AN EFFICIENT SENTIMENT ANALYSIS BASED DEEP LEARNING CLASSIFICATION MODEL TO EVALUATE TREATMENT QUALITY

Samer Abdulateef Waheeb[1], Naseer Ahmed Khan[2], Xuequn Shang[3*]

[1,2,3]School of Computer Science and Engineering, Northwestern Polytechnical University, Xi'an 710072, China

Email: samirabdulateef@mail.nwpu.edu.cn[1], naseerkhan@mail.nwpu.edu.cn[2], shang@nwpu.edu.cn[3*] (corresponding author)

DOI: https://doi.org/10.22452/mjcs.vol35no1.1

ABSTRACT

Extracting information using an automated system from unstructured medical documents related to patients discharge summaries in the health care centers is considered a big challenge. Sentiment analysis of medical records has gained significant attention worldwide to understand the behaviors of both clinicians and patients. However, Sentiment analysis of discharge summary still does not provide a clear picture of the information available in these summaries. This study proposes a machine learning-based novel sentiment analysis unsupervised techniques to classify discharge summaries using TF-IDF, Word2Vec, GloVe, FastText, and BERT as deep learning approaches with statistical methods, and clustering. Our proposed model is an unsupervised sentiment framework that provides good understanding and insights of the clinical features that are not captured in the electronic health data records. Moreover, it's a hybrid sentiment model consisting of clustering technique and vector space models for selecting the distinctive terms. The main intensity of measured sentiment is captured using the polarity of positive and negative terms in the discharge summary. The combination of SentiWordNet platform and our approach is used to build a lexicon sentiment dataset (assignment polarity). Experiments shows that our suggested method achieves 93% accuracy and significantly outperforms other state of the art approaches based on the inspiration of sentiment analysis technique to examine the treatment quality for discharge summaries.

Keywords: Clustering, Discharge summary, Sentiment analysis, SentiWordNet, Quality evaluation, Health surveillance

1.0 INTRODUCTION

Discharge summary (DS) is a collection of medical documents that are maintained on the patients' health data in the health care centers, written by the medical centers' personal in an unstructured fashion. Currently, the investigations on using a huge data set of health electronic data records have raised challenging questions extending from risk stratification to pharmacovigilance. However, electronic health records in terms of coded data bring a limitation schema for the clinical status and blurred decision-making [1]. Besides, most of the health records data are not related to sentiment analysis when their discharge summary got examined. Generally, health records data provides unstructured data, which means this data is usually captured in providing analysis as description notes. Besides, this data is rich in information and in providing decision making and analysis [2]. However, it is hard to quantify this type of data, which reduces the required efficiency for the health-systems. Recently, some researchers started seeking text reflections on specific health topics based on "feelings" that the author has. For example, the word 'excellent' is generally reflects a positive meaning, while the word 'worse' reflects a negative meaning. The main objective of text classification using sentiment analysis is to help classify DS related documents to neutral (normal condition), negative (abnormal condition), and positive (improved health condition) documents, which helps improve and evaluate treatment quality [3]. Hence, several tools and techniques are studied and validated to enable the quantization algorithm of this "feeling" impact, specifically in text documents, it is referred as a sentiment. This process has allowed more studies and investigations on the topics related to the health care applications such as satisfaction of the health care, impact of tobacco on health, discussions on the impact of cancer treatment, and health happiness on Twitter [4]. Evaluation of the quality of discharge summary is also crucial since based on quality of discharge summary documents treatment quality, health care standard and better management of the health care can be analyzed. Evaluation of DS based on the manual observation of the health care professionals for the acute coronary syndrome were discussed in [5]. In [6] a training program was launched to improve the quality of the discharge summary of the health care centers. So in the light of these discussions a less costly, more reliable and automated evaluation technique is required to assess the quality of the discharge summary in the health care centers.

Recently, the information obtained from discharge summary has dramatically increased and it is now considered as a significant data source in the development of several applications and domains due to peculiar usage in the medical fields. To have a full and good understanding of patients, medical reviews or opinions become one of the major areas of interest [7]. The data from medical institutions including hospitals and clinics has been broadly used to fulfill the demands of giving efficient reviews about the behavior of the patient after discharge stage. Several researchers have employed various discharge summaries to distinguish and anticipate possible patients who proceed further to retrieve their information from discharge summary [8]. Due to complexity and a huge number of these datasets, the research has now turned towards the data mining methods. The data mining methods have effectively modeled several techniques, algorithms and tools to handle with huge patterns ofdata, and increase insightful learning. In addition, analyzing patient's behaviors using sentiment analysis methods from discharge notes has a vital role in giving a reliable opinion for the physicians and clinicians, this is conducted by analyzing information retrieved from discharge summary notes [9]. In that matter, finding the polarity of clinician's reviews on a specific patient's discharge summary will utilize the sentiment analysis of these data.

Sentiment Analysis is nowadays getting more attention in the field of Natural Language Processing (NLP) due to its applications and significance in the health care setups, administration, decision process, and quality management. These administrative decisions and policies not only affect in improving the management and quality of the health care centers but also helps accurate diagnosis and treatment of the patients that visit those health care centers during their lifetime [10]. A human brain cannot extract and process a huge medical unstructured written data by the medical staff in order to improve the management and quality of the health care centers, which thereby necessitates the use of Machine Learning (ML) based automated systems that help in extracting and analyzing vast amount of data and presented information in a concise and summarized way to the health care experts which are at the helm of the crucial affairs related to those medical setups [11].

Recently, unsupervised techniques are on the rise in the literature to reduce the dimensionality of the dataset and due to the reason that labelled data is sometimes not available or is difficult to create. IF-IDF [12], Word2Vec [10], GloVe [11], FastText [13], and BERT [14] are unsupervised techniques, unsupervised in a sense that they are pre-trained vectors that the user does need to provide the labels of the dataset to get the "embedding" vectors, that are now being used more often in the area of Natural language processing to give the semantic meaning to words by converting words in the language to the fixed length vectors called "embedding" that are semantically more meaningful and useful. The ordinary Bag of Words (BoW) [15] model gets sometimes too long to handle and leads to bias, therefore vector space models help in reducing the dimension of the word vectors which can then be used as features leading to less biasness.

Our study hypothesized that sentiment analysis from discharge summary that has been written by doctors and clinicians can be measured as description notes, which helps understand the patient characteristics [16]. To be specific, if these discharge summaries are able to have predictive validity, it would be a great opportunity for utilizing the description notes to be used as an augmented data coded source. Hence, this study has applied a sentiment algorithm called "opinion mining" for quantifying the sentiment in a quantity of description hospital and clinical discharge summary [17]. The contributions of our work are summarized as follows:

- Pre-trained models are adopted, based on plaintext after (preprocessing or simplify text or normalization). The chosen pre-trained models for two main reasons: 1) these models are bypassing the time-consuming, available in different languages, pre-trained on a huge dataset from scratch, therefore, need focused on the right tuning only; 2) working on plaintext comes with improved performance.
- Unsupervised models Adopted IF-IDF, Word2Vec, GloVe, FastText, BERT, and Clustering techniques with sentiment analysis as unsupervised techniques for evaluating the quality discharge summary.
- Establishing a gold sentiment standard in a dataset (lexicon) of discharge summary by using SentiWordNet and statistical techniques.
- This research applied an idea that is inspired by deep learning for realizing text classification with the discharge summary to increase the model efficiency.

2.0 RELATED WORKS

2.1 Sentiment Analysis Methods

One of the challenging areas in the domain of NLP is sentiment analysis, which is classifying the orientation of a meaningful sentence into positive, negative, and neutral categories [18]. The methods that are built to analyze the

NLP data for the sentiment analysis can be divided into three main areas, which are techniques based on, lexicon, machine learning, and the combination of these two which are called hybrid methods [19].
A study [20] evaluated the treatment and diagnosis of the disease anorexia nervosa using the 52 healthcare related documents of 52 healthy patients and 15 control subjects. They a rule-based model that combined the famous lexicon analysis and Bag-of-Words technique tocategorize of each document to the predefined classes.

In the study [21], authors first compiled the data on patients' opinions regarding the 103 medicines from the various social medial platforms, which resulted in 22 million rows. After that, a deep learning model based on word-embeddings using 10-fold accuracy of used for classification using classifiers such as SVM, KNN, and logistic regression. They reported a maximum F1 measure of 0.645.

Of late, more and more studies are being reported using the unsupervised techniques of feature selection approach. F1 measure value of 0.787 using the Support Vector Machine, 0.78 using the Naïve Bayes, 0.77 using the Long Short-Term Memory (LSTM), and 0.87 using the decision trees were reported in the studies [22-25] respectively. However, sentiment analysis of discharge summary has been analyzed by dictionaries and lexical resources including popular SentiWordNet [11], MPQA [26], and WordNet effect [27]. These platforms are working as reliable tools for identifying sentiment in highly opinionated text. However, for clinical discharge summary notes, these platforms are not used explicitly [28]. Besides that, one of the major challenges faced by sentiment analysis of clinical records is the sentiment annotation high cost of huge medical records. Additionally, using a supervised learning method produces incorrect or impracticable settings due to the need of accessing labels into training data. Hence, using unsupervised learning method to develop the most suitable datasets in sentiment analysis of discharge summary.

2.2 Discharge Summary Quality Evaluation

DS is the product of health care centers which consists of a variety of documents, such as, patient's diagnosis, doctor's prescription, nurse written patients symptoms during the emergency pickup, electronic reports generated by the medical test, and the reviews about the drug by the health care centers in process of treating patients [29]. Key factors in the course of testing the quality of discharge summary such as poorly written nurses first hand reports of the patients, missing medical records in case of patient's abundance in the hospital, poorly written medical examination of the junior doctor's due to lack of proper training and collapse of medical record data in case of software failure were reported in the studies conducted by [30, 31].

In the study [32] conducted to test the quality of diagnosis, a comparison-based approach was employed between the reference list prepared by the health care clinicians and the diagnosis list obtained from the patients. Five key factors were considered in preparing and comparing against the reference list which was, missing diagnosis information, an incorrect diagnosis that was later rectified, a severe imprecise diagnosis that led to other complications, a partial diagnosis that required more follow up medical test, and the accurate diagnosis that led to the patient's gradual recovery. The study proposed in [33] used machine learning model based on text processing tools, word embeddings, and lexicon construction to come up with useful features for the training of the model. A cross-validation based F1 value of 0.885 was reported based on their novel features.

Doc2Vec, Word2Vec were applied in [34] to for the sentiment analysis task related to medical records . The authors of the study also used the Welsh statistic of the WordNet for the evaluation of unsupervised models for the medical domain. In the study of [35] a Bidirectional Encoder for Representation (BERT) was proposed to predict the ICD codes resulting in F1 value of 0.68. Named Entity Recognition (NER) was modeled using the bidirectional LSTM RNN [36] model and transfer learning technique was used for limited availability of labelled data for Chinese medical records.

2.3 Vector Space Models

Global vectors or GloVe is the fixed length embedding or a numerical vector that is calculated for a text (word) in a given vocabulary, the main idea of the unsupervised GloVe model is to model the ratio of the count co-occurrence probabilities of the words in a large context with the help of a bi-linear log model and minimizing the objective function that is weighted least squares [37]. A count co-occurrence matrix of the words, where the rows of the matrix represent the corresponding word and the columns of the matrix corresponds to the context of the word is first built and then the ratio of the probabilities of the words to the context is modeled so that this huge matrix could be reduced, resulting in a matrix that can be used as a lookup table for any word with its corresponding

embedding vector that is available in the columns. GloVe is working as metric space. This paper adopted a distance function. This research applied equation (1) with the Euclidean distance as: $1/2 \| W_i - W_k\|^2 + bi + \tilde{b}_k = \log (X_{ik})$, where engrossed the squared norms of the embeddings into the biases [11]. The cost function J for the GloVe model is given below:

$$J = \sum_{i,J=1}^{V} f(X_{ij}) \left(- c(d(w_i, \widetilde{w_j}) + b_i + \tilde{b}_j - \text{Log } (X_{ij})\right)^2$$

where c is a function to be selected as a hyperparameter of the model, and d can be any function of distance. Most of the direct correspondence with GLoVe would suggest $c(x) = x^2/2$, sometimes obtained improved results with other functions, such as $c = \cosh^2$..

FastText is based on the continuous skip gram model where the morphology of the words in the languages is also taken into consideration in addition to semantics. The main idea of this approach is that the morphology information is extracted using the "sub-word" units and words are formed by aggregating the n-grams characters. To increase the efficiency of this model, the authors also used a hashing function and optimization techniques that have enabled calculating these vectors representation more quickly and efficiently [13]. The cost function for the FastText is given below

$$\text{CostFunction} = \sum_{t=1}^{T} \left[\sum_{c \in C_t} l(s(w_t, w_c)) + \sum_{n \in N_{t,c}} l(-s(w_t, n)) \right] \quad (2)$$

l is a logistic cost function defined as $l: x \leftrightarrow \log (1 + e^{-x})$
s is an outcoming function $S(w, c) = \sum_{g \in G_w} z_g^T v_c$
w_t is a word at position t
w_c is a context around w_t
$N_{t,c}$ is a set of negative instances sampled from the terminology.
C_t is a context set which consists of indices of the words around the word at position t

Word2Vec (Skip-gram) model one of an unsupervised algorithm that is often used for feature learning. This model predicts the neighbouring words in each window for the target word. To train the words in each sentence, w_1, w_2,....,w_N , where N mentions to the whole word sum, the next impartial equation is exploited [10].

$$P = \frac{1}{N} \sum_{n=1}^{N} \left(\sum_{-c \le j \le c, j \ne 0} \log p(w_{n+j} \mid w_n) \right) \quad (3)$$

The words are represented as an outer summation from the training dataset. The inside summation extends the left context −c and the right context c, computing the log predicting the probability of the word context w_{n+j}, given the input word w_n. The equation of the basic skip-gram defines $p(w_{n+j} \mid w_n)$ using the softmax function. The distribution of probability is defined as

$$P(w_{n+j} \mid w_n) = \frac{\exp(u_{w_{n+1}}^T v_{wn})}{\sum_{v=1}^{V} \exp(u_v{}^T v_{w_n})} \quad (4)$$

V is the vocabulary, the word w vector is a representation of the input u w and output v w. For predicting context word by using input containing a list of words as a window.

BERT (Bidirectional Encoder Representations from Transformers) is used for representing each word into a constant length of numeric vector. There are two pre-trained sub models; the first one is contained 110M parameters architecture of NN, 12 layers, 768 hidden, and 12 head. The second model contained 340 M parameters architecture of NN, 24 layers, 1024 hidden, and 16 head, both models were trained on the 2.500M words English Wikipedia and 800M words Books Corpus [38]. Therefore, this is the state-of-the-art model, the performance of this model with NLP task needs more focus, recently and many researches have examined the performance of this model [39].

Recent studies proposed SentiWordNet methods as an alternative technique compared to other lexicons sentiment [18]. SentiWordNet is defined as an advanced lexicon of sentiment which proves its dependency performance over other sentiment lexicons. It is also proven as a reliable database sentiment among other sentiment competitions. A high performance is obtained when SemEval 2013 is applied to the sentiment analysis of dataset [19]. SentiWordNet is applied on Stanford Twitter Sentiment dataset. It gained the best accuracy of 58.99% and 58.25

% by using MPQA method respectively [20]. In another study, the best accuracy of 72.42% and a second-best accuracy of 70.75% are obtained by applying SentiWordNet MPQA method. The authors claimed that staff in the clinics have been clearly and indirectly defined their reviews and opinions patient by using sentiment analysis lexicon of the discharge notes. Used several Sentiment analysis of discharge notes to relate manually remarks written by clinicians in the medical reports. However, the accuracy of the results is 44.6% on nurse's remarks and 42% on radiology reports. They concluded that using this method is not proven enough to match and analyses the sentiment in discharge notes [21].

Studies on sentiment analysis for evaluating the quality of discharge summary, such as [24] used Hybrid method (Opinion, and AFINN) based on MIMIC database with Accuracy 62%, 68%, [22] they applied Semi-supervised (Random Fields (CRF)) based on Sample size 500 with Accuracy 60%, [25] they used Unsupervised (CNN and SVM) based on Sample size 5000 with Accuracy 58%, and [23] they employed Unsupervised (Word2Vec) based on Sample size 35000 with Accuracy 75%. Therefore, in this study two acute approaches are considered; the first one is the study deals with clinical notes that have much fewer sentiment terms than others. The second approach is that this study is applying unsupervised learning methods that cover more realistic datasets.

3.0 THE TEXT-CLASSIFICATION CHALLENGE

The task of classifying text is considered a challenging task in the Natural Language Processing and so far there is not a single method that can be used in all text classification task. The problems of a set of related documents like discharge summaries classification are formulated as assumed a set of documents, MD, that is, $\mathbf{MD} = (\mathbf{D1}, \mathbf{D2}, \ldots \ldots, \mathbf{DN})$ where Di indicates the ith document in MD, N epitomizes the total number of documents in the dataset. Then, tokenize a document, D_i, to a list of sentences, that is, D = (S1, …, Sn), where Si epitomizes the ith sentence in D and n epitomizes the total number of sentences in each document. The goal final classification goal is for predicting the polarity documents polarity based on polarity of a sentence from MD applying various correlated discharge summaries from related or similar diseases.

Using the technique of SA and evaluating the discharge summary quality by classifying DS documents obtained from the domain of health/medical. Based on the related works, our research is novel as it deals with the issue of classification (positive and negative) or can called (distinctive and non-distinctive). DS using sentiment analysis is suggested and employed the techniques such as TF-IDF, Word2Vec, GloVe, FastText, and BERT as hidden layers for classifier of sentiment analysis at the level of sentences. Five Multi-level features (sentence and word embedding, sentiment feature, medical concept, and linguistic knowledge) are considered an input vector to our model for encoding the features for representing the vector of sentence [40]. These algorithms performance on medical text is still not discovered yet, as only a few studies have been conducted in this area. The features combination such as sentence and word embedding, sentiment knowledge, sentiment shifter, statistical metrics, linguistic and medical concept have not been completely tested for sentiment analysis in the text-domain of healthcare by unsupervised techniques. This work has combined a set of features to deal with, word sense differences, polarity of the contextual, the limit of the word coverage in the general lexicon when dealing with sentiment analysis issues in the healthcare textual, and terms with the context of similar semantics with opposite polarity of sentiment [41].

4.0 METHODOLOGY

The proposed methodology covers five specific levels; the first level is the data collection. The second level covers the data preprocessing, the third level is used vector space models for transferring the list of words to list of vectors. The fourth stage is adopted our statistical method for building a specific lexicon. The last stage is used clustering technique for selecting distinctive documents (positive) and non-distinctive documents (negative). Then the discharge summary text and views are utilized for analysing the effectiveness of the results. In the classification of the results, the data mining method is applied with ''1' positive and '-1' negative. For analyzing results the unsupervised learning method; data modelling and data analysis TF-IDF, Word2Vec, GloVe, FastText, and BERT are used as opinion data mining performing techniques. The final step is evaluating our suggested method using Precession, Recall, Accuracy, and F1 measure and also compare the final results with state-of-the-art methods from the related works. Figure 1 shows the flow chart of the proposed methodology.

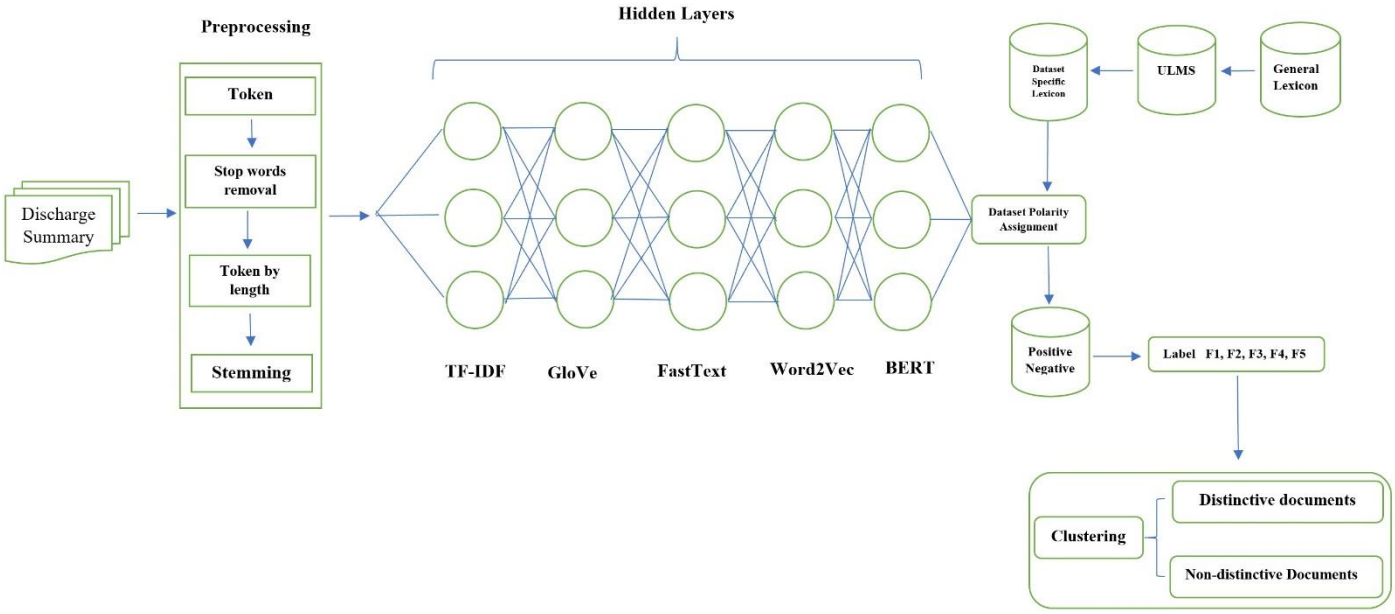

Fig. 1: The various levels of the proposed method.

We now define the structure of our adopted methodology by describing the steps that are used for the classification task.

4.1 Pre-processing

The first level we used before we could use the dataset in our model is to clean it so that the model gets focused on the specific parts of the dataset. For this we first tokenized the dataset based on white spaces and then removed the step words from it. In medical domain the length of the word is crucial as medical dataset contains has typical vocabulary so for this we again tokenized and finally removed the stem of the words so that non repeating tokens could be fed to the model.

4.2 Features Inclusion

We included five types of features in the proposed model which are TF-IDF, Word2Vec, GloVe, FastText and BERT. The idea of combining these features was to give more classification power to the model. The detailed justifications as to why we used these five sets of features is given in the section "justifications of the selected features".

4.3 Language Specific Lexicon for Polarity Assignment

Already available language lexicons like "General Lexicon", "ULMS" and then specific lexicon to the medical dataset are a valuable resource. These lexicons contain informative morphological and syntactical information of the language to assign the text to its polarity of positive and negative.

4.4 Clustering of Documents

To evaluate the quality of discharge summary we grouped the documents into two sets of documents that we called "distinctive" and "non-distinctive" documents using the K-Means clustering algorithm. The main idea using a simple clustering method like K-Means was to get the grouping of the documents as grouping of documents with the particular value of the K is an easy task for the K-Means algorithm and this method is more stable and converges in most scenarios.

4.5 Training and Classification

In the last step we build a classification model with the above described features and language lexicon related dataset. A five layer neural network where each layer corresponds to the features was trained to classify the discharge summary cleaned text in to the positive and negative classes.

5.0 EXPERIMENTAL RESULTS

The challenging task of sentiment analysis based on the proposed approach to check the quality of treatment in the DS has shown promising results. In following sub-sections, explain the suggested model and examine the English language corpus produced by i2b2.

5.1 Data Gathering

This research uses i2b2 server for data collecting, Our experiments on the 1237 de-identified DS, obesity, and 15 other related health diseases. To access the information located at this server which includes all the health data records of major hospitals and clinics. The i2b2 server software is an open-source framework that contains thousands of hospitals and clinics worldwide (https://www.i2b2.org/software/). It manages the electronic health data records and includes all the details related to the human health record data such as discrepancy notes, billing codes, medications, laboratory results, and problem health lists. The current study is only focused on discharge summary, which has written by doctors and clinicians [42]. Table 1 shows the statistics of this dataset; Hence, each existence of the annotations is divided into four categories: Absent, Current, Uncertain, and Unknown. Based on these given annotations, the discharge summary notes are separated into sub-category that links to a specific disease. Then, to determine the separation between each sub-category the category Current is applied. That means discharge summary notes are categorized into exact sub-categories when its category of the disease is presenting.

Table 1: i2b2 Obesity Dataset

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

Notes; “Absent“ means that each discharge summary includes details only information about particular diseases. ”Present” means that each discharge summary includes details on the information about a particular disease and also information about other correlated diseases. “Unmentioned” means that each discharge summary does not mention the information about other correlated diseases. “Questionable“ means that each discharge summary may have information about other correlated diseases, (https://www.i2b2.org/NLP/Obesity/).

5.2 Text Pre-processing

Normalization which is sometimes also called “preprocessing” in the domain of NLP is considered a crucial pre-step to clean the text and preparing in a standardized format that to be used in the training of machine learning or deep learning-based model for classification. Our strategy of developing the automated machine learning model is divided into various levels. In the first level, used steps like removing punctuation, filtering stop words

stemming, lemmatizing, fixed length of vocabulary of size 10,000, and length of sentences between the range of 3 to 25 words for each of those documents, the output of this stage is plaintext.
In the next level, fed our data, word wise to the five embedding models pretrained model to come with the numerical representation of each of the sentence. After that, using the Unified Medical Language System (UMLS) method for building a specific lexicon (medical domain) which helps to assign the polarity to each of the sentences, for more information see [43]. Figure 2 shows the F1 measure for our statistical approach. Lastly, using five key features such as word embeddings, sentiment rules, sentence embedding, shifter sentiment, linguistic, and statistical fields, fed these features representation of the data to the classifier model.

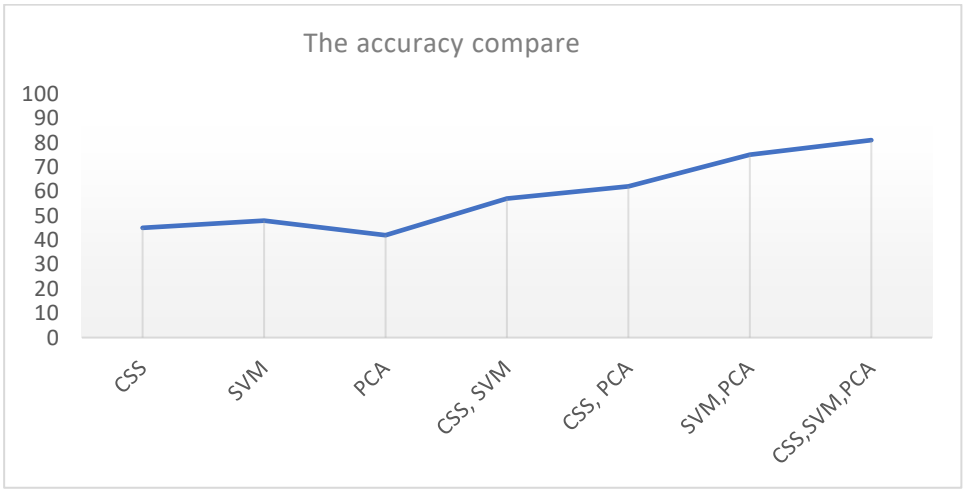

Fig. 2: The F1-measure results, when applied in three weighing approaches with seven various cases.

To analyze the degree to which features of medical notes were particular to a certain patient, two different cohorts have been pulled out from the health system. The first cohort is included all patients admitted to inpatient psychiatric unit with 30 beds between January 2019 to December 2019. The second cohort included all patients admitted to a general emergency unit of major hospital during the same period. Only people aged 18 and above are taken in this study; absolutely no additional exclusion or inclusion criteria have been applied. The main outcome measure of fascination with the psychiatric cohort was some time to psychiatric or maybe all cause clinic readmission, driven by checking out the time following index discharge to determine subsequent admissions. The general healthcare cohort is additionally examined of the time readmission, beside period to all-cause mortality are examined [29]. Figure 3 shows results of this phase, On Y-axis there is a list of terms, and on the X-axis, the dimension-1 of the chosen terms is displayed based on the vector.

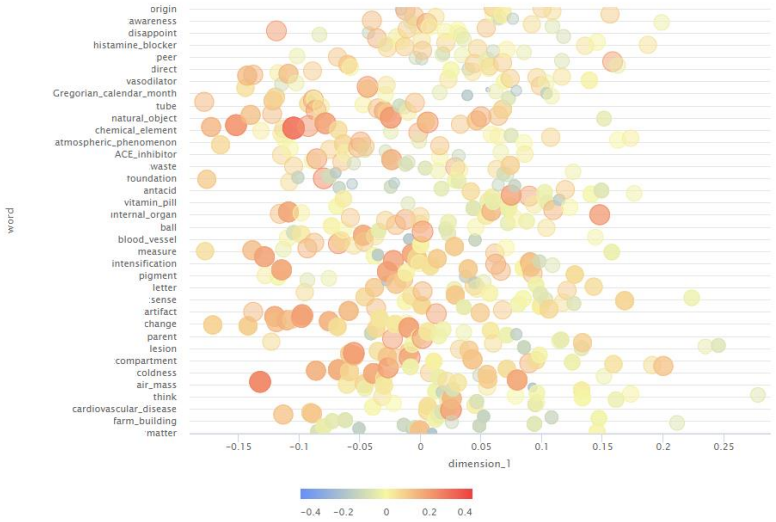

Fig. 3: Vector based words representation

5.3 Vector Space Models

Deep Neural Network (DNN) using the continuous bag-of-words method with other strategies that will be described in the following discussion to deal with the different problems in the model. Used Word2Vec model as pretrained on the famous dataset such as Wikipedia, Google News with the embedding dimension of size 200. The main parameters of our model are the word minimal vocabulary frequency, the layer size, the text window size, and the negatives number were set to 30, 200, 5 respectively. The values of 30 and 5 for the minimal vocabulary frequency and the text window size were chosen after doing experiments with range of values for both the parameters. Solving the issue 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 [12, 41].

This paper adapted GloVe algorithm to hyperbolic spaces and leverage a connection between statistical manifolds of hyperbolic geometry and Gaussian distributions, In order to improve interpret entailment relations between hyperbolic embeddings. Defend the products choice of hyperbolic spaces via this connection to Gaussian distributions and via computations of the hyperbolicity of the symbolic data upon which GloVe is based. [44].

One of the state-of-the-art character-level methods is FastText for embeddings, each term is modeled by sum of vectors, with each vector representing by n-gram. The benefit of this method is that the training process can then share strength across terms composed of common roots to train this approach on English datasets. This research uses the concatenation of WACKYPEDIA and UKWAC which consists of 3.358 billion words. Filter out term kinds that occur fewer than 6 times which results in a vocabulary size of 2,677,244. the main advantages of FastText, it can deal with the out-of-vocabulary (OOV) problem, suggested approaches for probabilistic word representations equipped with flexible sub-word structures, appropriate for out-of-vocabulary and rare words. The suggested probabilistic formulation incorporates uncertainty information and naturally allows one to uncover multiple meanings with multimodal density representations. Our approaches offer improved quality of semantic, outperforming competing approaches on word similarity benchmarks. Additionally, our multimodal density approaches can provide interpretable and disentangled representations, and are the first multi-prototype embeddings that can handle infrequent terms. Our experiments is devised as follows: train a model using FastText with the default setting unless specified otherwise. That is 2M buckets, 10 training epochs, and a learning rate of 0.1. The embeddings dimensionality of d is set to powers of 2 for avoiding edge effects that could make the understanding of the results more problematic [45].

BERT is the one of most popular pre-trained models, capability to deal with sequence task issues like textual classifier, question answering, relation extraction, and sentiment analysis. This model is trained with huge textual corpora, therefore, there is no need to deal with parameters tuning so we done the fine-tuning was adapted for hyper-parameters. Based on the above discussion there are two models namely; BASE and LARGE, the difference between each of them is the number of hidden layers, attention head, the feed-forward network size hidden, and maximum sequence length parameter (the size of accepted input vector), (12 or 24), (12 or 16), (768 or 1024), and (512 or 1024) respectively [14, 39]. In this research, the BASE model is adopted and the hyper-parameters are presented in Table 2. The justifications for choosing the BASE model instead of the LARGE model are the complexity , small dataset and the sequence length. The LARGE model contains too many parameters that leads to overfitting while the small dataset also does not train the LARGE model very well. Lastly, the sequence length is also a constraint as the small sequence length is not appropriate for the LARGE model as the model has issues in fine tuning the parameter and leads to poor training.

Table 2: The values of the fine-tuned Hyper-parameters

Hyper-parameter	Value
Hidden layers	12
Attention heads	12
Epochs	8
Learning rate	0.00003
Gradient accumulation steps	16
Hidden size	768
Parameters	110 M
Maximum sequence length	128

The above values of the parameters are chosen after experimentation with range of possible values. For Epoch and the learning rate it was observed that the high value of Epoch and learning rate did not reduce the loss value.

The architecture of BERT applied with two specific tokens namely; [SEP] and [CLS] the first one used for segment separation and the second one used for a classification task. For the classifier using the first input token, which representing the complete sentence sequence, and for hidden layer size H is the same size for output vector [46]. Therefore, the output of the transformer is the final hidden layer state used as an input for this first token, the vector can be presented as $C \in R^H$. for the fully-connected classification layer output used the vector C as input. The classifier layer matrix parameter set as $W \in R^{KxH}$, where, the K is categories number, the softmax function can be calculated the probability of each category, and present by equation 5;

$$p = sof\,(CW^T) \qquad (5)$$

Where P is the probability of each category, and *sof* is the softmax function.

The transformer is the BERT base. The words sequence is taken from two different sentences, presented by **y** and **x**. [SEP], After x and y, the token is located, while [CLS], before x, the token is located. Embedding function presented by E, and normalization layer presented by NL, the embedding function is given below;

$$\hat{h}_i^0 = E(x_i) + E(i) + E(1_x) \qquad (6)$$
$$\hat{h}_{j+|x|}^0 = E(y_j) + E(j + |x|) + E(1_y) \qquad (7)$$
$$\hat{h}_\cdot^0 = Dro\,(NL(\hat{h}_\cdot^0)) \qquad (8)$$

Where Dro is the Dropout layer [47].

The embedding techniques are passed through blocks of transformer M. Applying the activation function of Element-wise Gaussian Error Linear Units (GELU), function of Multi-Heads Self-Attention (MHSA), and Feed-Forward Layer (FFL) [48], by each block of transformer it is calculated as follow:

$$\hat{h}_\cdot^{i+1} = Skip\,(FFL, Skip\,(MHSA, h_\cdot^i)) \qquad (9)$$
$$Skip(f, h) = NL\,(h + Dro(f(h))) \qquad (10)$$
$$FF(h) = GELU\,(h\,W_1^T + b_1)\,W_2^T + b_2 \qquad (11)$$

Where $h^i \in \mathbb{R}^{(|y|+|x|)*d_h}$, $W_1 \in \mathbb{R}^{4d_h*d_h}$, $W_2 \in \mathbb{R}^{4d_h*d_h}$, $b_1 \in \mathbb{R}^{4d_h}$, $b_2 \in \mathbb{R}^{4d_h}$ and each new $\hat{h}_i$ position is equivalent to:

$$[\ldots, \hat{h}_i, \ldots] = MHSA\left([h_1, \ldots, h_{|y|+|x|}]\right) = W_0\,Concat(h_i^1, \ldots, h_i^N) + b_0 \qquad (12)$$

Instead, it is true in each head of attention that:

$$h_i^j = \sum_{k=1}^{|y|+|x|} Dro\left(\propto_k^{(i,j)}\right) W_{v^{h_k}}^j \qquad (13)$$

$$a_k^{(i,j)} = \frac{\exp \frac{(W_{Q^{h_i}}^J)^T W_{k^{h_k}}^J}{\sqrt{d_h}/N}}{\sum_{k1=1}^{|y|+|x|} \exp \frac{(W_{Q^{h_i}}^J)^T W_{k^{h_k}}^J}{\sqrt{d_h}/N}} \qquad (14)$$

Where $h_i^j \in \mathbb{R}^{(d_h/N)}$, $W_0 \in \mathbb{R}^{(d_h*d_h)}$, $b_0 \in \mathbb{R}^{d_h}$ and $W_Q^j, W_K^j, W_V^j \in \mathbb{R}^{d_h/N*d_h)}$, with N equal to attention heads number [49].

5.4 Clustering

The fourth level is clustering algorithm, which is mainly based on the output from the second phase. This process contains two main steps, namely: implementing the algorithm of a separated centroid-based clustering (k-means) and how to map the clusters. The key idea of this approach is to select the words randomly and distribute those words to the clusters, after that a calculation based on similarity metric is done with the rest of the points and the chosen clusters, this process is repeated until convergence on the chosen clusters similarity metric is achieved [30]. However, this algorithm based on some parameters which are K (for the clusters number), and the second parameter is measure type. For this research, suggest using cosine similarity as a measure for calculating the similarity vectors as it is a common similarity metric (term to term). Therefore, its vector-to-vector similarity metric depends on the angle between vectors is also used in several techniques for example clustering and summarization tasks [31]. However, the list of the word is represented by vectors to compute the cosine similarity [32].

The second issue of selecting the best clustering to solve our sentiment analysis problem. Moreover, the algorithm does not have any prior knowledge and completely unsupervised method. Suggest to adopt the clusters ordering technique, but there is a problem with this based on the size of each cluster, there are some clusters with the same size and the output of this method does not come with good performance. For this, suggest ordering of the cluster, the cluster-ordering relies on the cluster distance performance [50, 51]. The outcome for this phase is a list of the best discharge summaries for reducing the redundancy and dimensionality.

There are two steps for this level; the first step begins with the K-means clustering algorithm used in the experiment, which is considered one of unsupervised techniques. Furthermore, we used the discriminant document based on the embedding document. The main parameters for this algorithm are K, max runs, max optimization steps, and measure type. Assigned these parameters as 2, 10, 100, and cosine similarity for K, numerical values, max runs, and measure type respectively. Used value of K =2 for selecting discriminant (positive) and non-discriminant (negative) documents, the best dataset is based on the discriminant documents. Also, ranking of the clusters is based on the distance performance. Therefore, the low distance between the items in each cluster is considered as the better cluster. Table 3 shows the cluster selection results based on distance performance; the cluster will be selected depends on discriminant documents which shows the split percentage. Finally, Figure 4 contains words clustering based on the vector map. Figure 5. shows the classification of the documents as an example rely on the number of clusters.

Table 3: The clusters selection

Disease	**Cluster Distance Performance**	**The Best Cluster**	**Discriminant %**	**Non-discriminant %**
Asthma	Cluster 0 =1.378 Cluster 1 =1.629	Cluster 0	29	71
CHF	Cluster 0 =1.511 Cluster 1 =1.536	Cluster 0	54	46
CAD	Cluster 0 =1.612 Cluster 1 =1.367	Cluster 1	52	48
Diabetes	Cluster 0 =1.765 Cluster 1 =1.669	Cluster 1	77	23
Depression	Cluster 0 =1.642 Cluster 1 =1.644	Cluster 0	43	57
GERD	Cluster 0 =1.689 Cluster 1 =1.648	Cluster 1	68	32
Gallstones	Cluster 0 =1.646 Cluster 1 =1.623	Cluster 1	30	70
Gout	Cluster 0 =1.659 Cluster 1 =1.572	Cluster 1	60	40
Hypertension	Cluster 0 =1.644 Cluster 1 =1.715	Cluster 0	70	30
Hypercholesterolemia	Cluster 0 =1.665 Cluster 1 =1.436	Cluster 1	24	76
Hypertriglyceridemia	Cluster 0 =1.669 Cluster 1 =1.768	Cluster 0	77	23
Obesity	Cluster 0 =1.673 Cluster 1 =1.761	Cluster 0	77	23
OA	Cluster 0 =1.485 Cluster 1 =1.676	Cluster 0	35	65
OSA	Cluster 0 =1.614 Cluster 1 =1.574	Cluster 1	23	77
Venous Insufficiency	Cluster 0 =1.668 Cluster 1 =1.764	Cluster 0	77	23
PVD	Cluster 0 =1.764 Cluster 1 =1.669	Cluster 1	77	23

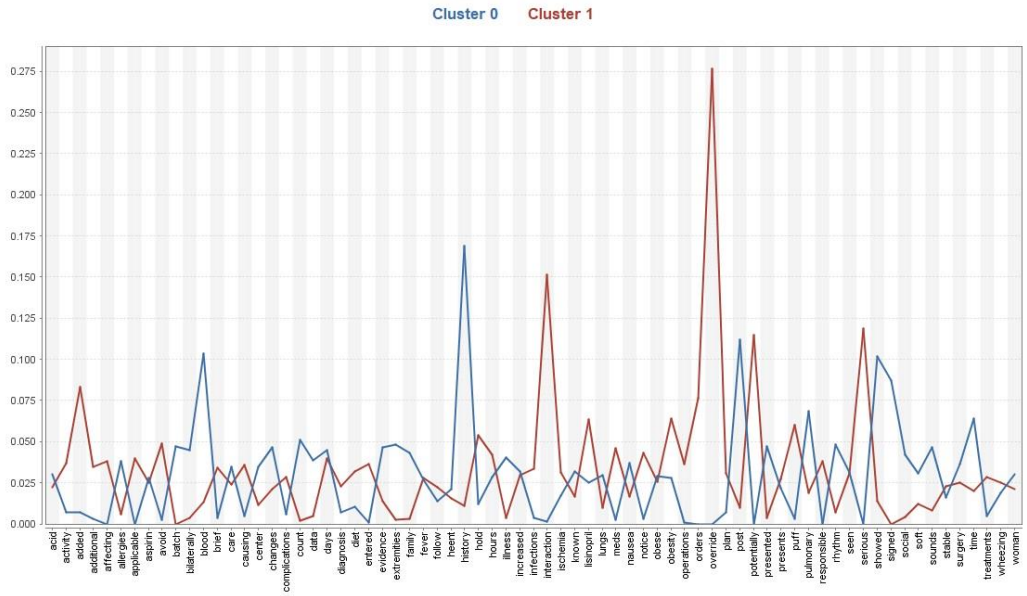

Fig. 4: Shows the clusters number which are represented by two various colors, the blue represents the first cluster, and red represents the second cluster. The y-axis shows the percentage of each word accrue in the clusters. Meanwhile, the x-axis shows the most significant words represented by vector space model

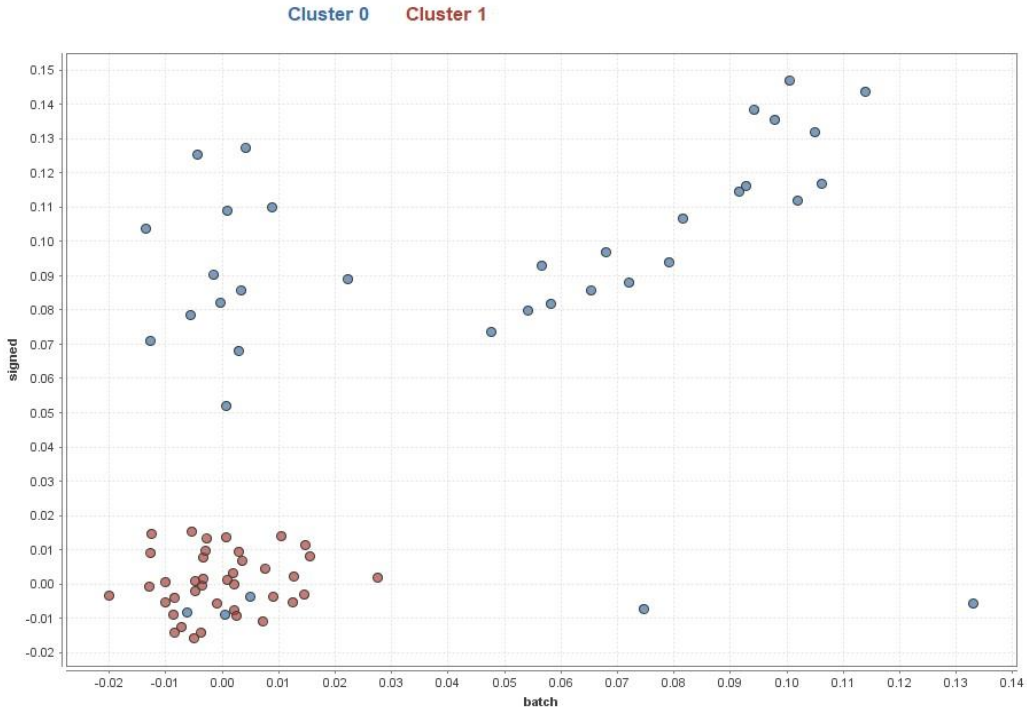

Fig. 5: Illustrates the documents classification as an example rely on the number of clusters.

The unsupervised text sentiment analysis for testing the treatment quality that relies on discharge summary into sentence level is necessary by the use of Algorithm 1 as follows. We used three weighting scheme to give more power to the generated Lexicon as the single weighted technique did not result in good accuracy for the examination of the discharge summary. This proves that a combination of weighting scheme converts the final score into the latent dimension that has more efficiency and is more robust as opposed to a single weighting technique.

Algorithm 1. examined the quality of treatment based on discharge summary

- **INPUT**: D (documents normalization).
- **OUTPUT**: Distinctive (positive =1) and Non-distinctive (negative = -1).

1: Pre-processing text (toke, stop word removal, token by length, and stemming).
2: Fed the list of words to deep neural network.
3: Transfer the words to vector by (TF-IDF, Glove, FastText, Word2Vec, and BERT).
4: Generated specific lexicon by using UMLS, weighted by (SVM), Chi-squared statistic (CSS), and (PCA).
5: for each W_i in Sentence S **do**
6: **SET** Total score (W_i) = 0
7: **SET** Index = polirty (1 or -1)
8: **for** each method M_j **do**
9: score **(**M_j , W_i **)** = compute score (M_j , W_i, S)
10: Total score (W_i) = Total score (W_i) + score (M_j , W_i)
11: **end for**
12: List.add (Total score (W_i))
13: **if** Total score (W_i) > 0 set 1
14: Total score (W_i) < 0 set -1 **then**
15: Polarity assignment
16: **end if**
17: end for
18: Assign the polarity for each sentence.
19: Transfer the sentences to fixed length of vectors by five features (sentence and word embedding, sentiment feature, medical concept, and linguistic knowledge).
20: Used K-means algorithm for classification the sentences to 1 or-1.

5.5 Evaluation of the Final Sentiment Analysis Result

The evaluation metrics are used to measure the usefulness and trustfulness of the unsupervised suggested model. Sentiment analysis for testing the qualities, like, redundancy, noisy information, completeness, and legibility. There are three main measures for evaluating the worth of any approach of recall, precision, F- measure, and accuracy [52-54]. These methods used the following standard metric as listed:

$$\text{Recall} = \text{True positive}/(\text{True positive} + \text{False negative}) \quad (15)$$

$$\text{Precision} = \text{True positive}/(\text{True positive} + \text{False positive}) \quad (16)$$

$$\text{F} - \text{measure} = 2 * (\text{Precision} * \text{Recall})/(\text{Precision} + \text{Recall}) \quad (17)$$

$$\text{Accurcy} = \frac{\text{True positive+True negative}}{(\text{True positive+True negative+False positive+False negative})} \quad (18)$$

Unsupervised results are reported by accuracy metric and the outcomes of varying our approach are presented in Table 4. The final results for our method based on the list of features are reported in Table 5. Table 6 shows the comparison of this work with other state-of-the-art researches. The efficiency was also calculated for compared approaches based on running time, see Figure 6. This research working with obesity and 15 related diseases, the patterns of connectivity are required between positive and negative labels related to doctor opinions, especially when the patient has more than one disease, the network connection as shown in Figure 7. We have adopted some other state-of-the-art techniques from related work for examining the reliability and efficiency of our proposed model as shown the results in Figure 8.

Table 4: Results based on proposed approach

Our Method	M1	M2	M3	M4	M5	Pr	Rec	F-measure	Accuracy
Experiment 5	+	+	+	+	+	0.982	0.934	0.958	0.930
Experiment 4	+	+	+	+		0.970	0.910	0.938	0.890
Experiment 3	+	+	+			0.964	0.897	0.896	0.876
Experiment 2	+	+				0.937	0.854	0.850	0.816
Experiment 1	+					0.991	0.865	0.784	0.680

Notes: Discharge summary (DS), Sentiment Analysis (SA), Precision (Pr), and Recall (Rec)
DS-SA-TF-IDF =M1 , DS-SA-GloVe =M2, DS-SA-FastText =M3, DS-SA-Word2Vec = M4, , and DS-SA-BERT = M5

Table 5: Results based on proposed features set

Our Method	X1	X2	X3	X4	X5	Pr	Rec	F-measure	Accuracy
Experiment 5	+	+	+	+	+	0.982	0.934	0.958	0.930
Experiment 4	+	+	+	+		0.8677	0.6572	0.7477	0.878
Experiment 3	+	+	+			0.8178	0.6195	0.7050	0.818
Experiment 2	+	+				0.8406	0.6371	0.7248	0.682
Experiment 1	+					0.7869	0.5998	0.6807	0.601

* **X1** = feature 1, **X2** = feature 2, **X3** = feature 3, **X4** = feature 4, **X5** = feature 5.

Table 6: Comparison of this work with related approaches

Study	Methods	Approaches	Sample size	Accuracy
[25]	Unsupervised	CNN (SVM)	5000	58%
[22]	Semi-supervised	Random Fields (CRF)	500	60%
[24]	Hybrid	Opinion, AFINN	MIMIC database	62%, 68%
[23]	Unsupervised	Word2Vec	35000	75%
Our work	Unsupervised	TF-IDF, Word2Vec, GloVe, FastText, BERT & Clustering	5500	93%

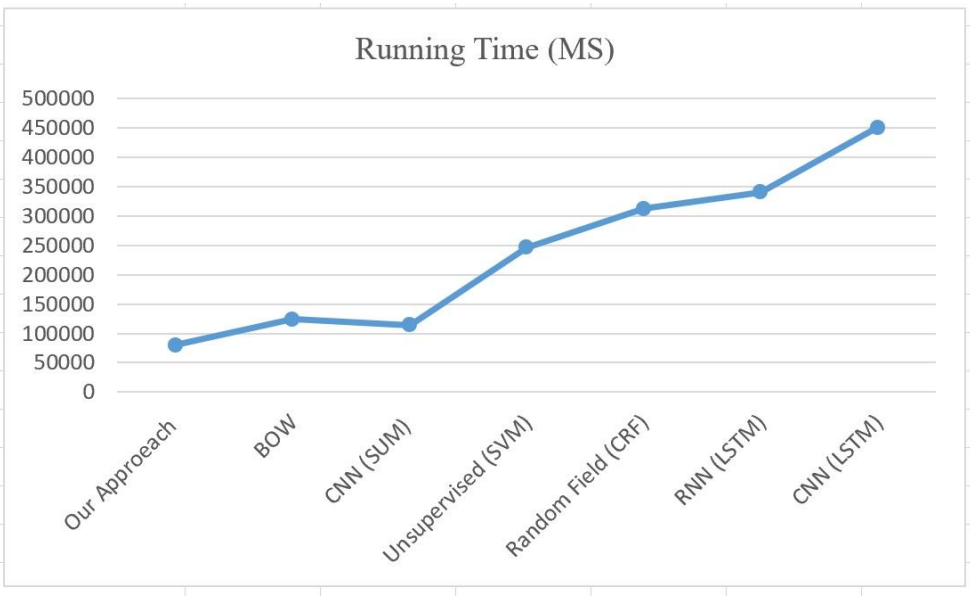

Fig. 6: The time complexity for each approach

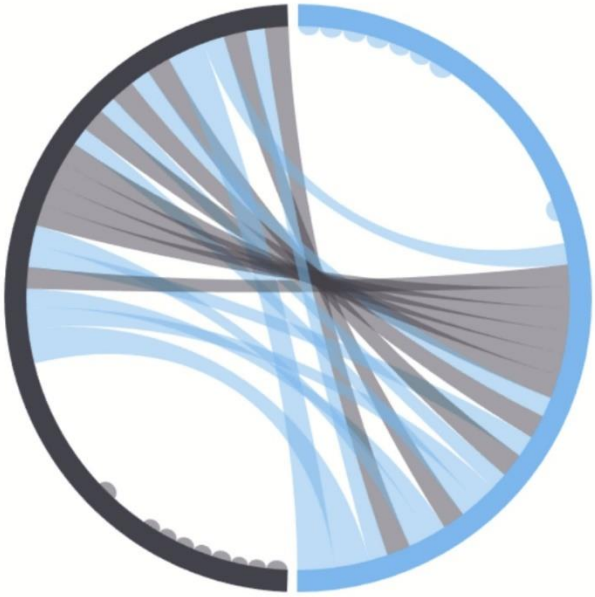

Fig. 7: The polarity (positive and negative) labels are connectivity based on diseases related. The blue color represented the positive label and the black color represented the negative label.

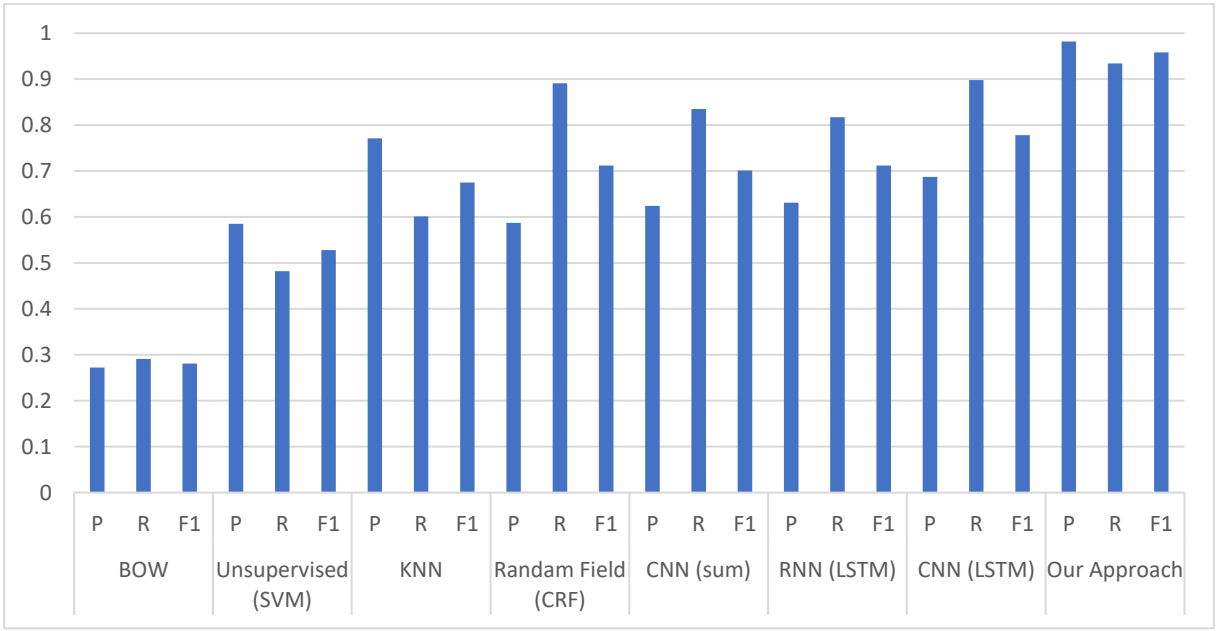

Fig. 8: The comparison based on statistics Performance between our method and other approaches, using, precision (P), recall (R), and F1 measures (F1)

6.0 DISCUSSION

The suggested method applies the unsupervised approaches which depend on the vector space models, statistical, and clustering. Our combined multi-vector space models at the medical text field, which contains the TF-IDF, Word2Vec, GloVe, FastText, and BERT has been fully investigated as hidden layers based on deep learning idea. The combined multi-level of the feature set at the medical text field, which includes the sentence and word embedding, sentiment knowledge, rules of sentiment shifter, medical concept, linguistic, and statistical knowledge has been fully investigated. Based on the Previous discussion cases examined, our proposed approach uses binary classification cases based on unsupervised method and reports the evaluation results in terms of precision, recall, F1, and accuracy metrics. For our experiment, the highest F1 measure is 0.958 and accuracy is 0.930. This paper suggested an approach for verifying that unsupervised (TF-IDF, Word2Vec, GloVe, FastText, BERT, and

clustering) techniques proofs an enhancement particularly for our experiments when compared between each of them and enhancement particularly when compared with different state-of-the-art methods.

6.1 Justifications of Selected Features

We now discuss the reasons for inclusion and the significance of each of the features that are used in our study.

6.1.1 TF-IDF

Our first feature used was the TF-IDF and this feature is used in many tasks on NLP due to the simplicity of this feature in mathematical calculation. In our study since the dataset is specifically related to the medical domain so the terms related to the medical domain like “blood cancer”, “ulcer” and “leukemia” are more important rather than those words that occur frequently in various other domains.

6.1.2 Word2Vec

Related or similar words are more important than the words that are not relevant to each other and since Word2Vec finds the embedding of words based on relatedness and closeness so it is more appropriate. Our medical dataset also have various medical terms that are more related for example terms like “disease” and “illness” , “cancer” and “leukemia” are more related to each other

6.1.3 GloVe

Although both GloVe and Word2Vec are used for the same purpose ,that is to get the embeddings of the words present in the corpus but the training methodology of the GloVe is different from the Word2Vec. GloVe is trained using the word concurrence matrix rather than the neighboring or local context used in Word2Ve. We believe that both of these models would produce features that when trained on our Dataset would give valuable information from both local and global perspective. For example the term “disease” and “illness” are more close to each other whereas the term “ulcer” and “disease” are far away from each other but still are related.

6.1.4 FastText

Sub words information of a word is also important many a times as these sub words contains meaningful morphological characteristics of the whole world. This is the idea proposed by FastText based embedding’s that we used in this study. It is specifically the case in medical domain that many words have different sub words that can be joined to forms related words. To give an example consider the word “Pathology”, “Gynecology” and “Ophthalmology” as all of these words include sub words that are meaningful in their own and are related to each other. Due to this specific morphological extraction, we have used this feature in our study for the classification task.

6.1.5 BERT

Bi-Directional models that base models in both left to right and right to left direction, extracts meaningful information from the model. Pre-trained BERT model that is already trained on the huge corpus so that it could lean meaningful language related information is fine tuned in our medical data so that both the general language features and features specific to the medical domain could be combined is used in this study.

6.2 Comparison with Approaches in the Literature

We have compared our models with both unsupervised and semi-supervised models that are used in the literature. We now discuss why our approach is superior to the methods that we compared.

6.2.1 Combining Multiple Features

We have used multiple features in our proposed approach whereas the techniques that are used in the literature did not combine multiple approaches and most importantly embeddings techniques were never combined in the literature. Moreover, features that are considered individually redundant when combined with the bunch of other features also many a times give more power to the classification model.

6.2.2 Power of Large Pre-Trained Models

Of late there are a large number of pre-trained models that are provided by Google, Facebook and OpenAI , all these model are trained on huge corpus with expensive and massive computational power that can be trained on personal computers. These models contain hundreds of millions of parameters that are adjusted in the training phase and capture valuable insight on the dataset. Utilizing these publicly available resources, we have incorporated BERT pre-trained based features in our model for better classification.

It is pertinent to discuss here the computational complexity of our proposed approach as it is directly proportional to the size of dataset or the number of documents in the dataset to be exact. The more the number of documents in the dataset, the more it will take to extract useful information from those documents. But, a large dataset of documents is always useful in the domain of NLP as the large set of documents-based machine learning model in the preprocessing stage extract the diverse set of useful information from the documents which leads to better cross validation-based accuracy.

7.0 CONCLUSION

In this work, the sentiment analysis of discharge summary notes based on SentiWordNet and building specific medical text domain by statistical techniques and used Word2Vec, GloVe, FastText, and BERT as unsupervised learning method is presented. The study is performed over 1000 discharge summary notes collected from medical institutions. (TF-IDF, Word2Vec, GloVe, FastText, and BERT) as deep learning techniques and clustering are used to reduce the dimensionality of the word vectors by calculating the fixed length embedding vectors that are able to differentiate between datasets sentiment terms consisting of positive and negative by using the sentiment scores. Sub-categories of summary notes relating to certain diseases are statistically analyzed by using SentiWordNet including vector space models. Our proposed method is promising in a sense that it can be deployed in health care centres and can be used by the medical health centre staff to produce effective administration results and better treatment quality for the patients.

For future research in this problem, we would focus on the end-to-end approaches where these hand crafted features get replaced by the automatic feature selection approaches. A real challenge would be to design the strategy which can not only give the good classification accuracy but also could give interpretable set of features the can be explained in accordance with the medical domain.

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, administration and moral support.

REFERENCES

[1] F. López-Martínez, E. R. Núñez-Valdez, V. García-Díaz, and Z. J. A. Bursac, "A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management," vol. 13, no. 4, p. 102, 2020.

[2] N. Garcelon, A. Burgun, R. Salomon, and A. J. K. I. Neuraz, "Electronic health records for the diagnosis of rare diseases," 2020.

[3] S. H. J. J. o. C. P. Kollins and Psychiatry, "From risk prediction to action: leveraging electronic health records to improve pediatric population mental health," vol. 61, no. 2, pp. 113-115, 2020.

[4] I. Liu, S. Ni, K. J. T. Peng, and e-Health, "Happiness at Your Fingertips: Assessing Mental Health with Smartphone Photoplethysmogram-Based Heart Rate Variability Analysis," 2020.

[5] R. A. Archbold, K. Laji, A. Suliman, K. Ranjadayalan, H. Hemingway, and A. D. J. B. j. o. g. p. Timmis, "Evaluation of a computer-generated discharge summary for patients with acute coronary syndromes," vol. 48, no. 429, pp. 1163-1164, 1998.

[6] R. N. Axon *et al.*, "A hospital discharge summary quality improvement program featuring individual and team-based feedback and academic detailing," vol. 347, no. 6, pp. 472-477, 2014.

[7] J. M. Steinkamp, W. Bala, A. Sharma, and J. J. J. J. o. B. I. Kantrowitz, "Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes," vol. 102, p. 103354, 2020.

[8] B. Starmer, M. Barton, and H. J. J. o. p. u. Corbett, "Automated electronic discharge summary for patients undergoing acute scrotal exploration: does it improve accuracy and quality?," vol. 15, no. 6, pp. 609. e1-609. e4, 2019.

[9] M. Cui *et al.*, "Risk assessment of sarcopenia in patients with type 2 diabetes mellitus using data mining methods," vol. 11, 2020.

[10] Y. Goldberg and O. J. a. p. a. Levy, "word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method," 2014.

[11] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532-1543.

[12] S. A. Waheeb, N. A. Khan, B. Chen, and X. J. I. Shang, "Multidocument Arabic Text Summarization Based on Clustering and Word2Vec to Reduce Redundancy," vol. 11, no. 2, p. 59, 2020.

[13] A. Joulin, E. Grave, P. Bojanowski, M. Douze, H. Jégou, and T. J. a. p. a. Mikolov, "Fasttext. zip: Compressing text classification models," 2016.

[14] V. Sanh, L. Debut, J. Chaumond, and T. J. a. p. a. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," 2019.

[15] F. B. Silva, R. d. O. Werneck, S. Goldenstein, S. Tabbone, and R. d. S. J. P. R. Torres, "Graph-based bag-of-words for classification," vol. 74, pp. 266-285, 2018.

[16] Y. Chen, E. A. Silva, J. P. J. R. S. P. Reis, and Practice, "Measuring policy debate in a regrowing city by sentiment analysis using online media data: a case study of Leipzig 2030," 2020.

[17] S. Ruseti, M.-D. Sirbu, M. A. Calin, M. Dascalu, S. Trausan-Matu, and G. Militaru, "Comprehensive Exploration of Game Reviews Extraction and Opinion Mining Using NLP Techniques," in *Fourth International Congress on Information and Communication Technology*, 2020, pp. 323-331: Springer.

[18] M. Emmert, N. Meszmer, and M. J. B. h. s. r. Schlesinger, "A cross-sectional study assessing the association between online ratings and clinical quality of care measures for US hospitals: results from an observational study," vol. 18, no. 1, p. 82, 2018.

[19] J. Xu *et al.*, "Sentiment analysis of social images via hierarchical deep fusion of content and links," vol. 80, pp. 387-399, 2019.

[20] G. Mujtaba *et al.*, "Clinical text classification research trends: systematic literature review and open issues," vol. 116, pp. 494-520, 2019.

[21] A. Yadollahi, A. G. Shahraki, and O. R. J. A. C. S. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining," vol. 50, no. 2, pp. 1-33, 2017.

[22] K. P. Chodey and G. Hu, "Clinical text analysis using machine learning methods," in *2016 IEEE/ACIS 15th International Conference on Computer and Information Science (ICIS)*, 2016, pp. 1-6: IEEE.

[23] K. Sugathadasa *et al.*, "Synergistic union of word2vec and lexicon for domain specific semantic similarity," in *2017 IEEE International Conference on Industrial and Information Systems (ICIIS)*, 2017, pp. 1-6: IEEE.

[24] G. E. Weissman, L. H. Ungar, M. O. Harhay, K. R. Courtright, and S. D. J. J. o. b. i. Halpern, "Construct validity of six sentiment analysis methods in the text of encounter notes of patients with critical illness," vol. 89, pp. 114-121, 2019.

[25] S. Yadav, A. Ekbal, S. Saha, and P. Bhattacharyya, "Medical sentiment analysis using social media: towards building a patient assisted system," in *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.

[26] S. N. Murphy *et al.*, "Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2)," vol. 17, no. 2, pp. 124-130, 2010.

[27] D. T. Heinze, M. L. Morsch, B. C. Potter, and R. E. J. J. o. t. A. M. I. A. Sheffer Jr, "Medical i2b2 NLP smoking challenge: the A-Life system architecture and methodology," vol. 15, no. 1, pp. 40-43, 2008.

[28] I. E. Waudby-Smith, N. Tran, J. A. Dubin, and J. J. P. o. Lee, "Sentiment in nursing notes as an indicator of out-of-hospital mortality in intensive care patients," vol. 13, no. 6, 2018.

[29] Y. Xiong, B. Tang, Q. Chen, X. Wang, and J. Yan, "A Study on Automatic Generation of Chinese Discharge Summary," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2019, pp. 1681-1687: IEEE.

[30] M. Ashouri, F. Haghighat, B. C. Fung, A. Lazrak, H. J. E. Yoshino, and Buildings, "Development of building energy saving advisory: A data mining approach," vol. 172, pp. 139-151, 2018.

[31] H. Liang *et al.*, "Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence," vol. 25, no. 3, pp. 433-438, 2019.

[32] S. Gholizadeh, A. Seyeditabari, W. J. B. D. Zadrozny, and C. Computing, "Topological signature of 19th century novelists: Persistent homology in text mining," vol. 2, no. 4, p. 33, 2018.

[33] F. J. I. A. Bu, "A high-order clustering algorithm based on dropout deep learning for heterogeneous data in cyber-physical-social systems," vol. 6, pp. 11687-11693, 2017.

[34] Q. Chen and M. J. S. C. S. Sokolova, "Specialists, Scientists, and Sentiments: Word2Vec and Doc2Vec in Analysis of Scientific and Medical Texts," vol. 2, no. 5, pp. 1-11, 2021.

[35] T.-S. Heo, Y. Yoo, Y. Park, and B.-C. J. a. p. a. Jo, "Medical Code Prediction from Discharge Summary: Document to Sequence BERT using Sequence Attention," 2021.

[36] X. Dong *et al.*, "Deep learning for named entity recognition on Chinese electronic medical records: Combining deep transfer learning with multitask bi-directional LSTM RNN," vol. 14, no. 5, p. e0216046, 2019.

[37] K. Chen *et al.*, "Defect Texts Mining of Secondary Device in Smart Substation with GloVe and Attention-Based Bidirectional LSTM," vol. 13, no. 17, p. 4522, 2020.

[38] S. Sun, Y. Cheng, Z. Gan, and J. J. a. p. a. Liu, "Patient knowledge distillation for bert model compression," 2019.

[39] Y. Liu *et al.*, "Roberta: A robustly optimized bert pretraining approach," 2019.

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

[41] S. A. Waheeb, N. A. Khan, B. Chen, and X. J. I. Shang, "Machine Learning Based Sentiment TextClassification for Evaluating Treatment Quality of Discharge Summary," *MDPI Information*, vol. 11, no. 5, 2020.

[42] P. C. Nair, D. Gupta, B. I. Devi, and N. R. Bhat, "Automated Clinical Concept-Value Pair Extraction from Discharge Summary of Pituitary Adenoma Patients," in *2019 9th International Conference on Advances in Computing and Communication (ICACC)*, 2019, pp. 258-264: IEEE.

[43] 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," vol. 11, no. 5, p. 281, 2020.

[44] A. Tifrea, G. Bécigneul, and O.-E. J. a. p. a. Ganea, "Poincar\'e GloVe: Hyperbolic Word Embeddings," 2018.

[45] T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. J. a. p. a. Joulin, "Advances in pre-training distributed word representations," 2017.

[46] J. Canete, G. Chaperon, R. Fuentes, and J. J. P. D. a. I. Pérez, "Spanish pre-trained bert model and evaluation data," vol. 2020, 2020.

[47] J.-S. Lee and J. J. a. p. a. Hsiang, "Patentbert: Patent classification with fine-tuning a pre-trained bert model," 2019.

[48] D. Hendrycks and K. Gimpel, "Bridging nonlinearities and stochastic regularizers with gaussian error linear units," 2016.

[49] T. Pires, E. Schlinger, and D. J. a. p. a. Garrette, "How multilingual is multilingual bert?," 2019.

[50] F. J. I. A. Bu, "A high-order clustering algorithm based on dropout deep learning for heterogeneous data in cyber-physical-social systems," vol. 6, pp. 11687-11693, 2018.

[51] S. A. Waheeb, H. J. I. J. o. A. i. S. Husni, and Technology, "Multi-Document Arabic Summarization Using Text Clustering to Reduce Redundancy," vol. 2, no. 1, pp. 194-199, 2014.

[52] A. Abdi, S. M. Shamsuddin, S. Hasan, J. J. I. P. Piran, and Management, "Deep learning-based sentiment classification of evaluative text based on Multi-feature fusion," vol. 56, no. 4, pp. 1245-1259, 2019.

[53] H. Sankar *et al.*, "Intelligent sentiment analysis approach using edge computing-based deep learning technique," 2019.

[54] A. Wong, J. M. Plasek, S. P. Montecalvo, L. J. P. T. J. o. H. P. Zhou, and D. Therapy, "Natural language processing and its implications for the future of medication safety: a narrative review of recent advances and challenges," vol. 38, no. 8, pp. 822-841, 2018.