]
14. Guerreiro, J.; Rita, P. How to predict explicit recommendations in online reviews using text mining and sentiment analysis. *J. Hosp. Tour. Manag.* **2019**, *43*, 269–272. [CrossRef]
15. Denecke, K.; Deng, Y. Sentiment analysis in medical settings: New opportunities and challenges. *Artif. Intell. Med.* **2015**, *64*, 17–27. [CrossRef]
16. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [CrossRef] [PubMed]
17. Vij, A.; Pruthi, J. An automated Psychometric Analyzer based on Sentiment Analysis and Emotion Recognition for healthcare. *Procedia Comput. Sci.* **2018**, *132*, 1184–1191. [CrossRef]
18. Moreira, L.B.; Namen, A.A. A hybrid data mining model for diagnosis of patients with clinical suspicion of dementia. *Comput. Methods Programs Biomed.* **2018**, *165*, 139–149. [CrossRef]
19. Marir, F.; Said, H.; Al-Obeidat, F. Mining the Web and Literature to Discover New Knowledge about Diabetes. *Procedia Comput. Sci.* **2016**, *83*, 1256–1261. [CrossRef]
20. Abirami, A.M.; Gayathri, V. A survey on sentiment analysis methods and approach. In Proceedings of the 2016 Eighth International Conference on Advanced Computing (ICoAC), Chennai, India, 19–21 January 2017; pp. 72–76. [CrossRef]
21. Barlas, P.; Adam, S.; Chatelain, C.; Paquet, T. A Typed and Handwritten Text Block Segmentation System for Heterogeneous and Complex Documents. In Proceedings of the 2014 11th IAPR International Workshop on Document Analysis Systems, Tours, France, 7–10 April 2014; pp. 46–50. [CrossRef]
22. Zeng, D.; Peng, J.; Fong, S.; Qiu, Y.; Wong, R. Medical data mining in sentiment analysis based on optimized swarm search feature selection. *Australas. Phys. Eng. Sci. Med.* **2018**, *41*, 1087–1100. [CrossRef]
23. CRAN—Package Tidytext. Available online: https://cran.r-project.org/web/packages/tidytext/index.html (accessed on 23 February 2022).
24. Wickham, H. Tidy Data. *J. Stat. Softw.* **2014**, *59*, 1–23. [CrossRef]
25. Data Visualization with R and ggplot2 | the R Graph Gallery. Available online: https://www.r-graph-gallery.com/ggplot2-package.html (accessed on 1 March 2022).
26. Rathore, A.K.; Kar, A.K.; Ilavarasan, P.V. Social Media Analytics: Literature Review and Directions for Future Research. *Decis. Anal.* **2017**, *14*, 229–249. [CrossRef]
27. Emmert-Streib, F.; Dehmer, M. High-Dimensional LASSO-Based Computational Regression Models: Regularization, Shrinkage, and Selection. *Mach. Learn. Knowl. Extr.* **2019**, *1*, 359–383. [CrossRef]
28. Text Mining and Word Cloud Fundamentals in R: 5 Simple Steps You Should Know—Easy Guides—Wiki—STHDA. Available online: http://www.sthda.com/english/wiki/text-mining-and-word-cloud-fundamentals-in-r-5-simple-steps-you-should-know (accessed on 1 March 2022).
29. Dimitri, G.M.; Beqiri, E.; Placek, M.M.; Czosnyka, M.; Stocchetti, N.; Ercole, A.; Smielewski, P.; Lió, P.; Anke, A.; Beer, R.; et al. Modeling Brain–Heart Crosstalk Information in Patients with Traumatic Brain Injury. *Neurocrit. Care* **2021**, 1–13. [CrossRef] [PubMed]
30. Ternès, N.; Rotolo, F.; Michiels, S. Empirical extensions of the lasso penalty to reduce the false discovery rate in high-dimensional Cox regression models. *Stat. Med.* **2016**, *35*, 2561–2573. [CrossRef]
31. Kan, H.J.; Kharrazi, H.; Chang, H.-Y.; Bodycombe, D.; Lemke, K.; Weiner, J.P. Exploring the use of machine learning for risk adjustment: A comparison of standard and penalized linear regression models in predicting health care costs in older adults. *PLoS ONE* **2019**, *14*, e0213258. [CrossRef] [PubMed]
32. Khanji, C.; Lalonde, L.; Bareil, C.; Lussier, M.-T.; Perreault, S.; Schnitzer, M.E. Lasso Regression for the Prediction of Intermediate Outcomes Related to Cardiovascular Disease Prevention Using the TRANSIT Quality Indicators. *Med. Care* **2019**, *57*, 63–72. [CrossRef] [PubMed]

33. Shan, C. Research of Support Vector Machine in Text Classification. In *Future Computer, Communication, Control and Automation*; Zhang, T., Ed.; Advances in Intelligent and Soft Computing; Springer: Berlin/Heidelberg, Germany, 2012; Volume 119, pp. 567–573. [CrossRef]
34. Hassani, H.; Beneki, C.; Unger, S.; Mazinani, M.T.; Yeganegi, M.R. Text Mining in Big Data Analytics. *Big Data Cogn. Comput.* **2020**, *4*, 1. [CrossRef]
35. Tabe-Bordbar, S.; Emad, A.; Zhao, S.D.; Sinha, S. A closer look at cross-validation for assessing the accuracy of gene regulatory networks and models. *Sci. Rep.* **2018**, *8*, 6620. [CrossRef] [PubMed]
36. Iversen, R.T.B. The mental health of seafarers. *Int. Marit. Health* **2012**, *63*, 78–89. [PubMed]
37. Pencina, M.J.; D'Agostino, R.B.; D'Agostino, R.B.; Vasan, R.S. Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **2008**, *27*, 157–172. [CrossRef]
38. Woldaregay, A.Z.; Walderhaug, S.; Hartvigsen, G.; Guitton, M.; Seale, D. Telemedicine Services for the Arctic: A Systematic Review. *JMIR Med. Inform.* **2017**, *5*, e16. [CrossRef] [PubMed]
39. Mair, F.; Fraser, S.; Ferguson, J.; Webster, K. Telemedicine via satellite to support offshore oil platforms. *J. Telemed. Telecare* **2008**, *14*, 129–131. [CrossRef] [PubMed]
40. Dehours, E.; Vallé, B.; Bounes, V.; Girardi, C.; Tabarly, J.; Concina, F.; Pujos, M.; Ducassé, J.L. User satisfaction with maritime telemedicine. *J. Telemed. Telecare* **2012**, *18*, 189–192. [CrossRef] [PubMed]
41. Wang, D.; Hu, B.; Hu, C.; Zhu, F.; Liu, X.; Zhang, J.; Wang, B.; Xiang, H.; Cheng, Z.; Xiong, Y.; et al. Clinical Characteristics of 138 Hospitalized Patients With 2019 Novel Coronavirus—Infected Pneumonia in Wuhan, China. *JAMA* **2020**, *323*, 1061–1069. [CrossRef] [PubMed]
42. The Novel Coronavirus Pneumonia Emergency Response Epidemiology Team. The Epidemiological Characteristics of an Outbreak of 2019 Novel Coronavirus Diseases (COVID-19)—China, 2020. *China CDC Wkly.* **2020**, *2*, 113–122. [CrossRef]
43. Paraskevis, D.; Kostaki, E.; Magiorkinis, G.; Panayiotakopoulos, G.; Sourvinos, G.; Tsiodras, S. Full-genome evolutionary analysis of the novel corona virus (2019-nCoV) rejects the hypothesis of emergence as a result of a recent recombination event. *Infect. Genet. Evol.* **2020**, *79*, 104212. [CrossRef] [PubMed]
44. Tsumoto, S.; Kimura, T.; Iwata, H.; Hirano, S. Mining Text for Disease Diagnosis. *Procedia Comput. Sci.* **2017**, *122*, 1133–1140. [CrossRef]
45. Metsker, O.; Bolgova, E.; Yakovlev, A.; Funkner, A.; Kovalchuk, S. Pattern-based Mining in Electronic Health Records for Complex Clinical Process Analysis. *Procedia Comput. Sci.* **2017**, *119*, 197–206. [CrossRef]
46. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: Text mining and data integration of disease–gene associations. *Methods* **2015**, *74*, 83–89. [CrossRef]
47. He, K.; Hong, N.; Lapalme-Remis, S.; Lan, Y.; Huang, M.; Li, C.; Yao, L. Understanding the patient perspective of epilepsy treatment through text mining of online patient support groups. *Epilepsy Behav.* **2019**, *94*, 65–71. [CrossRef]
48. Groenhof, T.K.J.; Koers, L.R.; Blasse, E.; de Groot, M.; Grobbee, D.E.; Bots, M.L.; Asselbergs, F.W.; Lely, A.T.; Haitjema, S.; van Solinge, W.; et al. Data mining information from electronic health records produced high yield and accuracy for current smoking status. *J. Clin. Epidemiol.* **2020**, *118*, 100–106. [CrossRef] [PubMed]
49. Lucini, F.R.; Fogliatto, F.S.; da Silveira, G.J.; Neyeloff, J.; Anzanello, M.J.; Kuchenbecker, R.S.; Schaan, B.D. Text mining approach to predict hospital admissions using early medical records from the emergency department. *Int. J. Med. Inform.* **2017**, *100*, 1–8. [CrossRef] [PubMed]
50. Kocbek, S.; Cavedon, L.; Martinez, D.; Bain, C.; Mac Manus, C.; Haffari, G.; Zukerman, I.; Verspoor, K. Text mining electronic hospital records to automatically classify admissions against disease: Measuring the impact of linking data sources. *J. Biomed. Inform.* **2016**, *64*, 158–167. [CrossRef] [PubMed]
51. Meaney, C.; Moineddin, R.; Voruganti, T.; O'Brien, M.A.; Krueger, P.; Sullivan, F. Text mining describes the use of statistical and epidemiological methods in published medical research. *J. Clin. Epidemiol.* **2016**, *74*, 124–132. [CrossRef] [PubMed]
52. Grantz, K.H.; Meredith, H.R.; Cummings, D.A.T.; Metcalf, C.J.E.; Grenfell, B.T.; Giles, J.R.; Mehta, S.; Solomon, S.; Labrique, A.; Kishore, N.; et al. The use of mobile phone data to inform analysis of COVID-19 pandemic epidemiology. *Nat. Commun.* **2020**, *11*, 4961. [CrossRef] [PubMed]