information

MDPI

Article

Machine Learning Based Sentiment Text Classification for Evaluating Treatment Quality of Discharge Summary

Samer Abdulateef Waheeb, Naseer Ahmed Khan, Bolin Chen and Xuequn Shang *

School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China; samirabdulateef@mail.nwpu.edu.cn (S.A.W.); naseerkhan@mail.nwpu.edu.cn (N.A.K.); blchen@nwpu.edu.cn (B.C.)

* Correspondence: shang@nwpu.edu.cn

Received: 8 March 2020; Accepted: 20 May 2020; Published: 23 May 2020

Abstract: Patients' discharge summaries (documents) are health sensors that are used for measuring the quality of treatment in medical centers. However, extracting information automatically from discharge summaries with unstructured natural language is considered challenging. These kinds of documents include various aspects of patient information that could be used to test the treatment quality for improving medical-related decisions. One of the significant techniques in literature for discharge summaries classification is feature extraction techniques from the domain of natural language processing on text data. We propose a novel sentiment analysis method for discharge summaries classification that relies on vector space models, statistical methods, association rule, and extreme learning machine autoencoder (ELM-AE). Our novel hybrid model is based on statistical methods that build the lexicon in a domain related to health and medical records. Meanwhile, our method examines treatment quality based on an idea inspired by sentiment analysis. Experiments prove that our proposed method obtains a higher F1 value of 0.89 with good TPR (True Positive Rate) and FPR (False Positive Rate) values compared with various well-known state-of-the-art methods with different size of training and testing datasets. The results also prove that our method provides a flexible and effective technique to examine treatment quality based on positive, negative, and neutral terms for sentence-level in each discharge summary.

Keywords: discharge summaries; text clustering; extreme learning machine; sentiment analysis; health surveillance; quality evaluation

1. Introduction

Discharge summaries (DS) are a part of clinical notes which have an unstructured design. These kinds of documents include information about drugs, treatments, and diseases. These summaries are written often under the work pressures that lead to incomplete sentences, misspellings, jargon, and non-regular abbreviations [1,2]. The automation of the medical information system (patient records) has enabled massive unstructured textual information storage. Extraction of useful knowledge from a huge number of unstructured discharge summaries has become a significant challenge in recent times [3]. This extracted knowledge is significant in various stages of a patient's life, for example, it is an important factor for the patient's health when physicians want to get the information about a patient's health progress (during the treatment stages). Extracting knowledge from these summaries allows the evaluation of the treatment quality in a better way that can benefit both patients and health centers [4,5].

One of the ways to examine the treatment quality is to transform unstructured data into structured data which can be done by steps such as, prepossessing (text cleaning), dimensionality reduction,

and then selecting the most relevant and meaningful text (text presentation) which is used to build a structured discharge summary. The quality of evaluation is measured to examine the accuracy of these structured documents and the main objective of a structured discharge summary is to examine the accuracy of the diagnostic measures and the quality of treatments given to patients. Incorrect diagnostics and treatment could result in incorrect and poor care quality, readmission risk, and incorrect management decisions that could severely affect patient's health in particular, and a health center's reputation in general. The evaluation of discharge summary has the capability to determine the limitations of the procedures available at the health centers and to provide better diagnostic and treatment to the patients [6].

Based on previous works, there are studies conducted to examine and evaluate the quality of discharge summary, such as evaluating the discharge summary based on the reporting of adverse events in thoracic treatment [7], patient readmission risk with heart failure [8], examining, the timeliness and accuracy of the treatments [9], the accuracy of the respiratory diagnosis [6], clinical discharge letters, risks related to patient safety [10], and evaluating the quality of diagnostic for pediatrics diseases [11]. Although, there are some studies on discharge summary, for example, extracting information from an unstructured format, they have mainly focused on extracting the medical entities like drugs, diseases, and treatment, and relationships like phenotype identification, disease classification, and disease diagnosis [12,13]. Previous studies suffer from various limitations, for example, most of the works focused on extracting the medical terms and the relationships between these terms in the medical domain and on methods which help determine quality of discharge summary. However, these limitations could be overcome by proposing a new approach for extracting useful and relevant knowledge and information hidden within this enormous amount of clinical data [14].

Sentiment analysis (SA) is one of the new techniques in natural language processing (NLP) [15]. Sentiment classification for health care deals with the diagnosis of healthcare-related problems which are diagnosed and treated by health centers' expert staff, like doctors, nurses, and laboratory technicians. It takes their opinions into perspective to make policies and modifications that could directly address patient's problems related to health and follow-up. Sentiment classification is used generally with HMIS (hospital management information system) and other software products like mobile applications designed for health centers expert, and it is having a significant, positive effect in the health care industry to ensure the quality of services [16]. Aspect-based analysis of health care not only recommends the quality services and focused treatments, but also helps in developing a technological system in the health care centers that is proved to be more effective. Machine learning techniques are used to analyses millions of reviewed documents so that efficient, accurate, and fast treatment and decisions get implemented in the health care centers. Using both health care expert decisions and analytical based models on the text data available at the health care centers, a quality-based treatment can be deployed in the health care centers [17].

Text classification based on sentiment analysis helps us classify DS to positive (improved health status), negative (degraded health status), and neutral (stable health status) documents, which helps evaluate and improve treatment quality. Previous studies on sentiment analysis focus on extracting information from the subjective corpus like Twitter and company reviewer by using machine learning methods [18–22]. The main idea of our DS-SA-extreme learning machine autoencoder (ELM-AE) is to overcome the problems [18,23,24], which are as follows: (1) this study will focus on the objective corpus and sentiment analysis problems that deal with DS, such as data with a large feature set, data sparsity, polarity shift, multi-classification, the lexicon of a specific domain, and accuracy; (2) one of the challenging tasks is the polarity of context in sentiment analysis, it is related to sentiment analysis of the context where the previous word polarity changes with respect to contexts, (3) the problem with word embedding is the word senses, which means this model cannot recognize the synonyms for a word and generates a single word form of representation [25]; (4) the other most common problem is the lack of lexicon specific domain in discharge summaries; (5) we used the vector space model to convert each word or sentence to vector and then got a representation of these vectors in a low-dimensional

space. Nevertheless, the word embedding model has its own significant problems, like ignoring the words sentiment polarity. However, this model has drawbacks with sentiment information and cannot capture the whole sentiment information of sentences [24].

We have devised a novel model, which is described as follows:

- A deep learning model with a combination of features consisting of word-embeddings, sentiment-shifter rules, sentiment based extracted knowledge, linguistic and statistical based extracted knowledge that has not been used before based on our literature review on the relevant studies.
- Meanwhile, we experimented with multiple strategies like statistical models, weighted principal component analysis (W-PCA), Chi-squared statistics (CSS), and weighted-support vector machine SVM that help us generate the specific lexicon.
- We have adopted an ELM as a multi-classification problem (positive, negative, and natural), by using multi-level of features with discharge summaries.

2. Related Work

2.1. Quality Evaluation for Discharge Summary

Medical institutions work as a provider for discharge summaries, which have contents on various aspects of patient's details, like nurse letters, radiological reports, and drug reviews [26]. These documents are used as a source for extracting the valuable information, which can help evaluate quality of the treatment and enhance the medical care decisions. In general, there are two parts of the work to examine the discharge summary quality. The first one is focused on how to enhance the quality of discharge summary itself and the second one is to examine quality of the treatment based on this summary. Moreover, the second kind of work can be separated into: (1) examining the quality of summary based on the medical factors and (2) testing the quality for discharge summary through information technology (computer-aided). The methods suggested by [27,28] examined the quality of discharge summary that relies on some medical factors. They listed some risk factors like delayed delivered summary, lack of information (low quality), that writing discharge summary is missing in medical learning, and the absence of patient understanding. The summary should be submitted to physicians to ensure that they wrote the last treatment after patient discharge from hospital to avoid losing important information. Insufficient and incomplete discharge summaries led to increased readmission risk and many other complications. The junior doctors informed that they did not receive sufficient training and guidance for how to write a discharge summary so they suggested translating medical terms and formulation for the readability and understanding of the patients.

The research proposed by [6] also examines the quality of diagnosis written in the discharge summary. This research focused on making the comparison between diagnosis lists written in the summary and the reference list of diagnosis for patients. This comparison was implemented based on five factors, namely: wrong diagnosis, missing diagnosis, correct diagnosis, serious imprecise diagnosis, and partial imprecise diagnosis. The recommendation of this research is to improve the discharge summary quality by involving the junior physicians in activities related to writing the discharge summary to avoid common mistakes, deliver the summary on time, avoid repeated diagnosis, and share records among the specialist doctors electronically.

The research proposed by [11] tested the quality of diagnoses for pediatric diseases using an artificial intelligence approach and examined the quality of three different kinds of diagnosis. The first diagnosis comes from the machine learning suggested model for extracting the most relevant features from the summary to mimic the medical human resources (physicians), the second one comes from junior physicians, and the third comes from senior physicians. The model has components from text prepossessing framework, schema constructional, word embedding, lexicon constructional, and sentences classification based on long short-term memory. The final result of examining the quality

of diagnosis (model diagnoses) improved the F1 measure to 0.885, as compared to junior doctors' diagnosis that obtained a lower F1 measure than senior physicians diagnosis.

2.2. Sentiment Analysis

One of the dynamic research areas related to NLP is sentiment analysis, this tries to identify objective information and define the orientation of sentiment for a given document, for example to neutral, positive, and negative classification of the sentences [29]. The related work divided the tradition sentiment analysis approaches into three groups: machine learning, lexicon, and hybrid methods. One of the classifying orientations of sentiment is the machine learning-based method based on common algorithms of machine learning. Additionally, the lexicon method works with machine learning methods in order to extract the related features of sentiment and enhance accuracy, this approach is called the hybrid approach [22].

Mohammed et al. [25] proposed a sentiment classification system based on the supervised machine learning approach, the logistic regression algorithm. They used the groups of features that were used by various methods, such as character n-grams, lexicons, word senses, etc. The basic linguistic functions like a spelling corrector and a word sense disambiguate were also applied in the text pre-processing. Recently, many works have been based on sentiment analysis and feature selection use classifiers, such as naive Bayes with an accuracy of 78.02% [30], support vector machine (SVM) with an F1 measure of 0.787 [31], decision tree with an accuracy of 87.7% [32], and LSTM(Long Short Term Memory) with an F1 measure of 0.778 [33].

Spinczyk and Dominik [34] developed an approach based on the anorexia nervosa with sentiment analysis to examine and evaluate the treatment and diagnosis. They used 67 cases for their study and divided them into the ill cases, containing 15 documents, and healthy cases, with 52 documents, as a control group. Their work relies on the bag-of-words method and they used the lexicon method to get the polarity of each document, based on the rule-based model. One of the main limitations of their study is quantitative results and that they just used the technique of text categorization to classify the documents. Their paper showed a p-value = 0.0025 but ignored the comparison with the state-of-the-art techniques.

Jiang and Keyuan [35] developed an approach based on the word embedding and deep learning model to examine the medical date that rely on the social media API. Their paper tested the health issues based on the patient's opinion on the 103 medicines. They selected 12,331 random rows from 22 million rows for the classifier model. They used 10-fold cross-validations with logistic regression, decision tree, KNN (K-Nearest Neighbors), SVM, and bag-of-words (BoW), for testing their approach and got a performance of 0.645 using the F1 measure.

2.3. Deep Learning

Deep learning has of late, been more popular than ever before due to the stepwise approach of adding more computational neural network layers with non-linear functions to represent almost any complicated computational problem, the idea is to use a technique called "back-propagation" that works with the training data and the parameters or weights of the computational layered model are tuned gradually with each training iteration. In the beginning, deep learning was limited to image and video processing, but with the advent of 1D-CNN(1-Dimensional Convolution Neural Networks) and LSTM it is now being extensively used in the domain of text and natural language processing. Deep learning, which is also sometimes called representation learning, is a set of methods by which raw data or input is fed to some non-linear functions organized in a layer wise fashion, and a representation of the data is thereby learned. Similarly, more layers are added in this fashion and a more abstract representation is learned based on the raw data, which could finally be used to detect patterns in the data to classify objects in the data, called supervised learning, or learning useful patterns, called unsupervised learning from the raw data [36].

The extreme learning machine autoencoder (ELM-AE) mean objective is to represent the meaningful of the input features with three various representations: compressed representation, sparse representation, and equal dimension representation, hence, ELM-AE is competent of learning a suitable feature representation. However, selecting random biases and random weights has obtained good performance [37].

In general, based on the above discussion, most papers are focused on examining the treatment quality based on the level of discharge summary quality (medical factures) or focused on extracting the medical terms and the relationships between these terms in the medical domain [12]. Our work will go deeper to extract other relationships based on the sentiment terms. This kind of work helps to evaluate the treatment quality in the clinical domain. Moreover, most of the sentiment analysis works are implemented with customers review and patient's opinion as subjective corpus based on general lexicon, and this lexicon ignores the neutral terms. Based on our opinion, we need more investigation to examine the quality of treatment with discharge summary as an objective corpus and involve sentiment analysis techniques to this end.

However, this research paper suggests a semi-supervised method that applies sentences vector (SV)-BOW separately for document embedding. Based on these two models, we generated the feature sets as a list of the words and a list of sentences. The extreme learning machine is applied for reducing the feature set and the classification task. A comparison with the state-of-the-art classification approaches: SVM, random forest, naïve Bayes, logistic regression, embeddings language models (ELMo), CNN (LSTM), and RNN (LSTM) was also performed.

3. The Challenge of the Text Classification Task

The discharge summary classification problem is formulated as given a set of documents, MD, that is, $\mathbf{MD} = (\mathbf{D}1,\ \mathbf{D}2, \ldots\ldots,\ \mathbf{DN})$ where Di indicates the *i*th document in MD, *N* represents the total number of documents in the dataset. Then, we tokenize a document, D_i, to a list of sentences, that is, D = (S1, . . . , Sn), where Si represents the *i*th sentence in D and *n* represents the total number of sentences in each document. The goal of the final classification is to predict the polarity of documents based on sentence polarity from MD using various related discharge summary from similar or related diseases.

Using the technique of sentiment analysis, we are evaluating the quality of the discharge summary by classifying discharge summary documents obtained from the health/medical domain. Based on the previous results, our study is novel as it deals with the multi-classification problem (positive, negative, and natural), by using a multi-level of features within discharge summaries. The discharge summaries with sentiment analysis are based on extreme learning machine-autoencoder (ELM-AE) multi-classification task (DSSA-ELM-AE) for sentiment analysis classifier at the level of sentences. Multi-level of features (word and sentence embedding, linguistic knowledge, association rules, sentiment feature, and medical concept) consider an input vector to our model for encoding the features to represent the sentence vector [22]. The performance of these algorithms on clinical text is still not explored yet, as only a few studies have been conducted in this field. The combined multi-level of feature set which includes the word and sentence embedding, rules of sentiment shifter, sentiment knowledge, linguistic, medical concept, and statistical knowledge has not been completely investigated for sentiment analysis in the clinical text-domain. We integrate a multi-level of features to deal with the following problems, namely: polarity of the contextual, word sense differences, the limit of the word coverage in the general lexicon when dealing with sentiment analysis problems in the clinical text, and words with the context of similar semantic but reverse sentiment polarity.

4. Proposed System

For this study, a novel sentiment analysis method is proposed for evaluating the treatment quality. There are four main stages for this approach, namely: data collection, text cleaning, lexicon generation, and extreme learning machine autoencoder experiments. The final step is to evaluate our suggested approach using metrics such as recall, precession, and F1, and also to compare the final

results with state-of-the-art methods from the literature. Figure 1 demonstrates the main steps for the suggested approach.

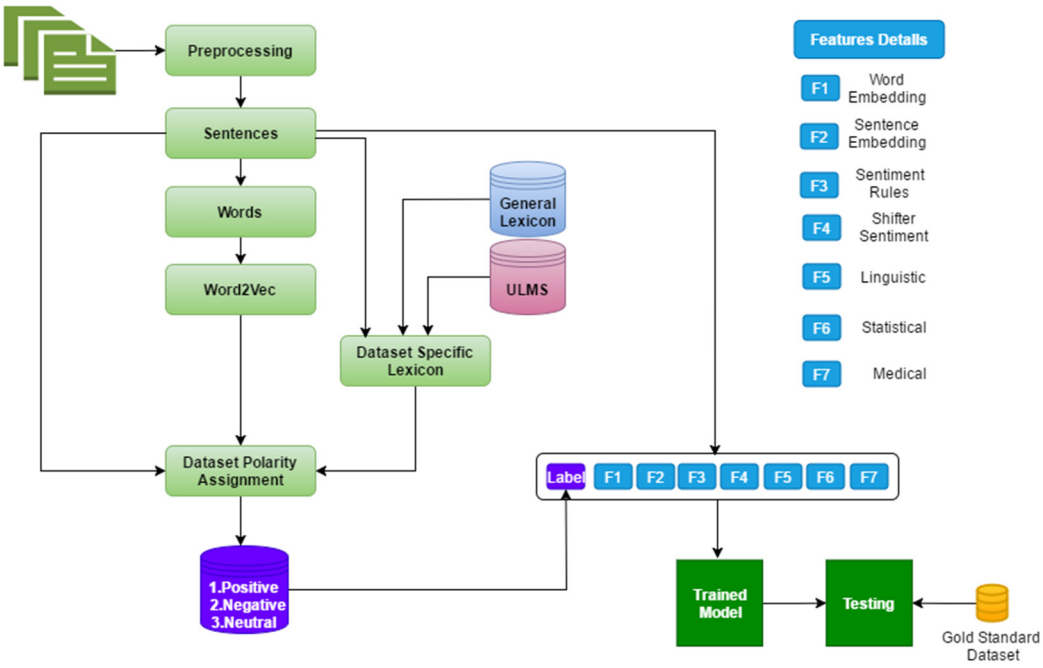


Figure 1. Steps of the proposed method.

The first stage is data collection as data is gathered from the medical records of the health and medical centers in the form of discharge text summaries. After that, the second stage is to simplify the text, which contains six main steps, namely: Tokenizing process, this step with sub-steps, which are tokening the document into paragraphs, then into sentences and words and finally a Word2Vec model is used to so that each word is represented into a fixed length numeric vector. The tokenizing process uses symbols like spaces and punctuation for separating the discharge summaries text into words [38,39]. The third phase is to assign polarity to the sentences which are done by first, generating the lexicon, called general lexicon, for the discharge summaries in addition to the available ULMS (Unified Medical Language System) lexicon which can help solve the polarity problem in the clinical text-domain. In the fourth phase, with the help of seven features (word embedding, sentence embedding, sentiment rules, shifter sentiment, and three features based on linguistic, statistical, and medical domains), knowledge are to be used in training the model, for training we propose ELM-EA, it is a feed-forward neural network with a hidden layer, the weights for this layer are set randomly and the weights for the output layer are calculated from the training dataset [40]. The last stage is evaluating if method is suitable for evaluating the usefulness and trustfulness of the DS-SA-ELM-AE for testing qualities such as noisy information, eliminating redundancy, comprehensibility, and readability. There are three main measures for evaluating the worth of any method of F-measure, precision, and recall.

5. Experimental Results

The important issue for this research is whether the DS-SA-ELM-AE method performs better in sentiment analysis and achieves quality-of-services (QoS). In the next sub-sections, we will discuss the proposed model, and test the corpus with the English language generated by i2b2.

5.1. Data Collection

In order to examine the DSSA-ELM-AE, we performed the experiments on the 1237 de-identified discharge summaries, obesity disease, and 15 comorbidities as shown in Table 1, we adopted this table

from https://www.i2b2.org/NLP/Obesity/. The dataset is constructed for extracting the relationship among the medical terms. Our research is focused on experimenting with the health care and treatment quality based on techniques of sentiment analysis. In this research, the idea from8 is adopted. To obtain a more realistic result of our proposed approach, we need to build a gold standard data, which consists of the corrected or true results or labels. For this reason, we asked five annotators, three Ph.D. students in linguistics (English language), a teacher and a lecturer with good skills for reading, understanding, and teaching English language. All annotators were given the dataset and they labeled each sentence in the dataset using three terms, positive, negative, and neutral. The final polarity or label of the sentence was decided based on the majority of the votes of the annotators, in this way the dataset was labeled sentence-wise. We selected two diseases from sixteen diseases randomly, obesity with 606 summaries and asthma with 606 summaries. The goal of annotators is to build a gold standard for training data that is labeled at a sentence level, with 5320 sentences. The tags of sentiment for each sentence completed with polarity positive = 1, neutral = 0 and negative = −1 as shown in Table 2. We used these datasets for evaluating the final results.

Table 1. Statistics of i2b2 obesity corpus.

Disease	Present	Unmentioned	Absent	Questionable	Total
Asthma	75	529	1	1	606
CAD	331	240	16	4	591
CHF	239	344	7	0	589
Depression	90	519	0	0	609
Diabetes	396	181	12	6	595
Gallstones	93	513	3	0	609
GERD	98	500	1	3	602
Gout	73	534	0	2	609
Hypercholesterolemia	246	343	9	1	599
Hypertension	441	149	10	0	600
Hypertriglyceridemia	15	594	0	0	609
OA	89	513	0	0	602
Obesity	245	354	3	4	606
OSA	88	510	0	7	604
PVD	83	525	0	0	608
Venous Insufficiency	14	592	0	0	606
Sum	2616	6940	62	28	9644

"Present" means that each discharge summary content on the information about specific disease and also information about other related diseases. "Unmentioned" means that each discharge summary does not mentioned on the information about other related diseases. "Absent" means that each discharge summary contains only the information about specific diseases. "Questionable" means that each discharge summary may have information about other related diseases (https://www.i2b2.org/NLP/Obesity/).

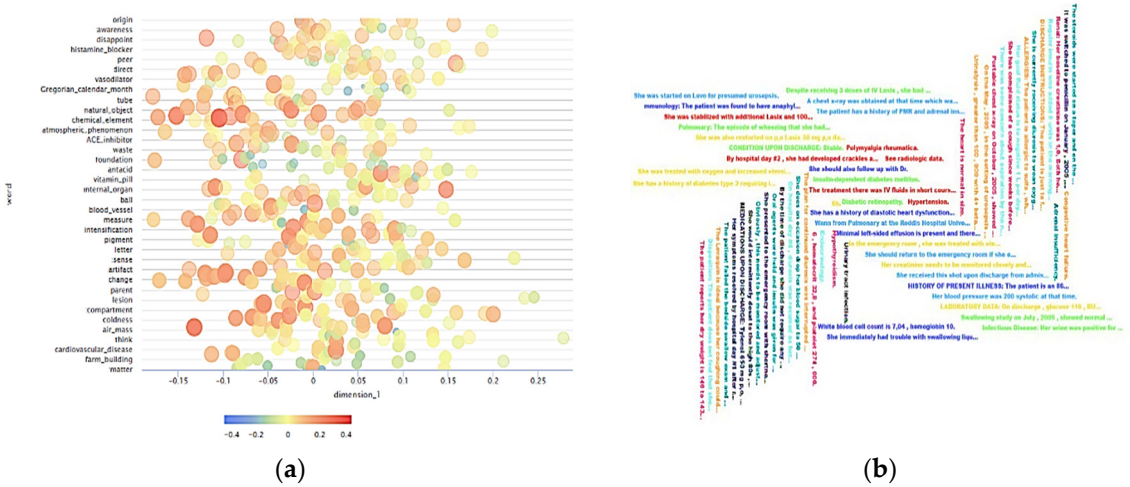
Table 2. Statistics of our gold standard corpus.

Disease	Positive	Negative	Neutral	Total
Asthma	331	37	238	606
Obesity	320	43	243	606

5.2. Text Presentation and Feature Extraction

Normalization [41] is the crucial stage of text processing and is generally called "pre-processing", this is done to the raw text so that textual data can be cleaned of spurious effects that are not important for further analysis. Pre-processing is called the text simplification method, and it generally consists of six sub-steps. We first started with text tokenization. In this step, we tokenized the text to a set of sentences. Then, we applied filtering to remove the stop word list, deleting all unnecessary words like "the", "in" and "on". After that, we applied a tokenization by length technique and selected 3 to 25 characters in each term. This led us to choose words with three letters at least. The fourth step was

to use stemming to transfer all the selected words, for example delete "ed", "ing", etc.). The fifth step is to use a fixed length vocabulary size, in our case, we used the simple bag-of-words idea that is that each word was assigned a corresponding numeric integer, we used the vocabulary size of most frequent first 10,000 words so that each word in the text got mapped to the corresponding integer for further processing on the numerical data. In the last step, we used the Word2Vec model to map those words to a fixed length vector so that semantic information between the words could be used in the further analysis as naive TF-IDF and fixed length word representation using naïve bag-of-words has now become obsolete because of not utilizing the semantic relations between the related words and is being replaced with more advanced embedding-based technologies consisting of Word2Vec, Glove and Gensim, etc. The results of this phase, after considering all the pre-processing steps (normalize corpora), are used as a feature set of word and document embedding based on word2vec and SV-BOW, sentiment-based rules knowledge, shifter, linguistic and statistical knowledge. One of the well-known techniques for word embedding is Word2Vec, it is divided into two models, namely: skip-grams, and a continuous bag-of-words model to obtain the word embedding vectors. Additionally, skip-grams are used for predicting the neighboring words given the target word. While continuous bag-of-words predicts the target words that rely on the context of text embedding [42,43]. In our analysis, we used a continuous bag-of-words approach and included some other strategies that will be described in the following discussion to deal with the various issues in the model. This approach was trained by utilizing Wikipedia, Google News and PubMed (https://www.ncbi.nlm.nih.gov/pubmed/) with 200 as the dimension size of each word in word embedding. The main parameters of the the model are the mininal vocabulary frequency of the word, the size of the layer, the size of text window, and the number of negative samples. After experimentation, the final values of those parameters used for the model were 30, 200, 5 and 5 for the above parameters, respectively. To solve the problem related to shifter words, parameters values of minimum confidence were set to 0.8, gain theta to 2.0 and Laplace k to 1.0. In Figure 2a the sample output for this phase is shown, where *Y*-axis illustrates the list of words while the *X*-axis illustrates the dimension−1, where the words appeared based on the information in the word vector, Figure 2b shows the sentence textual presentations based on the vector.



(**a**) (**b**)

Figure 2. (**a**) Words presentation based on the vector, and (**b**) sentence presentation based on the vector.

5.3. Lexicon Generation

One of the main issues in sentiment analysis is lexicon limitation. We divided this part into two steps as follows:

- The first step begins by comparing our bag-of-words method with UMLS and SentWordNet lexicon based on semantic sentiment approach, which suffers from a drawback that it neglects a neutral score. To remedy this, we used part of speech (PENN) tagging system (JJ.* |NN.*

|RB.* |VB.*) (https://cs.nyu.edu/grishman/jet/guide/PennPOS.html). Then constructed two lists of the words, the first one was our BOW, the second one is SentWordNet that was based on hypernyms technique.

- The second step is to transfer both lists with weights 1, 0, −1 pos, neu, neg respectively, based on the feature weights approaches like Chi-squared statistic (CSS), weighted by (SVM), and (PCA) [39]. These three different schemes offer us more flexibility to select the best weights for each term. Then, we used cosine similarity as a numerical measure to extract the similarity weights between these methods and both the lists. Based on the discussion above, the experiment was implemented on these three weight approaches to get better accuracy. The output of this phase will be used for assigning the training dataset based on document.
- sentiment polarity as in Equation (1). For more details see the (Supplementary file S1) which includes a full picture for each method and the main idea for the second and fourth phases present as workflow. Table 3 shows the samples from our dictionary list. Figure 3 shows the compared results based on seven cases.

$$\text{Document sentiment polarity} = (\text{neg} - \text{pos})/(\text{neg} + \text{pos} + \text{neu}) \quad (1)$$

Table 3. The positive, negative and neutral keyword list for each discharge summaries.

ID	Positive Keywords	Negative Keywords	Neutral Keywords
1	Positive, improve, increase	Failure, fatigue, weakness	Notify, totally, planned
2	Accept, strong	Disease, losing weight	Data, side, center
3	Effective, improve, clear, safe, good	Swollen, illness, loose, droop	Move, weigh, event
4	Bright, gain, stable, consistent, satisfactory	Itching, headache, smoker	Site, totally, tubes
5	Significant, thought, advanced, aspiration	Complained, work, deviation, upset	System, steadily, discussion

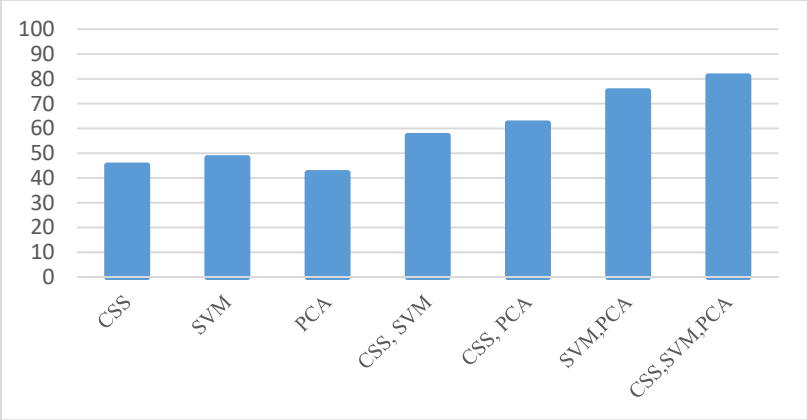


Figure 3. Illustration of the F1 measure when used in three weighing methods with seven different cases.

5.4. Evaluation and Comparison

The extreme learning machine autoencoder with 10-fold cross-validations for multi-classification task is then applied. The model has an input layer which contains the number of neurons as a number of features, N_{elm} parameter, the hidden layer which contains a number of nodes, these parameters need tuning, together with λ, and C that need to be tuned also. Weight and bias parameters for hidden layer are generated randomly [44]. Based on these parameters, we propose an optimized parameters grid search which has the capability for tuning the key parameters for the model in the sub-process to get the optimal parameters. Figure 4 shows the optimal key parameters optimization process based on grid search, whereas Figure 5a shows the λ optimal parameter values for asthma and obesity, respectively; Figure 5b shows the C optimal parameter values for asthma and obesity respectively; and Figure 5c shows the *Nelm* optimal parameter values for asthma and obesity diseases, respectively,

for two cases of our gold standard dataset. For the output layer, nodes number and activation function parameter values were chosen as 3 and sigmoid, respectively.

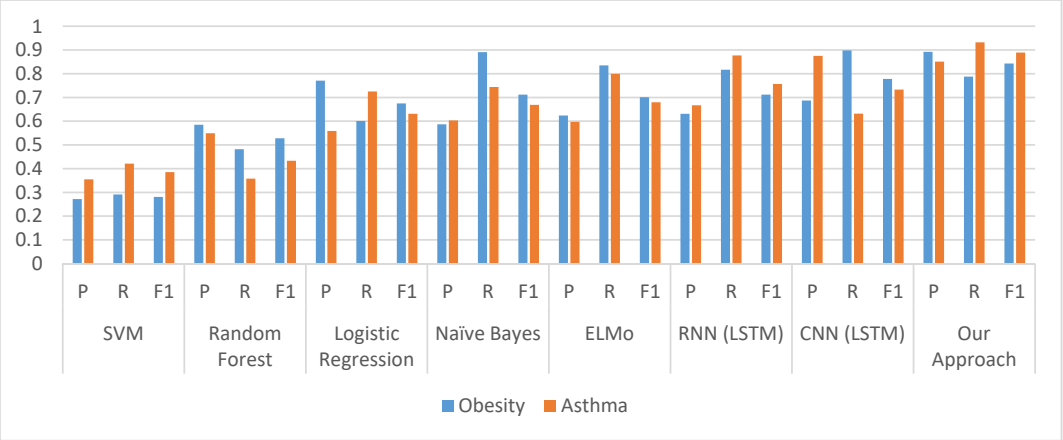
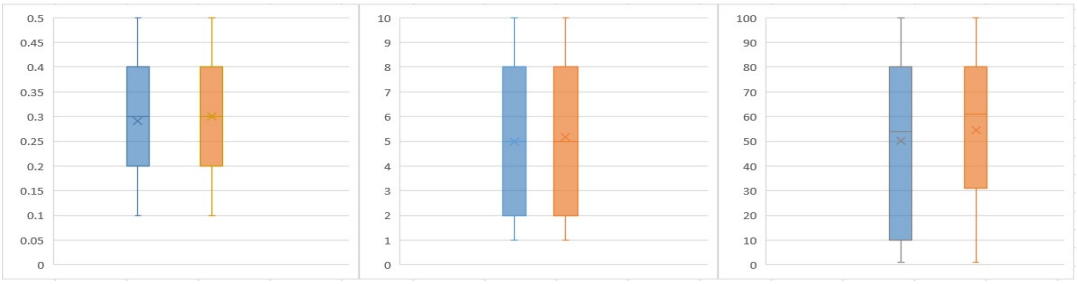


Figure 4. Performance statistics based on comparison between our approach and other approaches, using recall (R), precision (P), and F1 measures (F1).



(**a**) (**b**) (**c**)

Figure 5. The optimal parameters values optimization process, (**a**) values for λ parameter for asthma (blue color), and obesity (orange color), (**b**) values for C parameter for asthma and obesity, (**c**) values for *Nelm* parameter for asthma and obesity.

We examined the binary classification as the first experiment for our proposed model. This experiment performed four cases: asthma disease with positive and neutral labels, asthma disease with negative and neutral labels, obesity disease with positive and neutral labels, and obesity disease with negative and neutral labels. Figure 6a shows the results in terms of recall, precision, and the F1 measure, where Pos refers to positive, Neu refers to neutral, and Neg refers to negative.

In the second experiment for our proposed model, we used some neutral labels as the positive label (randomly) for the first round, then used them as negative labels (randomly) for the second round. We examine this case because there is a probability that when giving a score for the neutral from −1 to 1, maybe it is nearer to −1, then we put it into negative group, or it is nearer to 1, then we put it into positive group. Figure 6b shows the results in terms of recall, precision, and the F1 measure, where Pos refers to positive, Neu refers to neutral, and Neg refers to negative.

The third experiment for our proposed approach is implemented with multi-label. Figure 6c shows the results in terms of recall, precision, and the F1 measure. Then, we compare our final results with seven other approaches from the literature with our feature set and adopted the CNN (LSTM), and RNN (LSTM) methods for our comparison. Our gold standard data was also implemented to examine the quality of treatment (evaluating) and was compared with the results of our full-automatic polarity assign approach. To make this comparison unbiased, we applied 10-fold cross-validations with seven baseline approaches. Figure 4 shows the final result based on recall, precision, and the

F1 measure for each method compared with our approach that rely on our gold standard data set. Figure 7 shows the efficiency calculated for each method based on running time.

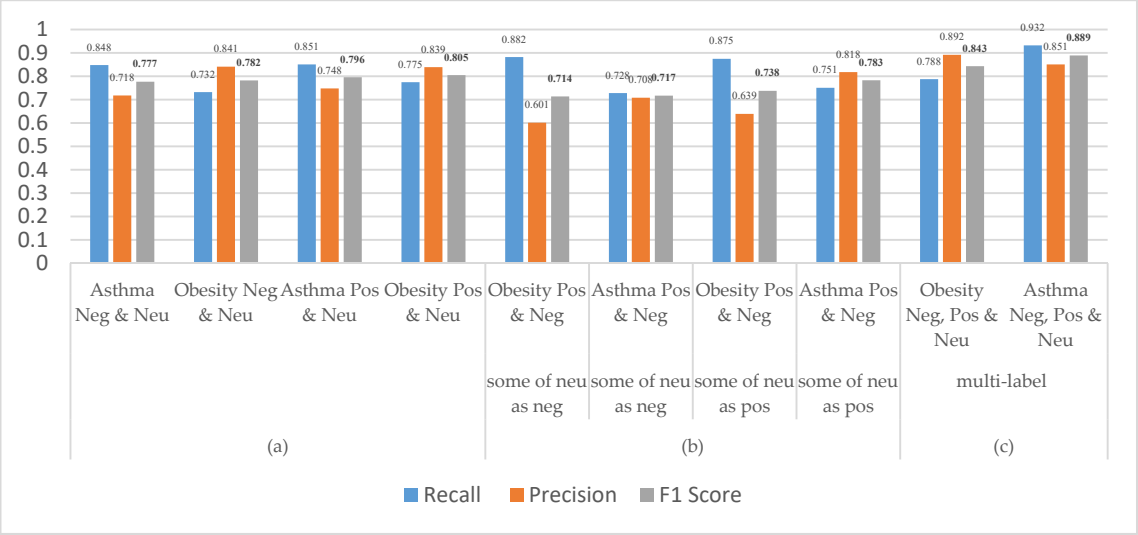


Figure 6. Performance statistics based on comparison between binary classification results with different cases (Pos, Neg or/and Neu), and multi-label classification, by using our method: (**a**) the results in terms of recall, precision, and the F1 measure; (**b**) the results in terms of recall, precision, and the F1 measure; (**c**) the results in terms of recall, precision, and the F1 measure.

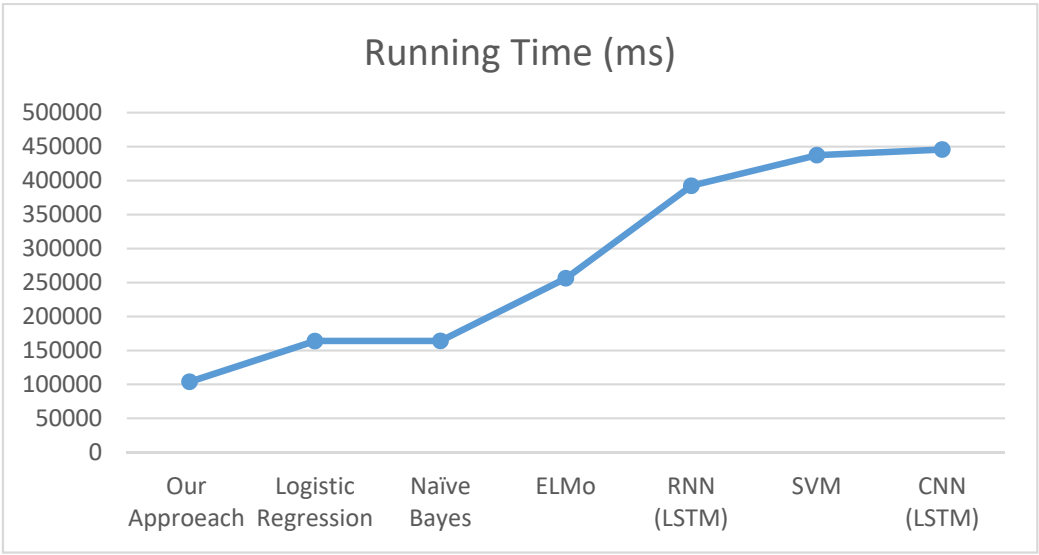


Figure 7. Time complexity for each method.

Error analysis in machine learning is generally reported using an accuracy metric and the results of varying our method are presented in Table 4, the experiments carried out varying based on the list of features. However, we get the best accuracy when we use all the features.

Overall accuracy also is calculated for each disease, the outcome for asthma is 0.747 and for obesity is 0.717.

Receiver operating curves (ROC)-based comparisons with multinomial logistic regression (MLR) and linear discriminant analysis (LDA) are shown in Figure 8 to analyze the results of our proposed approach with the inherent multi-class classifiers, that is, those classifiers that do not need on-vs-all like tricks to report accuracy for multi-class problems. A comparison of TPR and FPR on the axis is

displayed for each combination of the two-class subset from the three class set of "Positive", "Negative" and "Neutral" classes in Figure 5a–c, respectively. A similar pattern is observable here, that is, for the initial threshold our proposed method does not show promising results but gradually it performs much better than the rest of the inherent multi-class classifiers. Here, we have obtained evidence that our proposed approach and set of robust features outperform the classifiers that are inherently multi-class based.

Table 4. The results obtained from the experiments carried out varying the list of features.

Our Method	F1	F2	F3	F4	F5	F6	F7	Precision	Recall	*F*_measure
DS-SA-ELM-AE	+	+	+	+	+	+	+	0.851	0.932	0.889
DS-SA-ELM-AE	+	+	+	+	+	+		0.8675	0.6571	0.7478
DS-SA-ELM-AE	+	+	+	+	+			0.8176	0.6193	0.7048
DS-SA-ELM-AE	+	+	+	+				0.8404	0.6369	0.7246
DS-SA-ELM-AE	+	+	+					0.7869	0.5998	0.6807
DS-SA-ELM-AE	+	+						0.7826	0.6004	0.6795
DS-SA-ELM-AE	+							0.7469	0.5733	0.6487

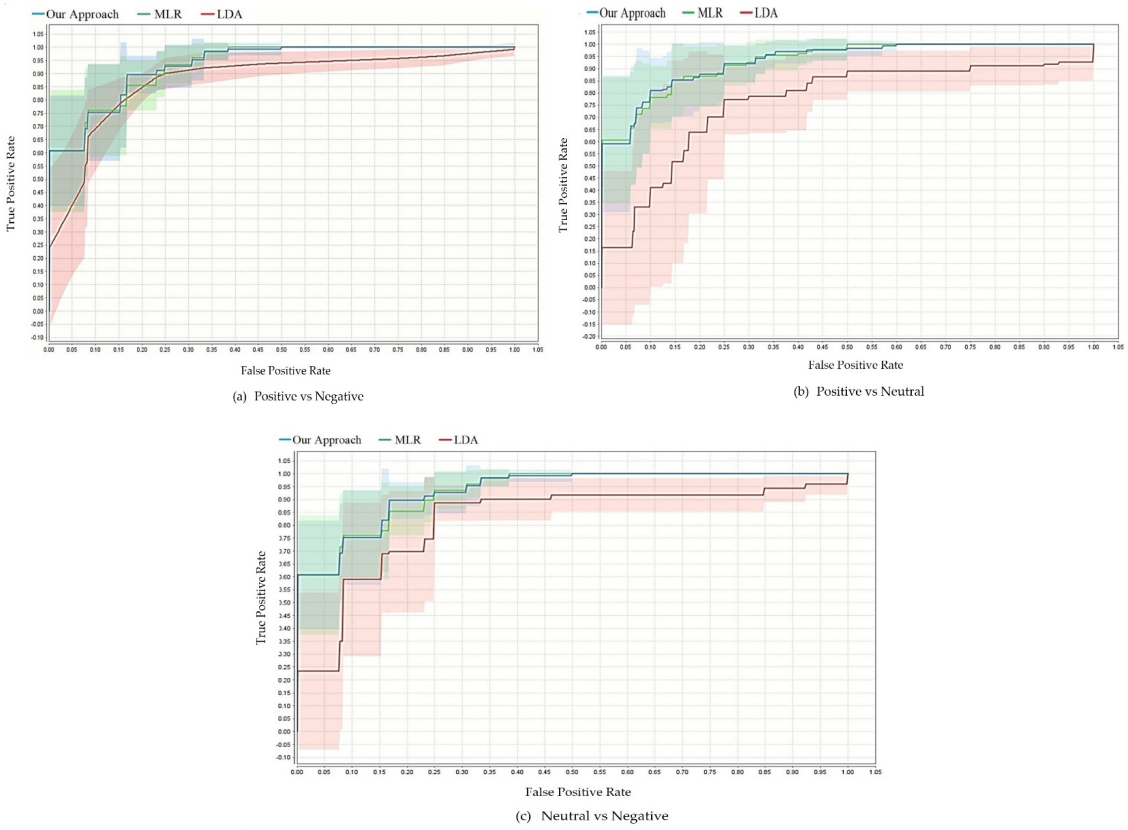


Figure 8. Receiver operating curves (ROC) curves comparison with the inherent multi-class classifiers.

These methods used the following standard metric as follows:

$$y_j = \sum_{i=1}^{\varphi} \beta_i g(w_i.x_i + b_i),\ j = 1, 2, \ldots, N \tag{2}$$

where, $w_i = [w_i, w_i, \ldots, w_{in}]^T$ links *i*th the input neurons and hidden neuron, *bi* presents the *i*th hidden neuron bias, and the connection between output neuron and the hidden neuron is presented by βi. Within matrix form.

$$y = \mathrm{H}\beta \tag{3}$$

where,

$$y = [y_1, y_2, \ldots y_n]^T \tag{4}$$

$$H = \begin{bmatrix} g(w_1 x_1 + b_1) & \ldots. & g\left(w_\varphi x_1 + b_\varphi\right) \\ \vdots & & \vdots \\ g(w_1 x_1 + b_1) & \ldots. & g\left(w_\varphi x_N + b_\varphi\right) \end{bmatrix}_{N \times \varphi} \tag{5}$$

$$\beta = [\beta_1, \beta_2, \ldots \beta_n]^T \tag{6}$$

here, **H** can be a matrix of non-square so there cannot exist βi, bi, $\mathbf{w}i$, where i = 1, 2, ..., N such as **y** = **H** β. The final square result of this method is

$$\beta = \mathrm{H}^{\dagger}\, y \tag{7}$$

here, $\mathbf{H}^{\dagger}$ is the Moore–Penrose popularize matrix **H** inverse [45].

6. Discussion

Our method used a semi-supervised approach which is based on vector space model and statistical techniques combined with ELM-EA. We focused on the text challenges based on the discharge summaries to examine and evaluate the treatment quality. This evaluation is implemented based on the list of sentences (list of features) for each discharge summary and a gold standard dataset with positive, negative, and neutral as labels, based on our novel method for constructing the lexicon (medical domain). These labels helped us to decide whether each document is positive, negative, or neutral and to evaluate quality of treatment. The combined multi-level of the feature set at the clinical text domain, which includes the word and sentence embedding, rules of sentiment shifter, sentiment knowledge, linguistic, medical concept, and statistical knowledge has been completely investigated.

Based on the abovementioned cases we examined, our suggested method uses two binary classification cases based on evaluation on gold standard dataset and reports the evaluation results in terms of precision, recall, and F1 metrics. For the first experiment, the highest F1 measure with obesity Pos and Neu labels is 0.805. For the second experiment, the highest F1 measure with asthma Pos and Neg (when some Neu as Pos) is 0.783. In the third experiment with multi-label, we performed the classification experiment with recall, precision, and F1 measure, the DSSA-ELM-AE outperforms other approaches, the high F1 measure based on Asthma data set evaluation is 0.889, and the high F1 measure with based on obesity data set evaluation is 0.843. Based on the above final results, we suggest using our proposed approach for predicting the quality of treatment for other diseases when providing the gold standard dataset for each disease.

The computational complexity of producing the threshold for our model depends on how to transform the list of sentences to a list of features, which was used to select for training and testing parts. Take for instance, if there are 100 sentences in each discharge summary, the computational complexity, a function of the total number of sentences in all the discharge summary, is a factor of 100, but if we produce the gist of each discharge summary by extracting a few sentences from each of them by applying list of features, the complexity of model as well as the density of the resultant classifier accuracy is predicted to a greater extent when compared with the golden standard dataset. Additionally, there is a significant decrease in the computational complexity in the prediction of the final result.

We used ELM because of its inherent capability to model complex phenomena and fast training, the combination the of weighted approach and feature selection that we used in our work has benefits or greater accuracy in the case of imbalanced datasets, as in our case, because traditional classifiers are biased towards the majority class and are not useful in the case of imbalanced datasets. Although the dataset is imbalanced, the features that we selected are more robust and compatible with the ELM

configuration as evident from the Figure. Moreover, our results also demonstrate from this behavior that ELM worked well as compared to other classifiers.

Moreover, our method is faster than other existing methods as explained in the time complexity section. This led us to use our proposed model for experiments with other available datasets to examine the quality of the treatment based on assigning labels as positive, negative, and neutral. On the other hand, for future work we will solely use deep learning-based methods, our proposed method will use different corpus and various languages, like Arabic and Chinese language, and also we will try to collect the real clinical data and build our carpus with three different languages Arabic, Chinese and English to give the research area more flexibility to examine various characteristic for these languages.

7. Conclusions

In our study, we developed a novel approach for evaluating the treatment quality by combining both unsupervised (skip-gram, SV-BOW, and association rules,) techniques and semi-supervised (extreme machine learning autoencoder) techniques. The word embedding, rules of sentiment shifter, sentiment knowledge, and linguistic and statistical knowledge as a feature set were used, these features resulted in better accuracy during evaluation. Moreover, we have extracted the most relevant feature set from the corpus. Our proposed classification technique proves an enhancement in terms of recall, precision, and F1 measure when compared with state-of-the-art approaches. Our method outperformed other methods based on the inspiration of sentiment technique proposed for examining the quality of treatment for discharge summaries as an objective corpus. In other word, the ELM-AE has reduced time complexity and greater capability for exploring an aspect of feature learning that encourages us to propose this model for clinical data as a novel machine learning method.

Supplementary Materials: The following are available online at http://www.mdpi.com/2078-2489/11/5/281/s1, File1, Figure S1: shows the full picture of attribute weights (Chi), Figure S2: shows the full picture of attribute weights (PCA), Figure S3: shows the full picture of attribute weights (SVM) and Figure S4: shows the main idea for our study.

Author Contributions: X.S. conceived the idea. S.A.W. proposed the model, architecture, and devised implementing strategy for the proposed model. N.A.K. helped in editing manuscript and the implementation of the methodology. B.C. checked the overall progress of the methodology, results, and suggested edits. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. U1811262).

Acknowledgments: We are very grateful to Chinese Scholarship Council scholarship (CSC) for providing us financial and moral support.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Kilgour, C.; Bogossian, F.; Callaway, L.; Gallois, C.J.D.R. Experiences of women, hospital clinicians and general practitioners with gestational diabetes mellitus postnatal follow-up: A mixed methods approach. *Diabetes Res. Clin. Pract.* **2019**, *148*, 32–42. [CrossRef] [PubMed]
2. O'Connor, R.; O'Callaghan, C.; McNamara, R.; Salim, U.J.I.J.O.M.S. An audit of discharge summaries from secondary to primary care. *Ir. J. Med. Sci.* **2019**, *188*, 537–540. [CrossRef] [PubMed]
3. Sun, W.; Cai, Z.; Li, Y.; Liu, F.; Fang, S.; Wang, G.J.J.O.H.E. Data processing and text mining technologies on electronic medical records: A review. *J. Healthc. Eng.* **2018**, *2018*, 4302425. [CrossRef] [PubMed]
4. Hanauer, D.A.; Mei, Q.; Law, J.; Khanna, R.; Zheng, K.J.J.O.B.I. Supporting information retrieval from electronic health records: A report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). *J. Biomed. Inform.* **2015**, *55*, 290–300. [CrossRef] [PubMed]
5. Rumshisky, A.; Ghassemi, M.; Naumann, T.; Szolovits, P.; Castro, V.; McCoy, T.; Perlis, R.J.T.P. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Nat. Transl. Psychiatry* **2016**, *6*, e921. [CrossRef] [PubMed]

6. Tsopra, R.; Wyatt, J.C.; Beirne, P.; Rodger, K.; Callister, M.; Ghosh, D.; Clifton, I.J.; Whitaker, P.; Peckham, D.J.J.O.E.I.C.P. Level of accuracy of diagnoses recorded in discharge summaries: A cohort study in three respiratory wards. *J. Eval. Clin. Pract. Wiley Online Libr.* **2019**, *25*, 36–43. [CrossRef]
7. Graham, A.J.; Ocampo, W.; Southern, D.A.; Falvi, A.; Sotiropoulos, D.; Wang, B.; Lonergan, K.; Vito, B.; Ghali, W.A.; McFadden, S.D.P.J.B.Q.S. Evaluation of an electronic health record structured discharge summary to provide real time adverse event reporting in thoracic surgery. *BMJ Qual. Saf.* **2019**, *28*, 310–316. [CrossRef]
8. Goldgrab, D.; Balakumaran, K.; Kim, M.J.; Tabtabai, S.R.J.H.F.R. Updates in heart failure 30-day readmission prevention. *Heart Fail. Rev.* **2019**, *24*, 177–187. [CrossRef]
9. Gilbert, A.V.; Patel, B.K.; Roberts, M.S.; Williams, D.B.; Crofton, J.H.; Morris, N.M.; Wallace, J.; Gilbert, A.L.J.J.O.P.P. An audit of medicines information quality in electronically generated discharge summaries–evidence to meet the Australian National Safety and Quality Health Service Standards. *J. Pharm. Wiley Online Libr.* **2017**, *47*, 355–364. [CrossRef]
10. Schwarz, C.M.; Hoffmann, M.; Schwarz, P.; Kamolz, L.P.; Brunner, G.; Sendlhofer, G.J.B.H.S.R. A systematic literature review and narrative synthesis on the risks of medical discharge letters for patients' safety. *BMC Health Servres* **2019**, *19*, 158. [CrossRef]
11. Liang, H.; Tsui, B.Y.; Ni, H.; Valentim, C.C.; Baxter, S.L.; Liu, G.; Cai, W.; Kermany, D.S.; Sun, X.; Chen, J.J.N.M. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. *Nat. Med.* **2019**, *25*, 433. [CrossRef] [PubMed]
12. Reátegui, R.; Ratté, S.J.B.M.I.; Making, D. Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med. Inform. Decis. Mak.* **2018**, *18*, 74. [CrossRef] [PubMed]
13. Servid, S.A.; Noble, B.N.; Fromme, E.K.; Furuno, J.P.J.J.O.T.A.G.S. Clinical intentions of antibiotics prescribed upon discharge to hospice care. *J. Am. Heart Assoc. Wiley Online Libr.* **2018**, *66*, 565–569. [CrossRef]
14. Xu, J.; Gan, L.; Cheng, M.; Wu, Q.J.J.O.H.E. Unsupervised medical entity recognition and linking in Chinese online medical text. *J. Healthc. Eng.* **2018**, *2018*, 1–13. [CrossRef] [PubMed]
15. Jiménez-Zafra, S.M.; Martín-Valdivia, M.T.; Molina-González, M.D.; Ureña-López, L.A.J.A.I.I.M. How do we talk about doctors and drugs? Sentiment analysis in forums expressing opinions for medical domain. *Artif. Intell. Med.* **2019**, *93*, 50–57. [CrossRef] [PubMed]
16. Abualigah, L.; Alfar, H.E.; Shehab, M.; Hussein, A.M.A. Sentiment Analysis in Healthcare: A Brief Review. In *Recent Advances in NLP: The Case of Arabic Language*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 129–141.
17. Melo, P.F.; Dalip, D.H.; Junior, M.M.; Gonçalves, M.A.; Benevenuto, F.J.J.O.T.A.F.I.S. 10SENT: A stable sentiment analysis method based on the combination of off-the-shelf approaches. *J. Assoc. Inf. Sci. Technol.* **2019**, *70*, 242–255. [CrossRef]
18. Al-Smadi, M.; Talafha, B.; Al-Ayyoub, M.; Jararweh, Y.J.I.J.O.M.L. Using long short-term memory deep neural networks for aspect-based sentiment analysis of Arabic reviews. *Int. J. Mach. Learn. Cybern.* **2019**, *10*, 2163–2175. [CrossRef]
19. Ghasemaghaei, M.; Eslami, S.P.; Deal, K.; Hassanein, K. Consumers' attitude toward insurance companies: A sentiment analysis of online consumer reviews. *Decision Support and Analytics (SIGDSA)*. 2016. Available online: https://aisel.aisnet.org/amcis2016/Decision/Presentations/10/ (accessed on 21 May 2020).
20. Rezaeinia, S.M.; Ghodsi, A.; Rahmani, R.J.A.P.A. Improving the accuracy of pre-trained word embeddings for sentiment analysis. *arXiv* **2017**, arXiv:1711.08609.
21. Sankar, H.; Subramaniyaswamy, V.; Vijayakumar, V.; Arun Kumar, S.; Logesh, R.; Umamakeswari, A.J.S.P. Intelligent sentiment analysis approach using edge computing-based deep learning technique. *Softw. Pract. Exp. Wiley Online Libr.* **2019**. [CrossRef]
22. Wang, Y.; Youn, H.J.A.S. Feature Weighting Based on Inter-Category and Intra-Category Strength for Twitter Sentiment Analysis. *Appl. Sci.* **2019**, *9*, 92. [CrossRef]
23. Dehkharghani, R.; Saygin, Y.; Yanikoglu, B.; Oflazer, K.J.L.R. SentiTurkNet: A Turkish polarity lexicon for sentiment analysis. *Lang. Resour. Eval.* **2016**, *50*, 667–685. [CrossRef]
24. Wang, Y.; Rao, Y.; Wu, L. A review of sentiment semantic analysis technology and progress. In Proceedings of the 2017 13th International Conference on Computational Intelligence and Security (CIS), Hong Kong, China, 15–18 December 2017; pp. 452–455.
25. Mohammad, S.M.; Kiritchenko, S.; Zhu, X.J.A.P.A. NRC-Canada: Building the state-of-the-art in sentiment analysis of tweets. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, PA, USA, 6–7 August 2019; Volume 1.

26. Gilmore-Bykovskyi, A.L.; Kennelty, K.A.; DuGoff, E.; Kind, A.J.J.B.H.S.R. Hospital discharge documentation of a designated clinician for follow-up care and 30-day outcomes in hip fracture and stroke patients discharged to sub-acute care. *BMC Health Servres* **2018**, *18*, 103. [CrossRef] [PubMed]
27. Mehta, R.L.; Baxendale, B.; Roth, K.; Caswell, V.; Le Jeune, I.; Hawkins, J.; Zedan, H.; Avery, A.J.J.B.H.S.R. Assessing the impact of the introduction of an electronic hospital discharge system on the completeness and timeliness of discharge communication: A before and after study. *BMC Health Servres* **2017**, *17*, 624. [CrossRef] [PubMed]
28. Ooi, C.E.; Rofe, O.; Vienet, M.; Elliott, R.A.J.I.J.O.C.P. Improving communication of medication changes using a pharmacist-prepared discharge medication management summary. *Int. J. Clin. Pharm.* **2017**, *39*, 394–402. [CrossRef] [PubMed]
29. Pereira-Kohatsu, J.C.; Quijano-Sánchez, L.; Liberatore, F.; Camacho-Collados, M.J.S. Detecting and Monitoring Hate Speech in Twitter. *Sensors* **2019**, *19*, 4654. [CrossRef]
30. Flores, A.C.; Icoy, R.I.; Peña, C.F.; Gorro, K.D. An Evaluation of SVM and Naive Bayes with SMOTE on Sentiment Analysis Data Set. In Proceedings of the 2018 International Conference on Engineering, Applied Sciences, and Technology (ICEAST), Phuket, Thailand, 4–7 July 2018; pp. 1–4.
31. Ahmad, M.; Aftab, S.; Bashir, M.S.; Hameed, N.; Ali, I.; Nawaz, Z.J.I.J.A.C.S.A. SVM optimization for sentiment analysis. *Int. J. Adv. Comput. Sci. Appl.* **2018**, *9*, 393–938. [CrossRef]
32. Gupta, S.; Jain, S.; Gupta, S.; Chauhan, A.J.I.J.O.A.R.I.C.S. Opinion Mining for Hotel Rating through Reviews Using Decision Tree Classification Method. *Int. J. Adv. Res. Comput. Sci.* **2018**, *9*, 180. [CrossRef]
33. Ma, Y.; Peng, H.; Khan, T.; Cambria, E.; Hussain, A.J.C.C. Sentic LSTM: A hybrid network for targeted aspect-based sentiment analysis. *Cogn. Comput.* **2018**, *10*, 639–650. [CrossRef]
34. Spinczyk, D.; Nabrdalik, K.; Rojewska, K.J.B.E.O. Computer aided sentiment analysis of anorexia nervosa patients' vocabulary. *Biomed. Eng. Online* **2018**, *17*, 19. [CrossRef]
35. Jiang, K.; Feng, S.; Song, Q.; Calix, R.A.; Gupta, M.; Bernard, G.R.J.B.B. Identifying tweets of personal health experience through word embedding and LSTM neural network. *BMC Bioinform.* **2018**, *19*, 210. [CrossRef]
36. LeCun, Y.; Bengio, Y.; Hinton, G.J.N. Deep learning. *Nature* **2015**, *521*, 436–444. [CrossRef] [PubMed]
37. Sun, K.; Zhang, J.; Zhang, C.; Hu, J.J.N. Generalized extreme learning machine autoencoder and a new deep neural network. *Neurocomputing* **2017**, *230*, 374–381. [CrossRef]
38. Waheeb, S.A.; Husni, H.J.I.J.O.A.I.S. Multi-Document Arabic Summarization Using Text Clustering to Reduce Redundancy. *Int. J. Adv. Sci. Technol.* **2014**, *2*, 194–199.
39. Waheeb, S.A.K.N.A.; Chen, B.; Shang, X. Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy. *Information* **2020**, *11*, 59.
40. Huang, G.B.; Zhu, Q.Y.; Siew, C.K.J.N. Extreme learning machine: Theory and applications. *Neurocomputing* **2006**, *70*, 489–501. [CrossRef]
41. Reese, R.M. *Natural Language Processing with Java*; Packt Publishing Ltd.: Birmingham, UK, 2015.
42. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J.J.A.P.A. Efficient estimation of word representations in vector space. *arXiv* **2013**, arXiv:1301.3781.
43. Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.S.; Dean, J. Distributed representations of words and phrases and their compositionality. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 5–8 December 2013; pp. 3111–3119.
44. Yao, L.; Ge, Z.J.I.T.O.I.E. Deep learning of semisupervised process data with hierarchical extreme learning machine and soft sensor application. *IEEE Trans. Ind. Electron.* **2017**, *65*, 1490–1498. [CrossRef]
45. Huang, G.; Huang, G.B.; Song, S.; You, K.J.N.N. Trends in extreme learning machines: A review. *Neural Netw.* **2015**, *61*, 32–48. [CrossRef]