Computer Systems Science & Engineering
DOI: 10.32604/csse.2023.035753
Article

Tech Science Press

Discharge Summaries Based Sentiment Detection Using Multi-Head Attention and CNN-BiGRU

Samer Abdulateef Waheeb*

School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an, 710072, China
*Corresponding Author: Samer Abdulateef Waheeb. Email: samirabdulateef@mail.nwpu.edu.cn
Received: 02 September 2022; Accepted: 13 November 2022

Abstract: Automatic extraction of the patient's health information from the unstructured data concerning the discharge summary remains challenging. Discharge summary related documents contain various aspects of the patient health condition to examine the quality of treatment and thereby help improve decision-making in the medical field. Using a sentiment dictionary and feature engineering, the researchers primarily mine semantic text features. However, choosing and designing features requires a lot of manpower. The proposed approach is an unsupervised deep learning model that learns a set of clusters embedded in the latent space. A composite model including Active Learning (AL), Convolutional Neural Network (CNN), BiGRU, and Multi-Attention, called ACBMA in this research, is designed to measure the quality of treatment based on discharge summaries text sentiment detection. CNN is utilized for extracting the set of local features of text vectors. Then BiGRU network was utilized to extract the text's global features to solve the issues that a single CNN cannot obtain global semantic information and the traditional Recurrent Neural Network (RNN) gradient disappearance. Experiments prove that the ACBMA method can demonstrate the effectiveness of the suggested method, achieve comparable results to state-of-arts methods in sentiment detection, and outperform them with accurate benchmarks. Finally, several algorithm studies ultimately determined that the ACBMA method is more precise for discharge summaries sentiment analysis.

Keywords: Sentiment analysis; lexicon; discharge summaries; active learning; multi-head attention mechanism

1 Introduction

Sentiment Analysis (SA) uses natural language processing, computational linguistics, and textual analysis to detect neutral, negative, or positive feelings from the text data. The main aim of SA is to identify sentiment in a text about a subject or predominant emotion of a document expressed by the authors. Mining into these opinions is not only a non-trivial task but a helpful process in obtaining or converting this vast amount of textual data into a valuable resource, a process known as converting unstructured data to structured data [1]. Although the conversion of unstructured text into structured or

meaningful information is the primary purpose of SA [2], it contains challenges related to classification and identifying meaningful patterns from the data [3].

Considering the fundamental and applied significance of SA in various domains, our proposed study applies it to the clinical text (health reports or medical history) based on a patient's Discharge Summaries (DS). In the medical domain, such analyses can enable the doctors to determine the treatment quality of the patient based on the opinions written by the medical practitioners in the form of discharge summaries. Earlier studies on SA emphasized the data extraction from the corpus, such as company reviewers and Twitter, without considering limiting the information extraction process obtained from automated data scrappers [4]. The Deep Learning (DL) models containing (attention mechanisms, RNN, and CNN) used with Natural language processing (NLP) tasks are getting, of late, very common because of the representational power of deep learning-based models. These DL techniques should be improved to deal with DS-based analysis in the medical domain [5]. Our proposed hybrid approach tackles the task of SA based on the dataset obtained after processing blogs from the micro-blogging platforms related to clinical text [6].

This study proposes an "End-to-End" model [7] using a Convolution Neural Network and a Recurrent Neural to learn useful features from the sequences of data samples effectively. Cyclic Neural Networks, where there are both forward, and backward connections between neurons, have gradually developed into Long Short-Term Memory (LSTM) and a Bidirectional (Bi) BiLSTM with a capacity to remember longer data sequences. LSTM was explicitly designed to address the long-term dependency issue with conventional RNNs, namely vanishing and exploding gradients. The BiLSTM model can examine a significant amount of text data from the text more effectively than the unidirectional LSTM model because of both forward and backward connections [8]. The proposed composite model not only learns the valuable features related to discharge summary texts but also uses the attention focuses on those aspects of the features that are most related to the diseases.

Our contribution in this paper is a development of a DL model consisting of Active Learning (AL), Convolutional Neural Network (CNN), BiGRU, and Multi-Attention, which is henceforth named (ACBMA). Our algorithm initially extracts text features using CNN and then uses BiGRU to extract global text features so that the issues of single CNN weakening to gain all semantic data and the vanishing of conventional round neural network gradients can be resolved. Our deep learning technique is combined with a manually chosen feature template. In addition to simplifying the structure, it has performed well compared to state-of-the-art models. The significant contributions are listed below

- This paper proposes an End-To-End Deep Neural Network model for the classification of the discharge summary to examine the treatment quality based on feature selection and extraction. This research achieves promising results using the proposed approaches, and our techniques have outperformed the state-of-the-art methods found in the literature.
- This paper proposes a weak supervised neural network-based model to construct a domain-specific sentiment lexicon by leveraging the feature distribution of embedding space in a specific domain. The proposed model is trained unsupervised, thus facilitating the construction of various lexicons in low-resource languages.
- The proposed new method based on the cross-feature combinations that relied on the attention technique could efficiently fuse these features at various levels, thus offering more efficient and accurate data required for the sentiment classifier tasks.
- This paper conducts extensive experiments on English-language benchmark datasets for the sentence-level task. Our experimental results have shown that the proposed domain-specific lexicon can effectively boost the performance of both supervised and unsupervised models. Furthermore, the experimental results have also demonstrated that Attention-based Score Embedding can significantly improve the performance of Text Sentiment Detection compared to the baseline methods.

2 Related Work

Sentiment analysis is one of the most active research areas in NLP, which attempts to find intelligent models based on clinical textual sentiment data that can model the sentiment orientation, for instance, the positive, negative, and neutral categories of the sentences [9]. The sentiment analysis methods were classified into four major categories: lexicon, machine learning, hybrid approaches, and Graph-Based [10,11]. SA is being applied in a range of domains like the healthcare and medical field [12], Business intelligence [13], Recommendation systems [14], and Government intelligence [15].

Deep Learning techniques for sentiment analysis in recent years have been getting popular. The authors in [16] proposed a sentiment detection method using the BiLSTM model that uses multi-channel features and a self-attention mechanism. In [17], the authors suggested the attention mechanism that uses CNN-LSTM for learning universal sentence representations within embedded models and presented the attention mechanism. The encoder of CNN is tiny in size, suited for slightly embedded models, and has excellent performance, according to experimental data. The work in [18] proposed a sentiment detection approach that relies on a converter that enhances Twitter's tweet quality by utilizing deep, intelligent context embedding and encoding the representations in the converter. The words semantic knowledge, syntactic, polysemous, and emotional are taken into consideration at the same time. Cheng [19] proposed the bidirectional GRU multi-head attention capsule and Multi-channel convolution, which applied vector neurons to change scalar neurons to the sentiment of model text, and applied capsules for characterizing text sentiment, and the final F1 score was around 90% based on their experiments. Chu et al. [20] proposed a model of multi-scale convolution that relies on the BiGRU-Word2Vec and multi-head attention mechanism model. The absolute accuracy was around 91.3% based on their experiments.

The standard LSTM considers sentiment tasks as a text classification problem. ATAE-LSTM [21] introduced an attention-based LSTM to model the relation between the aspect and its context by learning the aspect matrix. SenticLSTM [22], an attentive biLSTM, was introduced by leveraging the external information as commonsense knowledge from SenticNet to address ABSA. Lexicon-based ML [23], an ML-based solution for ATSA, introduced modeling the connection between sentences and leveraging the explicit features. ATAE-LSTM [24] presented an attention-based LSTM to learn aspect representation, encouraging LSTM to predict polarity in response to an aspect.

Based on the research mentioned above, In the early stages of sentiment detection research, it was discovered that machine learning methods produced promising results. However, it is incredibly challenging for those models concerning sentiment detection issues due to the large dataset volume. In CNN-based models for sentiment classification, most deep learning-based sentiment analysis researchers failed to consider the context relation, whereas the LSTM models only view the above relation with a slow convergence rate. This paper solved these issues using the BiGRU model with a bidirectional sequential structure, but the BiGRU model's direct application might cause unnecessary computational overhead due to the dimension of highly high input. The CNN model was used to increase the vector-matrix formed dimension of the word by utilizing the original information, then combined with the BiGRU technique for sentiment detection. Finally, a multi-head attention mechanism was added to the model to boost its operational effectiveness and prediction precision. The experimental results demonstrate the effectiveness of the proposed ACBMA model.

3 Proposed Method

Our proposed ACBMA algorithm detects the sentiment of the documents related to the discharge summaries. The suggested model is a hybrid method based on the word embedding aspects, Lexicon Generation, Active Learning method, CNN, BiGRU, and multi-head attention technique, called ACBMA algorithm. Decomposing discharge summaries into several words and then using those words in word

vectors is essential to train the deep learning model effectively. The preprocessing steps are data collection, format conversion CSV-TXT, then word token removed stop word list, training word vector, and saving the results. The word vectors were trained using Word2Vec [25]. Word2Vec uses a shallow neural network technique with three layers: input, projection, and output. Different neural networks can use the resulting word vector as input in various tasks. The two primary models in Word2Vec are Skip-gram and CBOW (continuous bag of words). This paper is used the famous Skip-gram in our proposed model as Skip-gram produces word embeddings from the existing headword while CBOW produces the existing headword from the text information of the word. The Skip-gram technique aims to create datasets for input and output. Firstly, the known words and their context are established in a dataset, and the window size is set. Next, the input and the target words of window size are combined to create such a dataset. The structure of the Skip-gram model contains three layers input, mapping, and output layers. The known word w_t $(w_{t-n},\ w_{t-n-1},\ \ldots,\ w_{t-1},\ w_t,\ \ w_{t+1},\ \ w_{t-n+1},\ \ldots,\ \ w_{t+n}$ there are $2n$ words like goal words, and the attaining the goal word probability is $p(count(w)|w)$ [12,26,27]. The next Equation shows the objective function:

$$F = \sum_{w \in N} log_2^{p(count(w)|w)} \tag{1}$$

Polarity problems [28] are tackled in our study by polarizing the list of sentences and then creating a specific vocabulary (building a new lexicon) related to the discharge summary. The subjectivity and polarity of the text are then determined by applying unsupervised techniques and accessible terminologies, such as SentiWordNet, TextBlob, Unified Medical Language System (ULMS), and Valence Aware Dictionary and sentiment Reasoner (VADSR) [27]. In our work, we first used the textual data, a massive unstructured set of discharge summaries, to develop a structured and labeled dataset. Then created, the gold standard dataset was created using human input to compare the results of the automated approach with the gold standard dataset.

The CNN model inside our proposed approach consists of an input layer, convolutional layers, pooling layers, and fully connected layers. The input layer uses the textual data after first tokenizing the text so that a row in the input layer corresponds to each word vector in the sentence. For example, in processing the sentence in the comment text into n words, using Word2Vec to turn each word into a vector, and then mapping each vector into a vector of $m - dimensional$. The word order of the sentence is then divided up and mapped onto a matrix of $n\ \times\ m - dimensional$. The convolutional layer creates a feature map using a convolution kernel to perform convolution computation on input. h shows the convolution window size, $k\ \times\ m$ presents the convolution kernel size. Steps of k words in agreement with the stages of length t, and use the convolution kernel to execute the convolution operation to extract the text local features on the windows of the input word $x_1^h,\ x_2^{h+1},\ x_3^{h+2},\ \ldots,\ x_{n-h+1}^n$. d shows the input sentence is collected of n word vectors $x_1,\ x_2,\ \ldots,\ x_n$, the convolutional layer procedure can be expressed as

$$y_i = f\ (W.\ x_{i:i+h-1} + b), \tag{2}$$

where f refers to the function of the nonlinear weight matrix shown by W, the dimension of the convolution presents by h, $x_{i:i+h-1}$ presents vector combination $x_i,\ \ x_{i+1},\ \ \ldots,\ \ x_{i+h-1}$, and the bias vector shown by b. after convolution kernel extraction, the eigenvector y is

$$y = \{y_1,\ \ y_2,\ \ y_3,\ \ldots,\ \ y_{n-h+1}\}. \tag{3}$$

Feature extractors are convolution kernels; both a single convolution kernel and a multi-convolution kernel serve as feature extractors. The effectiveness of convolution kernels is often increased by feature extraction using a bunch of convolution kernels. For describing the combination of z convolution kernels, $[P_1,\ \ P_2,\ \ P_3,\ \ldots,\ \ P_Z]$ is used, where P_Z denotes the $z - th$ size convolution kernel or the longitudinal dimension of the window of the convolution kernel. The convolution kernel window's horizontal size is

referred to as the vector dimension of the word vector. Computing z convolution kernels will result in the creation of z feature map vectors. After going through the pooling process, a transformation of a multidimensional vector into a value is employed as a component of the pooled vector. The pooling layer applies pooling processing to each eigenvector. The pooling layer receives the output sequence from the convolutional layer and utilizes the maximum pooling algorithm as its preferred pooling method. The maximum pooling technique will choose the significant element from the series of $y_1,\ y_2,\ y_3,\ \ldots,\ y_{n-h+1}$ to construct a new vector of y.

$$y = \max(y_i). \tag{4}$$

This paper has utilized a version of RNN called GRU (Gated Recurrent Unit), which, compared to LSTM, is recommended to address issues like gradients in backpropagation and long-term memory. The issue of RNN gradient fading and its inability to learn long-term historical load characteristics were addressed by researchers, who proposed LSTM as a solution. The Long Short-Term Sequence (LSTS) data can be combined with LSTM to understand the correlation information. Recently, GRU was established in answer to the issue of LSTM having a slow convergence rate and too many parameters. The LSTM has significant learning capabilities. However, a variant known as the GRU has fewer parameters and demonstrates speedier convergence performance. The BiGRU network can learn the correlation between present load and factors that affect previous and future loads, which is more beneficial for obtaining the deep characteristics of load data [29].

The following Equations can be calculated in the hidden layer:

$$z_t = \sigma\ (W_Z.\ [h_{t-1},\ x_t]), \tag{5}$$

$$r_t = \sigma\ (W_r.\ [h_{t-1},\ x_t]), \tag{6}$$

$$\tilde{h}_t = tanh\ (W_Z.\ [r_t \times h_{t-1},\ x_t]), \tag{7}$$

$$h_t = (1 - z_t) \times\ h_{t-1} + z_t \times \tilde{h}_t, \tag{8}$$

Most researchers applied the cross-entropy loss function for text classification problems in neural networks, and the cross-entropy is often utilized as the loss function. The cross-entropy is used to calculate each category's probability; cross-entropy appears with the softmax function practically every time [30].

Recently, many deep learning models have made substantial use of the attention mechanism, and whether it be speech recognition, picture processing, or natural language processing, attention mechanisms are widely used in various tasks [31]. The secret to writing effective discharge summaries is to pay close attention to word vectors while disregarding word vectors that have no bearing on the discharge summary's context. The word vectors of important discharge summaries words can become the primary source of information by using an attention technique to the input context data, making the entire neural network model more accurate and effective. The attention technique, which also instructs the model to pay attention to the discharge summary phrases when other similar sentences occur, will improve the model's learning and generalization abilities. The general attention technique includes the self-attention technique as a particular case. The attention-related value, key, and query matrix are denoted by the letters V, K, and Q, respectively. V = K = Q in the self-attention process. It offers the advantage of immediately computing the dependency relationship without considering word spacing. It can also figure out the relationships between the words in a sentence and comprehend the fundamental structure of a sentence [32]. The CNN model application enhances the neural network's interpretability and model learning capabilities when combined with RNN. Fig. 1 shows the basic structure of multi-head attention. Adding two linear projection matrices $E_i,\ F_i \in \mathbb{R}^{n*k}$, it is the key concept to calculating value and key. The first step is to project the original $n * d - dimensional$ value and key layers $VW_i^v\ and\ KW_i^k$ into $k * d - dimensional$ projected the layers of value and key. Next, calculate $n * k - dimensional$ context mapping P applying scaled-dot product attention. The scaled

dot-product attention at the central position differs from the expected attention. Given matrices $V \in R^{n*d}$, $K \in R^{n*d}$, $and\ Q \in R^{n*d}$, scaled dot-product attention using the following Equation:

$$head_i = \text{Attention}\left(QW_i^Q,\ E_iKW_I^K,\ F_iVW_i^V\right) \tag{9}$$

$$head_i = \text{softmax}\left(\frac{QW_i^Q(E_iKW_i^K)^T}{\sqrt{d_k}}\right).F_iVW_i^V \tag{10}$$

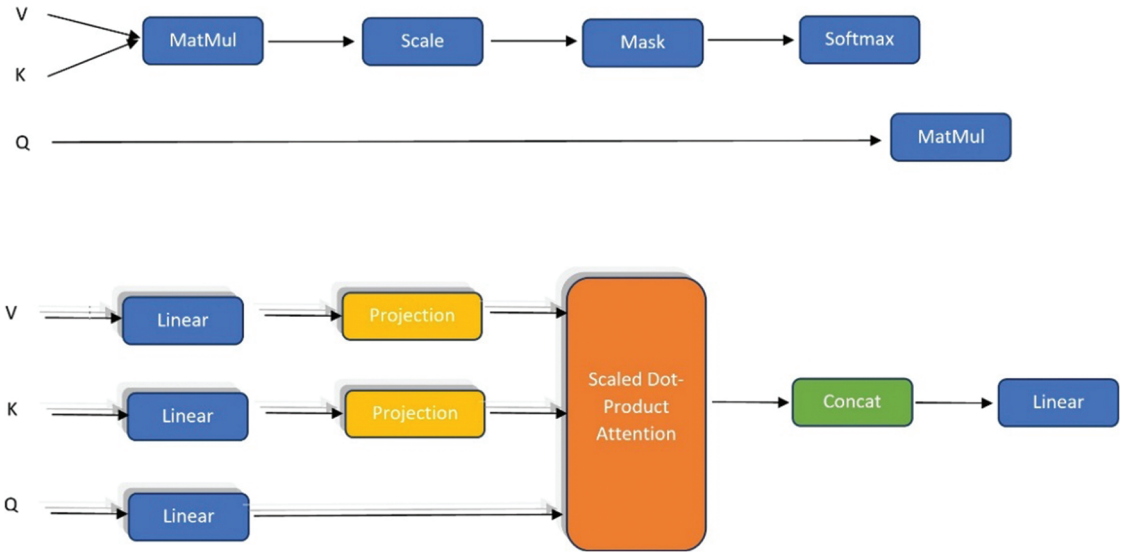

Figure 1: Illustrations of the model multi-head attention mechanism, structure of self-attention, and multi-head attention, respectively

Lastly, context embeddings have been calculated for each head applying $p.\ F_iVW_i^V$. Where d refers to the unit hidden number in the neural network. Multi-head attention uses the self-attention technique, which means $V = K = Q$ in the numeral, see Fig. 1. The advantage of this is that it allows for the calculation of the information of the current location and the information of every other position to determine the dependencies within the entire sequence. As an illustration, if the input is a sentence, then every word in the sentence needs attention calculated with all the other words in the sentence.

For this technique, multi-head attention achieves a linear transformation on the inputs of three vectors of $V,\ K,\ and,\ Q$, before executing calculations. Meanwhile, in the "multi-head attention" technique, the scaled dot-product attention part must be measured frequently. The "heads" number refers to the number of calculations, but the linear projections of $V,\ K,\ and,\ Q$ are varying for each head calculation. Yield the $i-th$ head as an instance:

$$Q' = \text{Q} * W_i^Q, \tag{11}$$

$$K' = \text{K} * W_i^K, \tag{12}$$

$$V' = \text{V} * W_i^V. \tag{13}$$

Meanwhile, this layer receives the BI-GRU output layer, therefore,

$$V = K = Q = y_t. \tag{14}$$

This head final result is calculated:

$$W_i = softmax\left(\frac{Q'K'^T}{\sqrt{d}}\right)V'. \tag{15}$$

The structure of the ACBMA model is shown in Fig. 2. Discharge summaries were divided into words before the model training, see Fig. 2. The word2Vec embedding model was then used to turn discharge summaries into word vectors, with the trained word vectors serving as the model's input. This model initially extracted features from the input word vector using a convolutional network. The output from the extraction of convolutional features was used as BiGRU's input. An attention mechanism module came after the BiGRU and was connected to a fully connected layer after processing pooling through the highest pooling layer. Finally, classification is performed using the sigmoid function. The model is then evaluated using the cross-entropy loss function, and the input sentence classification is determined. Finally, the performance of the proposed model was measured using the metrics of Precision, Recall, and F1 score.

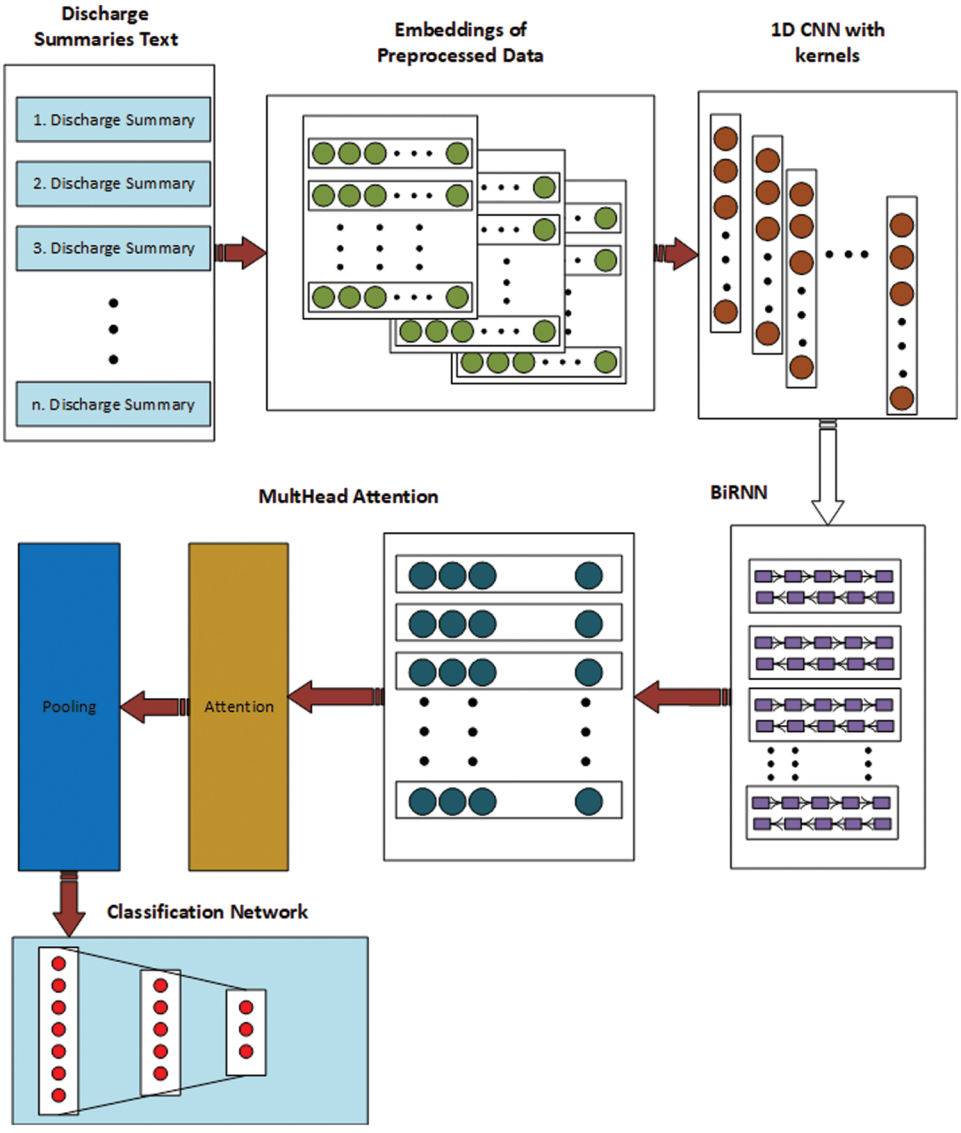

Figure 2: The ACBMA model structure diagram

4 Experiments and Results

4.1 Data Collection

The DS datasets as the objective corpus are examined in this research wherein 1237 de-identified DS, obesity illness, and 15 comorbidities were considered as an illustration in Table 1. This dataset was downloaded from the website of www.i2b2.org/NLP/Obesity/, which was used to determine a correlation among various medical terms. The experiment was performed to predict the treatment quality and health care based on the SA using the DS of the patient.

Table 1: The i2b2 obesity corpus statistics

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

Referring to the parameters in Table 1, the word “Absent” and “Present” correspondingly indicated that every DS provided the data on certain illnesses only and specific infections together with other associated diseases. The words “Questionable” and “Unmentioned” implied that each DS might contain data related to other illnesses and does not refer to data about related diseases, respectively.

4.2 Word Embedding

The DNN (Deep Neural Network) technique uses a constant bag-of-words with other approaches to describe various SA-related issues. Over the decades, the Word2Vec model has routinely been applied to multiple datasets like Google News and Wikipedia having a size of 200. In this study, the main parameters used in the model are the word minimum vocabulary frequency (30), sizes of layers (200), and text windows (5). The values of text window size and minimum vocabulary frequency were selected after conducting the experimentation over a wide range of values for them. The challenge regarding the shifter words was tackled by taking the value of Laplace k equal to 1.0, gain theta of 2.0, and minimum confidence of 0.8 [33].

The word clouds of frequently occurring sentiment terms in medical documents, see Fig. 3. The world or text cloud illustrates the textual data. The texts mining approaches enable highlighting of the texts of high-frequency terms as sentences, paragraphs, or documents, making more visual engagement than the ones represented manually. The sentiment (terms) word cloud of four health documents (Fig. 3) clearly shows the text cloud results. It was observed that most of the arrangements of sentiment terms in the medical documents are connected to the status, improve, stable, failure, etc. In general, these terminologies play a fundamental role in evaluating the treatment quality.

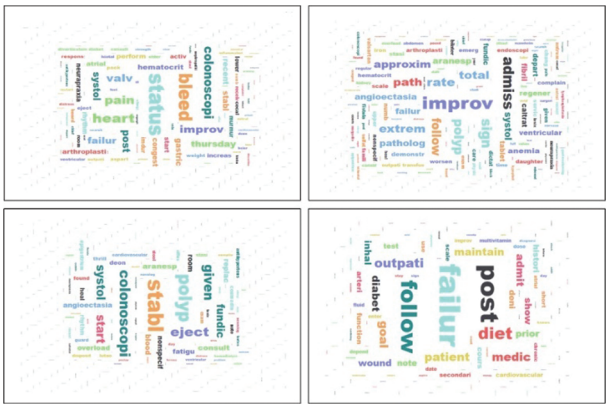

Figure 3: The word clouds frequently occurred in sentiment terms in medical documents

4.3 Lexicon Generation

This section presents one of the novelties of this work. The lexicon limitation is one of the most significant challenges in the medical domain. The novel mechanisms to build and integrate the lexicon-based sentimental scores have been introduced into the learning process of beep learning through an attention mechanism to address SA tasks. This section was separated into two sections, as follows:

i) This section compares the proposed BOW approach with SentWordNet [34] and UMLS [35], VADER [36], and TextBlob lexicon [37], relying on the semantic SA method that suffers from the issue of neglecting a neutral score. This problem is solved by applying the POS (PENN) tagging techniques like (JJ.* |NN.* |RB.* |VB.*) retrieved from (www.cs.nyu.edu/grishman/jet/guide/PennPOS.html). Next, two lists of the terms were generated, wherein BOW is the first, and four lexicons are fused as the second list that relied on the hypernym's procedure.

ii) In the second section, the sentiment-specific words embeddings models proposed to learn the sentimental orientation of features in the exiting language models, like GloVe, Word2Vec, FastText, BERT, and TF-IDF, from the global context in a specific domain. Inspired by this intuition, this paper introduces a weak supervised solution to build a domain-specific sentiment lexicon. Specifically, this paper proposes leveraging a tiny seed of sentiment words with the feature distribution in the embedding space of a specific domain to associate each word with a domain-specific sentiment score. The key idea is to learn a set of cluster embeddings used to build the lexicon by looking at their neighbors in the latent space. To achieve this, introduce an unsupervised neural network trained to minimize the error reconstruction, i.e., analogous to autoencoder, of a given input as a linear combination from the cluster's matrix. The model does not require labeled data for training purposes, so it enables constructing a sentiment lexicon in the low-resource language. Finally, the obtained results were applied to assign the training dataset relying on the medical documents. The input to the model is a list of sentence indexes in the vocabulary, which is modeled by simply averaging its corresponding features' vectors. The

modeled input dimension is reduced to k clusters to compute the relatedness probability to each cluster. The model is trained to approximate the modeled input as a linear combination of cluster embeddings from C. An example of the proposed model is shown in Fig. 4. The sentiment polarity was estimated via the following Equation:

$$\text{Polarity of text sentiment } = (\text{neg} - \text{pos})/(\text{pos} + \text{neg} + \text{neu}) \tag{16}$$

$$m_t = \frac{1}{n}\sum_{i=1}^{n} l_w^i, \tag{17}$$

where l_w^i denotes the feature vector of the word w_i. Hence, m_t denotes the sentence representation that belief captures the global sentiment representation of the input sentence.

$$r_t = C^T . v_{t,} \tag{18}$$

where r_t denotes the reconstructed vector, and v_t is a weighted vector over k clusters. Particularly, v_t can be read as the probability that the input belongs to the clusters. It is computed by reducing the sentence representation m_t from d dimension to k dimension and then apply a *softmax* nonlinearity to yield nonnegative weights:

$$v_t = softmax(W . m_t + b), \tag{19}$$

where W is the projection parameter and b is the bias, which is learned during the training process.

$$C \in R^{k \times d}. \tag{20}$$

where C denotes the cluster matrix.

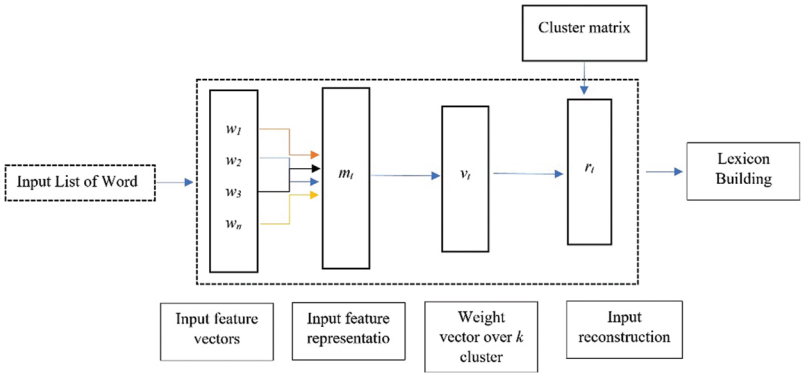

Figure 4: The steps of the proposed solution

In the next stage, the numerical representation for each sentence was obtained. The processed dataset was submitted words-wise to an embedding pre-trained model Word2Vec. After that, SentiWordNet, TextBlob, VADER, UMLS, and statistical techniques were used to develop a specialized vocabulary (medical domain) that determined the polarity of each sentence [38]. Various sizes of the lexicon (number of terms) were also examined to test this method's reliability [39]. The lexicon's capacities were also evaluated with different sizes of lexicons to investigate whether the largest or smallest lexicon of sentiment can produce better results [40]. The most extensive lexicon produced the highest performance, and the smallest lexicon resulted poorly. The obtained results rely on the number of words in each

lexicon, indicating that the values of lexicons 1, 2, 3, 4, 5, 6, and 7 are 10.000, 20.000, 30.000, 40.000, 50.000, 60.000, and 70.000, respectively, see Fig. 5.

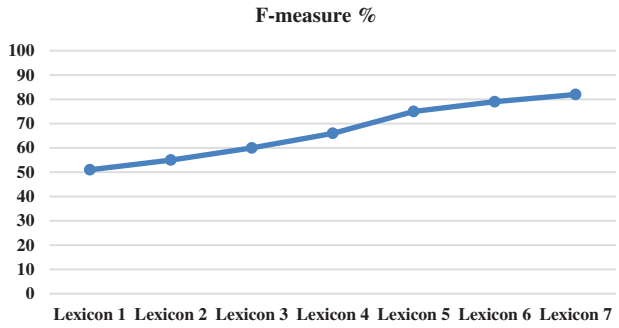

Figure 5: The size-dependent performance of each sentiment lexicon

4.4 Active Learning Method

The purpose of the AL method is to reduce the problems and costs related to the manual annotation step in supervised and semi-supervised machine learning approaches [41,42]. Reduction of the manual annotation burden becomes exceptionally critical in the medical domain because of qualified experts' high costs for annotating medical documents. This technique is applied for different biomedical tasks, for example, text classification, medical named entity recognition, and de-identifying medical records. The common AL technique is to choose the samples randomly. AL methods iteratively use ML approaches, and a human annotator can drastically decrease by involving in the learning process.

The AL general cycle to extract information from the document, see Fig. 6. The query strategy was used to select the informative samples from unstructured medical text documents as an iterative cycle. A human annotator does the labels, and these samples are used for extracting data and building an ML-based model at every recurrent cycle. This technique has not been fully explored for biomedical information extraction [43]. The key idea of AL is to test the effectiveness of the suggested model by decreasing the number of samples that need manual labeling. The major problem is identifying the practical examples available to train a model, producing better effectiveness and performance [44].

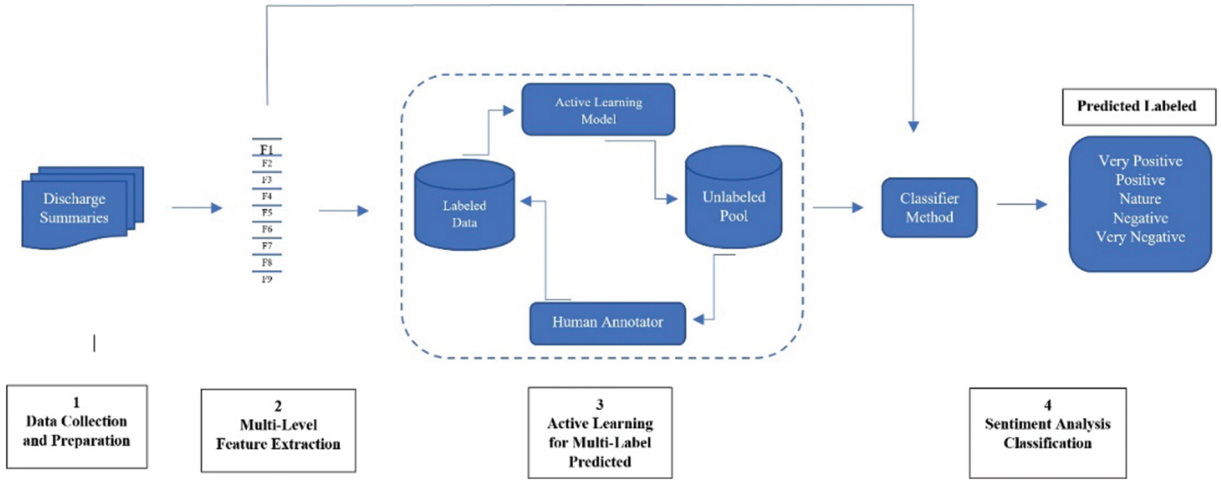

Figure 6: An overview of the ML when applied in AL

To accomplish the abovementioned objective, the authors collected the opinions and views (through questionnaires) from various teachers, lecturers, linguistics (English language) doctoral students, and annotators capable of teaching, understanding, and reading English. The dataset was supplied based on AL and human annotators, and they were asked to label each sentence using one of five terms: very positive, positive, neutral, negative, and very negative. The final label or polarity of the sentences was decided based on the annotators' majority vote, wherein the dataset was labeled sentence-wise. Two illnesses were chosen randomly from a list of sixteen diseases, asthma with 606 and obesity with 606 DSs. The annotator aimed to create a gold standard to train data labeled at the sentence level containing 5439 sentences. The sentiment tag for each discharge summary based on the sentences was assigned with the polarity of +1, 0.5, 1, 0, −0.5, and −1 corresponding to each positive, positive, neutral, negative, and very negative, see Table 2. These datasets were utilized to assess the final results. For more details and analyses of the used dataset, see Table 3.

Table 2: The gold standard corpus statistics

Diseases	Very positive	Positive	Neutral	Negative	Very negative	Total
Obesity	141	215	180	41	29	606
Asthma	213	150	211	21	11	606

Table 3: The statistics of the datasets are applied in our experiments

Dataset	Statistics	Training	Validation	Testing
Obesity	No. sentence	2,155	510	990
	No. token	31,120	7,240	10,560
	Aspect labels (%)	10.93	9.99	13.20
	Opinion labels (%)	9.02	7.29	8.90
Asthma	No. sentence	1,160	220	870
	No. token	30,90	8,210	9,620
	Aspect labels (%)	9.89	10.01	12.41
	Opinion labels (%)	9.15	7.40	8.97

4.5 Different Numbers of Convolution Kernels Experiment

Convolution kernels of various sizes and counts were used in the training model to examine the model. For the test results, see Table 3. The model was analyzed using a size of the convolution kernel two and a step size of 2, 4, and 6. The effectiveness of model training and the precision of the experimental outcomes have been impacted by the different sizes and numbers of convolution kernels. To experimentally test the resultis of the suggested model's results, eight convolution kernels of varied sizes and amounts were used. Table 4 illustrates the test results' F1 score, accuracy, precision, and recall. The recall, precision, accuracy, and F1 score of the suggested model were slightly higher when five convolution kernels (2, 4, 6, 8, and 10) were utilized than when other convolution kernel combination approaches were used, as can be shown from the above table and the experimental data. In the empirical procedure analysis of these eight different convolution kernel types, see Fig. 7. The variance in the correctness of the verification set after training is displayed. The experimental results of the convolution kernel allocation technique are shown

in the seven figures. These results primarily show how the set of training changes with the iteration number and how the validation set accuracy changes as the iteration number rises.

Table 4: Experimental outcomes for various conditions of convolution kernels

Kernels No.	Recall	Precision	Accuracy	F1 score
2, 4	0.9893	0.9716	0.9856	0.9753
2, 4, 6	0.9845	0.9724	0.9853	0.9751
2, 4, 6, 8	0.9880	0.9718	0.9851	0.9747
2, 4, 6, 8, 10	0.9894	0.9758	0.9861	0.9757
2, 4, 6, 8, 10, 12	0.9869	0.9717	0.9845	0.9741
2, 4, 6, 8, 10, 12, 14	0.9891	0.9705	0.9850	0.9746
2, 4, 6, 8, 10, 12, 14, 16	0.9881	0.9712	0.9853	0.9742

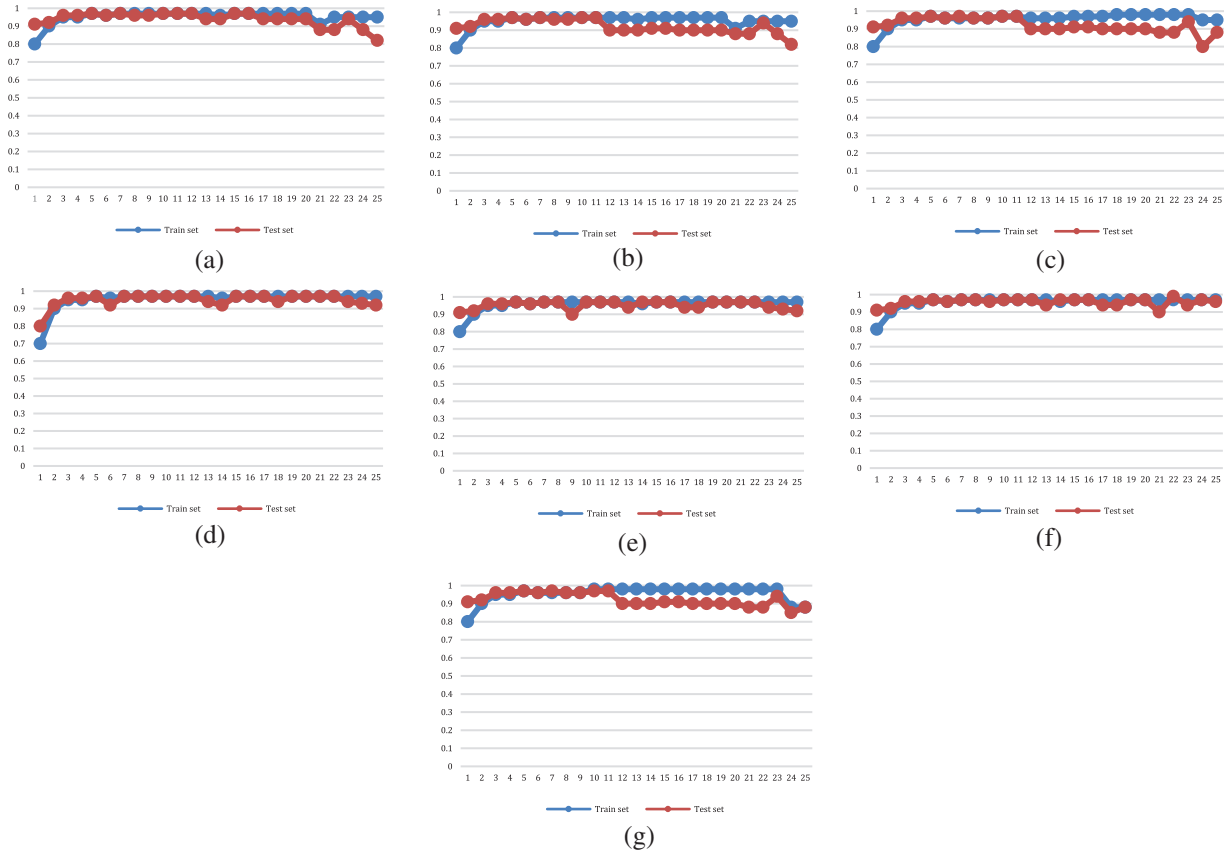

Figure 7: Illustrates the experiment results with different convolution kernels, the x-axis presents the accuracy, and the y-axis shows the iterations of the training and validation set differences with the iterations numeral. (a) [2, 4], (b) [2, 4, 6], (c) [2, 4, 6, 8], (d) [2, 4, 6, 8, 10], (e) [2, 4, 6, 8, 10, 12], (f) [2, 4, 6, 8, 10, 12, 14], and (g) [2, 4, 6, 8, 10, 12, 14, 16] convolution kernel

4.6 Different Numbers of Layers BiGRU Experiments

The learning rate controls whether and when the objective function converges to the local minimum, making it a significant hyperparameter in supervised learning and deep learning techniques. The objective function can combine with the local minimum in an acceptable amount of time with the correct learning rate. The convergence will happen too slowly if the learning rate is too slow. The cost function will oscillate if the learning rate is too high. This paper has utilized the 10-fold cross-validation method in our advanced solution. To make this comparison unbiased, this paper is applied 10-fold cross-validations with six baseline approaches and eight state-of-the-art deep learning methods. Multiple learning rates were tested in trials to determine the best learning rate. Traditional machine learning methods like Naive Bayes, logistic regression, KNN, SVM, decision trees, and random forest will be examined using a variety of approaches. The experiment's findings also included the F1 score, recall, precision, and accuracy. The examination results are displayed in Table 5. The deep learning methods have also been investigated at the same time. To validate the performance of the proposed solution, the research compared it with state-of-the-art techniques, including the convolution and GRU multi-head attention model [19], the BiGRU-Word2Vec and multi-head attention mechanism model [20], and the mixed methods of GRU and multi-head attention [45].

Table 5: Comparison of examined results with different traditional machine learning approaches

Methods	Recall	Precision	Accuracy	F1 score
SVM	0.7681	0.8292	0.8050	0.7975
KNN	0.7997	0.8254	0.8153	0.8124
DT	0.8488	0.8439	0.8460	0.8464
LR	0.8747	0.8642	0.8686	0.8694
RF	0.8864	0.8685	0.8761	0.8774
NB	0.8983	0.8822	0.8892	0.8902
AL-CNN-BiGRU-Word2Vec multi-head attention	0.9708	0.9103	0.9867	0.9853

Standard LSTM is the essential component of our proposed solution. To demonstrate the improvement of the proposed attention mechanism, this paper evaluates the performance of our proposed method by comparing it with the state-of-the-art LSTM methods that consider sentiment tasks as a textual classification issue. ATAE-LSTM [21] introduced an attention-based LSTM to model the relation between the aspect and its context by learning the aspect matrix. However, the aspect category is often implicitly expressed in the text and thus discourages learning aspect representation accurately. SenticLSTM [22], an attentive biLSTM, was presented leveraging the external information as commonsense knowledge from SenticNet to address ABSA. Lexicon-based ML [23], an ML-based solution for ATSA, proposed modeling the joining between sentences and leveraging the explicit features. ATAE-LSTM [24] presented an attention-based LSTM for learning aspect representation that encourages LSTM to predict the polarity in response to an aspect. The experimental findings are based on F1 score, recall, precision, and accuracy for Obesity and Asthma datasets, see Table 6.

Table 6: Examine outcomes compared using various state-of-the-art deep learning techniques

Approaches	Obesity		Asthma	
	Accuracy	F1 score	Accuracy	F1 score
Standard LSTM	85.26	90.64	81.20	88.30
ATAE-LSTM	88.11	90.49	84.12	87.65
SenticLSTM	89.25	91.82	85.34	88.28
Lexicon-based GML	90.86	92.07	88.76	89.10
ATAE-LSTM	91.37	93.78	88.37	89.10
GRU-multi-attention	92.37	95.36	94.38	95.80
BiGRU-multi-attention	94.49	96.47	95.89	96.41
CNN-GRU-multi-attention	95.51	94.49	95.11	95.23
AL-CNN-BiGRU-Word2Vec multi-head attention (our)	97.87	97.81	96.54	96.18

5 Discussion

The ACBMA algorithm method that is composite active learning, BiGRU, and CNN networks and presents a multi-head attention technique was suggested and relied on the relevant characteristics of CNN, bidirectional LSTM networks, and multi-head attention technique, which has been utilized in the sentiment analysis domain of discharge summaries. The discharge summaries dataset (medical field) was used for sentiment detection. This paper presents an unsupervised neural network model to learn a set of clusters embedded in the latent space. The model is trained in an unsupervised manner, i.e., analogous to autoencoders, and thus facilitates constructing a sentiment lexicon in low-resources languages. The CNN advantages to extracting local context features and the BiGRU to extracting global context features were fully taken into account in this method, along with the information in the text's context, and the context features were efficiently extracted. This study has performed analytical experimentation at every stage, including comparing convolutions and BiGRU. Additionally, several conventional machine learning techniques were evaluated. According to extensive experiments, the ACBMA model has an enhanced impact on the medical sector of discharge summaries. The comparison is done based on six-well know methods, can be observed the significant improvements (±10%), see Fig. 5. The other comparison is also done based on eight state-of-the-art procedures, which can be followed the significant improvements (±2%), see Fig. 6. Furthermore, the study discussed in this paper may be helpful in the area of sentiment analysis for discharge summaries. Our results are significantly based on utilizing multiple examines, handling missing values, removing/adding other variables from the proposed model, trying various statistical exams, and grouping similar and numeric variables. The final results illustrate that the ACBMA method improves the F1 score of 97.81 and 96.18 based on the accurate benchmarks of Obesity and Asthma, respectively. The experimental results demonstrate the effectiveness of the proposed ACBMA model.

6 Conclusions

Sentiment detection is of great value to the natural language process, which analyzes text to determine the expressed sentiment. The state-of-the-art techniques are built upon various deep neural network models. Despite the effectiveness of these approaches, the performance heavily depends on the quality of the massive labeled corpus, which is labor-intensive and readily available in real scenarios, so this paper is suggested a new unsupervised method to generate the specific lexicon. This paper proposes the ACBMA algorithm

approach, a fusion of active learning, BiGRU, and CNN networks, and presents a multi-head attention mechanism that was suggested and is utilized in the sentiment analysis domain of discharge summaries. This model relies on the characteristics of CNN, bidirectional LSTM networks, and multi-head attention techniques. In this method, the information in the text context and the CNN advantages to extract local context features and BiGRU to extract universal characteristics of the context were thoroughly considered. Effective text feature extraction was achieved. Each step of this investigation involved experimental analysis, including various convolutions and BiGRU. Different conventional machine learning techniques were also evaluated. Numerous trials showed that the ACBMA algorithm approach better impacts the medical domain of discharge summaries. Additionally, the study discussed in this essay may be helpful in discharge summary sentiment analysis.

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

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

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

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] H. Jelodar, "A NLP framework based on meaningful latent-topic detection and sentiment analysis via fuzzy lattice reasoning on youtube comments," *Multimedia Tools and Applications*, vol. 80, no. 3, pp. 4155–4181, 2021.

[2] S. Mukherjee, "Sentiment analysis," *ML. NET Revealed*, vol. 1, pp. 113–127, 2021.

[3] Z. Nabi, R. Talib, M. K. Hanif and M. J. C. S. S. E. Awais, "Contextual text mining framework for unstructured textual judicial corpora through ontologies," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 1357–1374, 2022.

[4] A. Alamoodi, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," *Expert Systems with Applications*, vol. 167, pp. 114155, 2021.

[5] A. E. de Oliveira Carosia, G. P. Coelho and A. E. A. J. E. S. w. A. da Silva, "Investment strategies applied to the Brazilian stock market: A methodology based on sentiment analysis with deep learning," *Expert Systems with Applications*, vol. 184, pp. 115470, 2021.

[6] L. Nemes and A. J. J. o. I. Kiss, "Social media sentiment analysis based on COVID-19," *Journal of Information and Telecommunication*, vol. 5, no. 1, pp. 1–15, 2021.

[7] P. Madan, "An optimization-based diabetes prediction model using CNN and Bi-directional LSTM in real-time environment," *Applied Sciences*, vol. 12, no. 8, pp. 3989, 2022.

[8] C. Liu, Y. Zhang, J. Sun, Z. Cui and K. J. I. J. o. E. R. Wang, "Stacked bidirectional LSTM RNN to evaluate the remaining useful life of supercapacitor," *International Journal of Energy Research*, vol. 46, no. 3, pp. 3034–3043, 2022.

[9] J. C. Pereira-Kohatsu, L. Quijano-Sánchez, F. Liberatore and M. J. S. Camacho-Collados, "Detecting and monitoring hate speech in twitter," *Sensors*, vol. 19, no. 21, pp. 4654, 2019.

[10] S. A. Waheeb, N. A. Khan and X. J. M. J. o. C. S. Shang, "An efficient sentiment analysis based deep learning classification model to evaluate treatment quality," *Malaysian Journal of Computer Science*, vol. 35, no. 1, pp. 1–20, 2022.

[11] N. Hong, "Developing a FHIR-based EHR phenotyping framework: A case study for identification of patients with obesity and multiple comorbidities from discharge summaries," *Journal of Biomedical Informatics*, vol. 99, pp. 103310, 2019.

[12] S. A. Waheeb, N. Ahmed Khan, B. Chen and X. Shang, "Machine learning based sentiment text classification for evaluating treatment quality of discharge summary," *Information*, vol. 11, no. 5, pp. 281, 2020.

[13] L. Rognone, S. Hyde and S. S. J. I. R. o. F. A. Zhang, "News sentiment in the cryptocurrency market: An empirical comparison with forex," *International Review of Financial Analysis*, vol. 69, pp. 101462, 2020.

[14] R. -P. Shen, H. -R. Zhang, H. Yu and F. J. E. S. w. A. Min, "Sentiment based matrix factorization with reliability for recommendation," *Expert Systems with Applications*, vol. 135, pp. 249–258, 2019.

[15] F. Falck, "Measuring proximity between newspapers and political parties: The sentiment political compass," *Policy & Internet*, vol. 12, no. 3, pp. 367–399, 2020.

[16] V. M. Joshi, R. B. Ghongade, A. M. Joshi and R. V. J. B. S. P. Kulkarni, "Deep BiLSTM neural network model for emotion detection using cross-dataset approach," *Biomedical Signal Processing and Control*, vol. 73, pp. 103407, 2022.

[17] H. Xiao, J. Qin, S. Jeon, B. Yan and X. J. M. Yang, "MFEN: Lightweight multi-scale feature extraction super-resolution network in embedded system," *Microprocessors and Microsystems*, vol. 35, pp. 104568, 2022.

[18] Y. Eum and E. -H. J. C. Yoo, "Imputation of missing time-activity data with long-term gaps: A multi-scale residual CNN-LSTM network model," *Computers, Environment and Urban Systems*, vol. 95, pp. 101823, 2022.

[19] Y. Cheng, "Sentiment analysis using multi-head attention capsules with multi-channel CNN and bidirectional GRU," *IEEE Access*, vol. 9, pp. 60383–60395, 2021.

[20] C. Chu, Y. Ge, Q. Qian, B. Hua and J. J. D. S. P. Guo, "A novel multi-scale convolution model based on multi-dilation rates and multi-attention mechanism for mechanical fault diagnosis," *Digital Signal Processing*, vol. 122, pp. 103355, 2022.

[21] W. Liao, J. Zhou, Y. Wang, Y. Yin, X. J. A. I. R. Zhang *et al.,* "Fine-grained attention-based phrase-aware network for aspect-level sentiment analysis," *Artificial Intelligence Review*, vol. 55, no. 5, pp. 3727–3746, 2022.

[22] B. Liang, H. Su, L. Gui, E. Cambria and R. J. K. -B. S. Xu, "Aspect-based sentiment analysis via affective knowledge enhanced graph convolutional networks," *Knowledge-Based Systems*, vol. 235, pp. 107643, 2022.

[23] Y. Wang, Q. Chen, J. Shen, B. Hou, M. Ahmed *et al.*, "Aspect-level sentiment analysis based on gradual machine learning," *Knowledge-Based Systems*, vol. 212, pp. 106509, 2021.

[24] H. Yan, B. Yi, H. Li and D. J. N. C. Wu, "Sentiment knowledge-induced neural network for aspect-level sentiment analysis," *Neural Computing and Applications*, vol. 67, pp. 1–12, 2022.

[25] S. Abdulateef, N. A. Khan, B. Chen and X. J. I. Shang, "Multidocument arabic text summarization based on clustering and word2vec to reduce redundancy," *Information*, vol. 11, no. 2, pp. 59, 2020.

[26] S. A. Waheeb, N. Ahmed Khan and X. Shang, "An efficient sentiment analysis based deep learning classification model to evaluate treatment quality," *Malaysian Journal of Computer Science*, vol. 35, no. 1, pp. 1–20, 2022.

[27] S. A. Waheeb, N. Ahmed Khan and X. Shang, "Topic modeling and sentiment analysis of online education in the COVID-19 era using social networks based datasets," *Electronics*, vol. 11, no. 5, pp. 715. 2022.

[28] E. Kavitha, R. Tamilarasan, A. Baladhandapani and M. J. J. C. S. S. E. Kannan, "A novel doft clustering approach for gene expression data," *Computer Systems Science and Engineering*, vol. 43, no. 3, pp. 871–886, 2022.

[29] B. Yin, R. Zuo and Y. J. N. R. R. Xiong, "Mineral prospectivity mapping via gated recurrent unit model," *Natural Resources Research*, vol. 31, no. 4, pp. 2065–2079, 2022.

[30] M. Yeung, E. Sala, C. -B. Schönlieb and L. J. C. M. I. Rundo, "Unified focal loss: Generalising dice and cross entropy-based losses to handle class imbalanced medical image segmentation," *Computerized Medical Imaging*, vol. 95, pp. 102026, 2022.

[31] J. Li, H. Sun and J. J. M. L. Li, "Beyond confusion matrix: Learning from multiple annotators with awareness of instance features," *Machine Learning*, vol. 91, pp. 1–23, 2022.

[32] A. Theissler, M. Thomas, M. Burch and F. J. K. -B. S. Gerschner, "ConfusionVis: Comparative evaluation and selection of multi-class classifiers based on confusion matrices," *Knowledge-Based Systems*, vol. 247, pp. 108651, 2022.

[33] F. Zhang, "A hybrid structured deep neural network with word2vec for construction accident causes classification," *International Journal of Construction Management*, vol. 22, no. 6, pp. 1120–1140, 2022.

[34] B. Mounica and K. J. I. J. o. S. A. E. Lavanya, "Feature selection method on twitter dataset with part-of-speech (PoS) pattern applied to traffic analysis," *International Journal of System Assurance Engineering and Management*, vol. 12, pp. 1–14, 2022.

[35] A. K. Chanda, T. Bai, Z. Yang and S. J. B. M. I. Vucetic, "Improving medical term embeddings using UMLS metathesaurus," *BMC Medical Informatics and Decision Making*, vol. 22, no. 1, pp. 1–12, 2022.

[36] J. S. Tan and W. C. Chia, "Research output to industry Use: A readiness study for topic modelling with sentiment analysis," in *Proc. of the 8th Int. Conf. on Computational Science and Technology*, Labuan, Malaysia, pp. 13–25, 2022.

[37] I. Gupta, T. K. Madan, S. Singh and A. K. J. a. p. a. Singh, "HiSA-SMFM: Historical and sentiment analysis based stock market forecasting model," arXiv preprint arXiv:2203.08143, 2022.

[38] S. A. Waheeb, N. A. Khan, B. Chen and X. J. I. Shang, "Machine learning based sentiment text classification for evaluating treatment quality of discharge summary," *Information*, vol. 11, no. 5, pp. 281, 2020.

[39] J. Zaghir, J. F. Rodrigues-Jr, L. Goeuriot and S. J. J. o. H. I. R. Amer-Yahia, "Real-world patient trajectory prediction from clinical notes using artificial neural networks and UMLS-based extraction of concepts," *Journal of Healthcare Informatics Research*, vol. 11, pp. 1–23, 2021.

[40] F. Bravo-Marquez, A. Khanchandani and B. J. C. C. Pfahringer, "Incremental word vectors for time-evolving sentiment lexicon induction," *Cognitive Computation*, vol. 14, pp. 1–17, 2021.

[41] I. M. El-Hasnony, O. M. Elzeki, A. Alshehri and H. J. S. Salem, "Multi-label active learning-based machine learning model for heart disease prediction," *Sensors*, vol. 22, no. 3, pp. 1184, 2022.

[42] Q. Zhuang, Z. Dai and J. J. C. I. Wu, "Deep active learning framework for lymph node metastasis prediction in medical support system," *Computational Intelligence and Neuroscience*, vol. 2022, pp. 13, 2022.

[43] N. S. Khan and M. S. J. W. P. C. Ghani, "A survey of deep learning based models for human activity recognition," *Wireless Personal Communications*, vol. 120, no. 2, pp. 1593–1635, 2021.

[44] X. Peng, X. Jin, S. Duan and C. J. I. T. Sankavaram, "Active learning assisted semi-supervised learning for fault detection and diagnostics with imbalanced dataset," *IISE Transactions*, vol. 55, pp. 1–29, 2022.

[45] Y. Wu and W. J. A. I. Li, "Aspect-level sentiment classification based on location and hybrid multi attention mechanism," *Applied Intelligence*, vol. 52, pp. 1–16, 2022.