12-2021

Development of an automated physician review classification system: A novel semi-supervised learning approach

sagarika suresh thimmanayakanapalya

Pavankumar Mulgund

Raj Sharman

Development of an automated physician review classification system: A novel semi-supervised learning approach

Keep one for your paper: Completed Research Paper

Sagarika Suresh
University at Buffalo
sthimman@buffalo.edu

Pavankumar Mulgund
University at Buffalo
pmulgund@buffalo.edu

Raj Sharman
University at Buffalo
rsharman@buffalo.edu

Abstract

Building automated text classifiers have assumed significant importance since the development of sizeable online information platforms. In addition, several compelling use cases have emerged in the field of artificial intelligence and analytics in recent years. However, building and training text classifiers become problematic in the healthcare context, which deals with a sensitive and limited volume of data. In this paper, we explore the development of a classifier and apply it to a specific case of classifying physician reviews into either clinical and non-clinical reviews. The primary purpose of this paper is to demonstrate the methodology using which the classifier has been developed, including a novel technique in curating datasets.

We leverage unsupervised guided Latent Dirichlet Allocation (LDA) method and supervised methods such as deep neural networks, Long-Short Term Memory (LSTM) networks, and Bi-directional LSTMs. Further, we compare the various models and choose the one with the best classification performance by validating the output results with the ground truth. Our methodology provides insights into making the best use of semi-supervised and supervised algorithms along with grounded data for developing classifiers that can be generalized for other novel contexts where dataset availability is limited.

Keywords

Classifier, Guided Latent Dirichlet allocation algorithm, LSMT, Bi-LSTM, Grounded theory

Introduction

Modern deep learning methods and unsupervised classification algorithms enable significant machine learning capabilities without the need for substantial feature engineering (Pouyanfar et al., 2018). While previously implemented approaches are powerful when massive amounts of training data are available to create models, the value of unsupervised algorithms comes from their capacity to learn general-purpose representations from vast unlabeled corpora (Shrestha & Mahmood, 2019). These representations provide structured input for future machine learning analysis that successfully captures lexical semantics without explicitly describing that meaning as features using natural language processing (NLP) techniques.

The widespread use of open-source industrial-standard toolkits for creating supervised and unsupervised classification models enables the development of machine learning components for various classification applications. However, unsupervised learning algorithms learn from raw data and do not require any prior knowledge. It is also time-consuming because the learning phase of the algorithm may take much time (Figueiredo & Jain, 2002). Supervised learning algorithms such as deep learning systems produce data output from the previous experience. However, in contexts where previous data is not available, the limitations of supervised learning algorithms have been well documented (Caruana & Niculescu-Mizil, 2006). We have tried to improve the combination of these algorithms by tuning feature selection for supervised algorithms and proposing better classification strategies using semi-supervised algorithms (Guided LDA) (Zhu & Goldberg, 2009). This paper has utilized the best features of semi-supervised and supervised learning and introduced a novel method of validating a built classifier with the *ground truth*. This strategy is beneficial in cases where classifiers are built in new contexts.

The paper demonstrates the methodology of building a classifier by implementing this in the healthcare context where access to a large volume of data is difficult. This paper attempts to build a classifier to categorize physician reviews based on a popular physician rating website (PRW) as clinical vs. non-clinical texts. Our future work will include a multi-level classification where reviews can belong to multiple categories. For example, clinical reviews mainly focused on treatment procedure, doctor competence, and knowledge, whereas non-clinical reviews mainly were about bedside manners, staff friendliness, ease of appointment scheduling. The theoretical foundation of what constitutes clinical and non-clinical reviews is based on prior work in the healthcare literature (Rothenfluh, F., & Schulz, P. J., 2018). In such specific contexts, it becomes challenging to find labeled datasets, and consequently, the use of unsupervised learning algorithms becomes inevitable. However, the significant limitations in such cases would be that these algorithms can be highly open-ended and do not necessarily help us build a classifier that classifies the physician reviews according to our needs. Therefore, we adopted a novel semi-supervised guided latent Dirichlet allocation algorithm to identify the two main classification topics (clinical and non-clinical) from the seed topics we(researchers) provided.

Additionally, the semi-supervised algorithm allowed us to recognize similar word patterns in the document that helped develop a labeled corpus. This labeled corpus was further used for the development of an LSTM and Bi-LSTM classifier. Furthermore, the results of the classifiers were tested on the physician reviews.

In the following sections, we will look at the literature review in Section 2. Then, we introduce our data collection procedure and experiment design in Section 3. Next, our model development process and will be presented in Section 4. Section 5 presents the results and comparison of models, and Section 6 speaks of some of the limitations of our paper. Finally, in Section 7, we bring out the conclusions of our paper.

Literature Review

For some period, text classifiers have been intensively developed. Several publications have created a text classification system, particularly for online physician reviews utilizing support vector machines and random forests ((Boser et al., 1992); (Breiman (2001); (Zhuo et al., 2008)) that leverage the statistical properties of the review text, such as the frequency of each word. However, our paper does not just include the frequency of words to find word similarity but also conducts a forward and background propagation using LSTM networks from a labeled corpus that has been qualitatively and automatically labeled using a hierarchical guided LDA approach.

Deep learning techniques, such as convolutional neural networks (CNN) (Zhang & Wallace, 2015), take the proximity of words into account but do not focus on the context of the words themselves but rather on the labeled corpora supplied to them. Additionally, work has been published that focuses on developing a

classifier utilizing natural language processing, leveraging the dependency tree of a review phrase (Li et al., 2011). Specifically, several studies examine current natural language processing (NLP) classifiers and suggest new ones, such as the Dependency tree-based classifier (DTC) (Li et al., 2011). However, dependency tree-based classifiers are known to focus exclusively on the syntactic structure of the word structure rather than the semantics.

Previously published articles have used text-mining methods to characterize patterns in physician reviews. For example, Wallace et al. (2014) created a probabilistic generative model to capture latent sentiment across many dimensions of care. This, however, places a premium on the emotion of the evaluations rather than on collecting patterns and categorizing contextual aspects across them. They demonstrated that including the output of their algorithm into regression models enhances correlations with state-level quality indicators. Hao and Zhang utilized topic modeling to identify similar themes in doctor reviews collected from Good Doctor Online across four specialties (Hao & Zhang, 2016). They discovered four common themes across the four specialties: the process of locating doctors, technical abilities or bedside manner, patient appreciation, and symptoms description. However, this article concentrates exclusively on subject modeling. Our article advances this step by utilizing guided topic modeling and then constructing a classifier by comparing three different neural network topologies. Similarly, Hao et al. compared reviews between Good Doctor Online and the US doctor review website RateMDs using topic modeling (Hao et al., 2017). While they discovered similarities between the two places, they also found variances representing the two countries' health care systems.

Hu and Liu used a four-step algorithm to judge features from customer reviews (Hu & Liu, 2004). This method identifies features using association rule mining, prunes uninteresting and redundant features, identifies uncommon features, and ultimately determines the semantic orientation of each opinion statement. Popescu and Etzioni developed an unsupervised method for extracting product features and opinions from product reviews (Popescu & Etzioni, 2007). After identifying an explicit characteristic in a sentence, they extracted the heads of probable opinion statements using manually designed extraction methods. This technique is only applicable when features are specified explicitly. Our study differs from theirs in terms of the labeled corpora that we created utilizing a hierarchical guided LDA method in addition to a qualitative data creation strategy.

Agarwal et al. extracted dependency tree patterns from phrases using numerous hand-crafted methods (Agarwal et al., 2015). Next, they combined this data with the semantic information in the Massachusetts Institute of Technology Media Lab ConceptNet ontology. Finally, they used the extracted concepts to train a machine learning model to recognize concept patterns in text, then classify documents into positive and negative categories.

Wawer (2015) generated dependency patterns by using target-sentiment (T-S) pairings and recording the dependency routes between T and S - words in their corpus's dependency tree. The patterns were augmented with conditional random fields to identify targets of opinion words. Our study differs in that it focuses on integrating qualitative and quantitative research, such as semi-supervised and supervised learning algorithms, on establishing a logical approach for developing a classifier. Additionally, it considers the algorithm's performance accuracy and validates the method with qualitatively grounded labeled data, which provides additional context for the evaluations.

Experimental Design

Dataset development

There were two types of datasets to be used for our methodology. One is the ground labeled data, and the other being the tagged corpora based on widely available clinical and non-clinical data.

Ground truth

The ground truth data was required to understand how physician reviews are broadly segregated. Earlier work by Tavakol et al. (2006) argued that reviews in RateMD could be classified into two main topics, which included clinical and non-clinical reviews. To develop this dataset, we collected datasets from a major PRW, which consisted of actual physician ratings. Three researchers qualitatively coded the data using grounded theory, and there was an inter-rater agreement reliability score that was checked after the coding was conducted. Each of the ratings was classified as either clinical or non-clinical based on the metrics manual for the coding developed and curated from previous literature (Tavakol et al., 2006).

Table 1 below gives a sense of how the data was coded as either clinical or non-clinical. Only some of the metrics of coding have been provided in Table 1. Full details of this metrics manual are given in the Appendix. The qualitative coders did not reveal the grounded data to the algorithm developer; this was so that the labeled corpora collected for supervised learning would not be manipulated.

Clinical	Non-clinical
Is the review related to a treatment procedure done by the physician?	Is the review related to the ease of appointment scheduling?
Is the review related to the physician's competence in the disease?	Is the review related to the general demeanor of the hospital staff?

Table 1: Sample of how reviews were coded for ground truth

Labeled corpora for supervised learning

Once the ground truth data was established and coded, we next set out to find corpora labeled as clinical and non-clinical. According to current literature, no classifier had segregated and labeled data into clinical and non-clinical data. Therefore, we curated the dataset by collecting and combining data from various verified sources.

For the non-clinical data labeling, we collected reviews on trips, hotels, airports, movies, books, and various general reviews on appointment scheduling at a lawyer's office and so on. The verification of these sources is listed in Table 2. By doing this, we were making sure that the corpora for their classifier consisted of plenty of verified reviews. Most importantly, this allowed us to structure words such as "great book," "great way of delivery," "hospitality was amazing" under non-clinical reviews.

The clinical data was harder to obtain because of compliance issues, and many hospitals are not willing to disclose private information. Additionally, medical data is tough to find due to HIPAA privacy regulations. However, we were able to obtain a publicly available dataset. This dataset offers a solution by providing medical transcription samples. It consisted of clinical review notes from the n2c2 NLP research data sets (Harvard medical school); the data was initially developed during the i2b2 project (Informatics for Integrating Biology & the Bedside - A National Center for Biomedical Computing) (Oleynik et al., 2019). This dataset contains sample medical transcriptions for various medical specialties and therefore allowed us to structure words such as "treatment was done in great detail," "diagnosis with relation to a heart condition was well done," as clinical reviews. A complete summary of the datasets collected is listed in Table 2.

Type of Data	**Source**	**Number of Reviews**	**Label**
Physician review dataset	Popular PRW Website	1614	Ground Labeling (Only known to the qualitative coders)
Labelled Corpora (Qualitative data)	Trip Advisor, Amazon reviews, Trivago, Wish reviews	5064	Non-clinical
Labelled Corpora (Qualitative data)	N2c2 NLP research data set	4121	Clinical

Table 2: Dataset Information

Experimental Model Development Process

In this sub-section, we will focus on how the model was developed. Figure 1 gives an overall picture of how the experimental design would look. In the following sub-section, we elaborate on the experimental model.

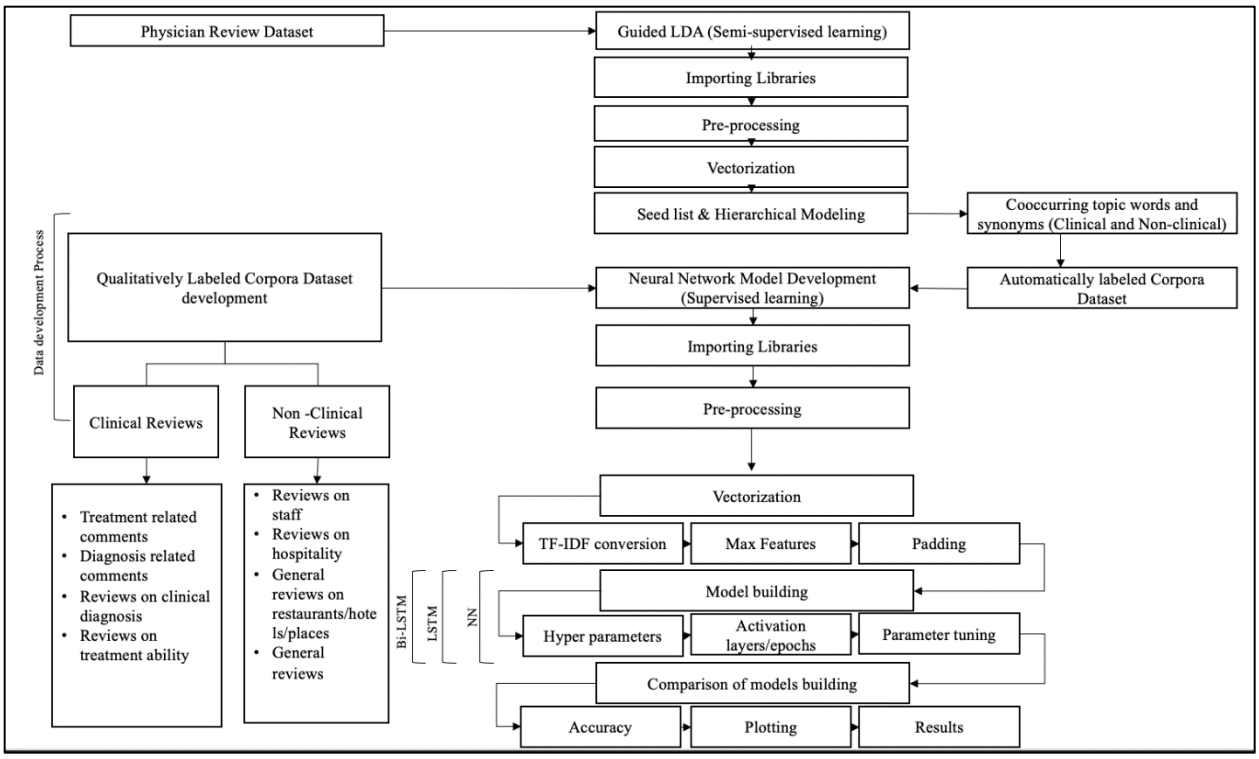


Figure 1: Experimental Model development process

Guided LDA

In figure 1, we see the experimental model development process. Once the data development phase was over, we used the physician rating dataset (Dataset 1) for the guided LDA development process. Guided LDA or SeededLDA implements latent Dirichlet allocation (LDA) using collapsed Gibb's sampling (Toubia et al., 2019). GuidedLDA can be guided by setting some seed words per topic, making topics converge in that direction. In our case, a good analysis of grounding the physician reviews was already done, and we had the metrics manual, which they used as seed words (Mansfield et al., 1999). The metrics manual (See Appendix) listed the seed words for all the physician reviews together. In table 3, we see the seed words and their topic labeling; these match with the metrics manual developed by us.

Clinical seed words	Non-clinical seed words
'knowledge','competance','correctness','diagnostic','ability','timely','referral','completeness', 'quality','cost','consciousness','testing','experience','responsible','systematic','correct','quality'	'environment','cleanliness','comfort','instrument','execution','treatment','procedure','reachability','punctuality','scheduling','waiting','time','notification','reachability','notification','appointment','teamwork','staff','monitoring','training','provisioning','comprehensiveness','social','skills','attentiveness','privacy','protection','shared','decision','communication','recommendation''satisfaction','efficiency','complication','follow-up'

Table 3: Seed words for the Guided LDA model from the metrics manual

Once the model was run with the seed list, the model results gave us new seed list words closely related to each given topic, as depicted in Figure 2. The pseudo algorithm for this model is given under code 1. Words with a high probability of co-occurring with the first iteration of seed-listed topics were listed as seed lists for the second iteration of the guided LDA model. The topic words were also compared with a synonyms dictionary to add synonyms for the next iteration. The iteration was carried out until the number of seed words for both topics was overlapping significantly. However, they individually could not establish a good relevance to the main initial topics. This was found to be that more than 48% overlap between the two topics could not confirm the uniqueness of the topics.

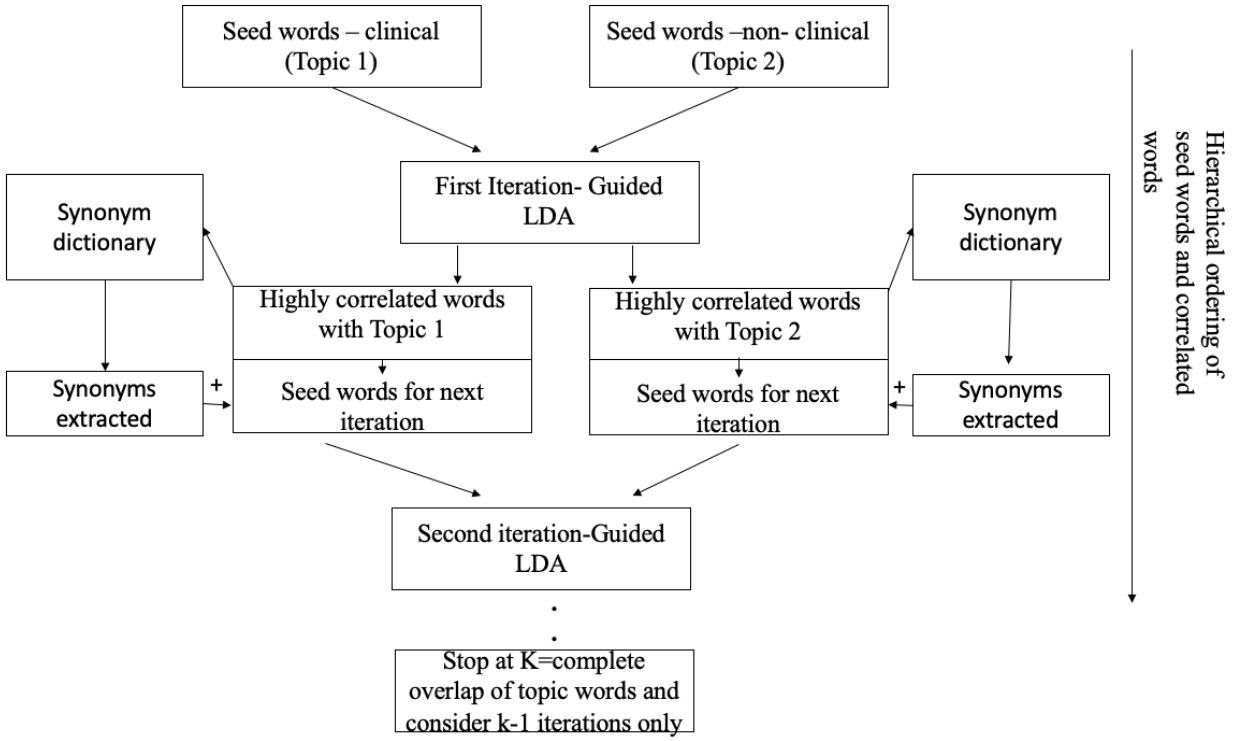


Figure 2: Hierarchical ordering of seed words for Guided LDA

```
For i in guided_LDA_topics:
        if word in topic 1:
        label_corpus=word ++
                For word in label_corpus :
                        find synonyms in synonym dictionary:
                        label_corpus=synonyms++
                        return;
```

Code 1: Pseudocode for Hierarchical corpus labeling using guided LDA

The words were now used as a part of modeling the classification dataset. The dataset now consisted of collected dataset S from (1) review websites, (2) medical transcripts from the n2c2 NLP research dataset (3) the seed words from the reviews' dataset, which was obtained from the guided LDA methodology. In the following sub-section, we demonstrate how the classifier was built.

Model Development process

Text pre-processing

Before we begin to develop the classifier, we need to pre-process the labeled corpora. Textual data such as reviews is different from numerical data, and such data is represented in human language and is not easy to directly convert into the quantitative format. In addition, processing raw text directly could be very noisy because some of the text content may not contain useful information. We used a natural language processing toolkit (NLTK) to process our data (Loper and Bird 2002). Detail processing steps are shown in Figure 3. Numbers, punctions, stop words were removed and then converted to lower case for uniformity. The texts were then stemmed and lemmatized.

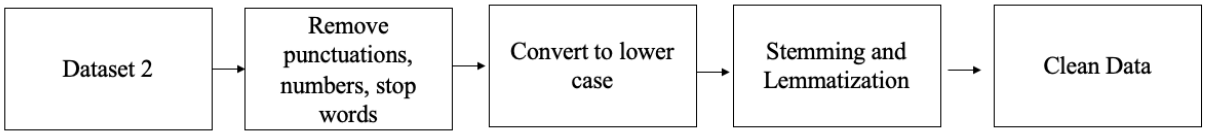


Figure 3: Text preprocessing

Vectorization

Since the machine had to work with an array of numbers rather than a set of strings while training the classification models, we converted the strings into TF-IDF format (text to numbers) and then extracted the max features. Following this, padding was done to extract meaningful features from the context of the corpora presented (Dwarampudi & Reddy, 2019). Out of the total corpora, the dataset was imbalanced to have 5064 clinical texts and 4121 non-clinical. Labeled corpora texts, each with close to 14,400 parameters to train once a max feature of 600 was selected. Since there was an imbalance of 5064 clinical texts vs. 4121 non-clinical texts, we balanced the texts to be 4121 each. Therefore, a total of 8282 reviews were used. We also added the list of words from the hierarchical guided LDA corpus to each review to enrich the comments with synonyms.

Model Building

We tested out three main models, including deep neural network learning models that followed supervised learning mechanisms. The models were (1) Deep Neural network, (2) Long short-term memory (LSTM) neural networks, and (3) Bidirectional -LSTM. We will describe each model building process separately and compare the three model classifiers in the results section.

Deep Neural network

We first built our classifier using a simple deep neural network model. Deep neural network represents machine learning when the system uses many layers of nodes to derive high-level functions from input information (Canziani et al., 2016). It means transforming the data into a more creative and abstract component. In our model, we make sequential calls for Keras's sequential model. Here, the deep neural network layers are added in a sequence (Liu et al., 2017). Our model's first layer, i.e., embedding layer, maps each word to an N-dimensional vector of real numbers. The embedding dimension is the size of this vector which is 16 in our case. The embedding layer indicates that the two words with similar meanings tend to have very close vectors. Because the embedding layer is the first hidden layer in our model network, we need to pass the shape of our input layer as defined by input length. The pooling layer that we added to our model helps reduce the number of parameters in the model hence helps to avoid overfitting. We have used average pooling here and converted the layer to 1 dimension. Next, we use a dense layer with activation function RELU followed by a dropout layer to avoid overfitting and a final output layer with a sigmoid activation function. As there are only two classes (clinical and non-clinical) to classify, we use only a single output neuron. The sigmoid activation function outputs probabilities between 0 and 1. The model summary provides the layer, shape, and number of parameters used in each layer.

Supervised Learning (LSTM)

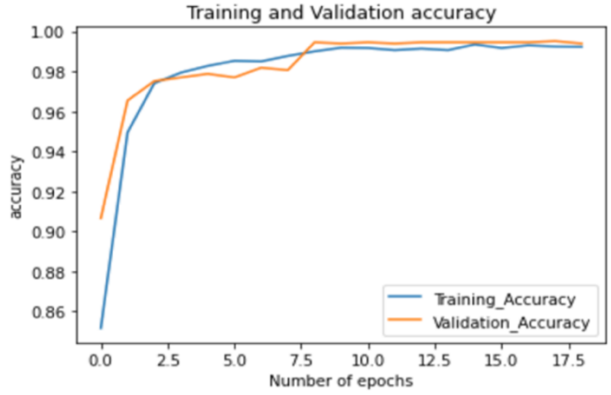
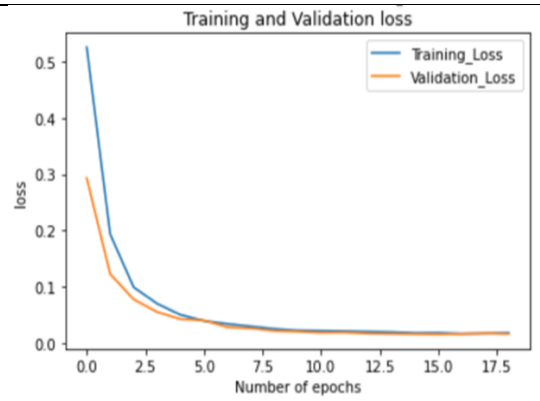
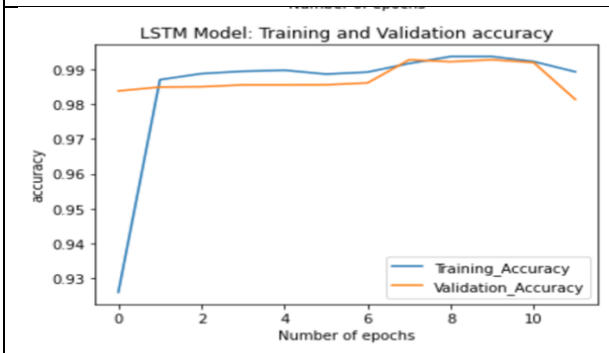
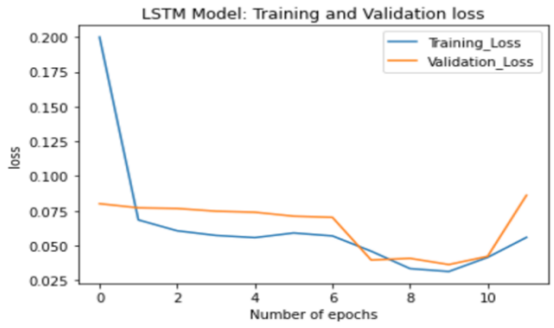
Next, we used the long short-term memory network (LSTM) to build the classifier. Long Short-term memory (LSTM) networks are a type of recurrent neural network capable of learning order dependence in sequence prediction problems. Recurrent neural networks are different from traditional feed-forward neural networks (Sundermeyer et al., 2012). his difference in the addition of complexity comes with the promise of new behaviors that the conventional methods cannot achieve. Recurrent neural networks have an internal state that can represent context information, the critical information about the past inputs for an amount of time that is not fixed a priori but depends on its weights and the input data. A recurrent neural network whose inputs are not fixed but rather constitute an input sequence can be used to transform an input sequence into an output sequence while taking into account contextual information in a flexible way (Lipton et al., 2015). The very reason for selecting a long-short term memory neural network is because our goal is to build a classifier that adapts to new contexts. Choosing an LSTM model helped us to additionally account for context adaptability. We fit the detection model using LSTM. Some new hyper-parameters used in LSTM were *the* number of nodes in the hidden layers, which we chose to be as 20 within the LSTM cell, and also the true value set for return sequences ensures that the LSTM cell returns all of the outputs from the unrolled LSTM cell through time. If this argument is not used, the LSTM cell will provide the result of the LSTM cell from the previous step.

Supervised Learning (Bi-LSTM)

Finally, we used the bidirectional LSTM (Bi-LSTM) networks to build our classifier. Unlike in LSTM, the Bi-LSTM learns patterns from both before and after a given token within a document (Zhang et al., 2020). The Bi-LSTM back-propagates in both backward and forward directions in time. Due to this, the computational time was increased compared to LSTM. However, in most cases, Bi-LSTM was said to result in better accuracy. We expected the model to perform well overall through the Bi-LSTM model because building this model helped us further investigate whether adding backward and forward propagation towards our curated datasets helped improve its context adaptability (Mughees et al., 2021).

Results

The results of the three models can be seen in Table 4. The Dense detection model reported an accuracy of 99.39%, the LSTM network model reported accuracy of 98.14%, and the Bi-LSTM network reported accuracy of 98.54%. Table 4 showed us that the validation accuracy was higher than the training accuracy for the dense neural network which pointed that the model worked well with new data (Table 4 1a). However, the validation accuracy was comparatively lower than the training accuracy for the LSTM and Bi-LSTM network models.

Accuracy	Loss
Dense Neural Network 1a	Dense Neural Network 1b
	
LSTM Network 2a	LSTM Network 2b
	
Bi-LSTM Network 3a	Bi-LSTM Network 3b

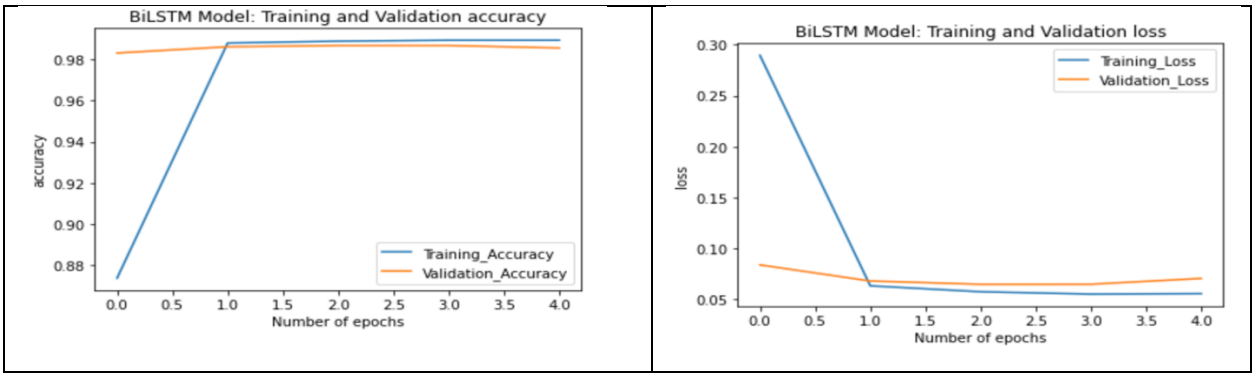


Table 4: Comparisons of the results of the three models

Validation and comparison of the models with the ground truth

In this subsection, we point out that although the accuracy of the dense neural network was higher when it came to classifying the reviews into either clinical or non-clinical based on the context, the Bi-LSTM model performed well. In table 5, we compare the model predictions with the ground truth and notice that in both clinical and non-clinical classification, the Bi-LSTM model does the best job at classifying the reviews. In review 1, which was non-clinical and purely about the expense, the neural network model predicted a relatively safe value. In contrast, the Bi-LSTM model predicted it as non-clinical (0 being clinical and 1 being non-clinical, the prediction value was continuous). Although the LSTM model does predict review 1 to be non-clinical, in review 2, we notice that LSTM classifies a clinical review comment to be non-clinical, this is because LSTM looks only at forward propagation and places more importance on the future words in the comment which point towards the comment being more non-clinical ('saved her life, he was good at explaining things, and answering questions), whereas the Bi-LSTM model looks at words like septic, explaining and purely classifies the review comment to be clinical.

Sl No.	Review	Model type	Model prediction 'clinical': 0, 'nonclinical': 1	Ground Truth classification
1.	"Expensive doctor, too much bills to pay"	Deep Neural network	0.47	Non-clinical
	"Expensive doctor, too much bills to pay"	LSTM	0.97	Non-clinical
	"Expensive doctor, too much bills to pay"	Bi-LSTM	0.98	Non-clinical

2.	"Dr Ahmed took care of my mom when rushed to the hosp cuz she was septic and had no pressure.. He figured out what was wrong with her and saved her life. He was good at explaining things and answering questions I had. "	Deep neural network	0.12	Clinical
	"Dr Ahmed took care of my mom when rushed to the hosp cuz she was septic and had no pressure. He figured out what was wrong with her and saved her life. He was good at explaining things and answering questions I had. "	LSTM	0.97	Clinical
	"Dr Ahmed took care of my mom when rushed to the hosp cuz she was septic and had no pressure.. He figured out what was wrong with her and saved her life. He was good at explaining things and answering questions I had. "	Bi-LSTM	0.009	Clinical

Table 5: Comparisons of the review comments with the ground truth.

The Bi-LSTM was the best in our case because LSTM in its core preserves information from inputs that have already passed through it using the hidden state (Mohan & Gaitonde, 2018). Unidirectional LSTM only holds the past information because the only inputs it has seen are from the past. Using bidirectional will run inputs in two ways, one from past to future and one from future to past, and what differs this approach from unidirectional is that in the LSTM that runs backward preserving information from the **future**, and using the two hidden states combined, you are able in any point in time to preserve information from **both past and future**.

Limitations

When we ran the whole Bi-LSTM model to predict the classification for the 1,614 review comments, we noticed that the total number of matches between the Bi-LSTM predicted

classification vs. the ground truth was 1,311, therefore leaving a total of 303 misclassified reviews. However, further analysis let us realize that there was a threshold from 0.46 to 0.62 (remember that 0 to 1 was the continuous predictor variable where a prediction closer to 0 was clinical and 1 was non-clinical) in the classification predictor, which led classified comments to have an equal number of non-clinical as well as clinical comments. In other, words there were about 604 comments out of 1614 comments which appeared in a threshold after examination that they contained equal distribution of context between clinical and non-clinical terminologies. We realize that further analysis of this dataset could be done where the review comments lying between this threshold could further split into other classes/categories. This is one of the limitations of our paper. In the future, we plan on bringing out theme analysis of the reviews using the hierarchical guided latent Dirichlet allocation algorithm we proposed earlier to give rise to more topics and themes and then create labels based on those themes. Future work would also consist automatic creation of multi-classes for the classifier.

Conclusion

To conclude, our paper focuses on the methodology in which the dataset can be curated and makes full use of the semi-supervised algorithm to label classes and further use that dataset to build a classifier that can be applied to a new context. The critical aspect of this paper is to introduce a novel approach to solving classification problems. Furthermore, adding a ground truth also helps validate our dataset. Additionally, our article also compared three supervised classification models and informed us that the Bi-directional LSTM performs best when a new context-based dataset is created. Future work that our research proposes is to use the hierarchical latent Dirichlet allocation algorithm demonstrated in the paper, develop thematic extraction of reviews, and build a classifier that can automatically recognize themes and multi-class labeling.

APPENDIX

Metrics for Manual Coding

All metrics under technical or medical sub-section were coded as clinical, and the remaining were mapped as non-clinical (Rothenfluh, F., & Schulz, P. J., 2018)

Dimension and indicators
Structure
Infrastructure-NonClinical
Office environment, cleanliness, comfort
Instruments in the practice to make the diagnosis or execute the treatment

Dimension and indicators
Reachability of the practice by car or public transport
Organization- NonClinical
Punctuality, wait time in practice
Scheduling or making appointments
Waiting time until the next appointment
Reachability of the practice via phone
Notification of patients in case of appointment delays or cancellations
Teamwork between physician and his team
Number of staff present in the practice to welcome and take care of patients
Staff -NonClinical
Staff friendliness and courteousness
Staff experience and training
Process
Interpersonal- NonClinical
Comprehensiveness and completeness of information provision
Social skills of the doctor (attentiveness, helpfulness, empathy)
Amount of time spent with the patient
Friendliness of the physician
Physician's (active) listening skills
Conversation climate with the doctor
Trust in physician

Dimension and indicators	
	Confidentiality, protection of privacy
	Information provision about how to handle the illness or disease
	Shared decision about the course of action together with the patient or shared decision making
	Doctor's effort to engage the patient in shared decision making
	Physician's skill to assess the patient's handicaps and presentation with appropriate information and treatment options
	Communication and narration during the treatment execution
Technical or medical-Clinical	
	Physician's knowledge
	Physician's competence
	Correctness of the diagnosis, diagnostic ability of the physician
	Improvement of the patient's health status
	Timely referral to a specialist or the hospital if needed
	Completeness and quality of anamnesis
	Quality and variety of treatment suggestions
	Cost consciousness of the physician when making tests or giving out medications
	Physician's experience
	Responsible medication prescription
	Systematic proceeding of physician to reach the correct diagnosis
	Timeliness or promptness of the diagnosis and initiation of the treatment
	Correctness of treatment execution by the physician and his team

Dimension and indicators
Quality of the information provided to the patient
Physician's competence to execute the treatment competently
Outcome- NonClinical
Likelihood of recommendation
Satisfaction with the doctor
Presence and quality of the follow-up care
Efficiency of the treatment or cost-benefit ratio
Price of the treatment
Cost coverage by the health insurance
Patient's increase in knowledge about his disease or injury
Number or kind of complications[a]
Patient loyalty or patient's intention to return for future or follow-up treatments[a]
Summative and other- NonClinical
Summative or overall score
Other organization scores
Other interpersonal scores
Other overall scores
Other technical scores

REFERENCES

Agarwal, B., Poria, S., Mittal, N., Gelbukh, A., & Hussain, A. (2015). Concept-level sentiment analysis with dependency-based semantic parsing: a novel approach. *Cognitive Computation*, *7*(4), 487-499.

Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. Proceedings of the fifth annual workshop on Computational learning theory,

Breiman, L. (2001). Random Forest, vol. 45. *Mach Learn*, *1*.

Canziani, A., Paszke, A., & Culurciello, E. (2016). An analysis of deep neural network models for practical applications. *arXiv preprint arXiv:1605.07678*.

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning,

Dwarampudi, M., & Reddy, N. (2019). Effects of padding on LSTMs and CNNs. *arXiv preprint arXiv:1903.07288*.

Figueiredo, M. A. T., & Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on pattern analysis and machine intelligence*, *24*(3), 381-396.

Hao, H., & Zhang, K. (2016). The voice of Chinese health consumers: a text mining approach to web-based physician reviews. *Journal of medical Internet research*, *18*(5), e108.

Hao, H., Zhang, K., Wang, W., & Gao, G. (2017). A tale of two countries: International comparison of online doctor reviews between China and the United States. *International Journal of Medical Informatics*, *99*, 37-44.

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining,

Li, P., Zhu, Q., & Zhang, W. (2011). A dependency tree based approach for sentence-level sentiment classification. 2011 12th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing,

Lipton, Z. C., Kale, D. C., Elkan, C., & Wetzel, R. (2015). Learning to diagnose with LSTM recurrent neural networks. *arXiv preprint arXiv:1511.03677*.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11-26.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Mansfield, J. R., McIntosh, L. M., Crowson, A. N., Mantsch, H. H., & Jackson, M. (1999). LDA-guided search engine for the nonsubjective analysis of infrared microscopic maps. *Applied spectroscopy*, *53*(11), 1323-1330.

Matsumoto, S., Takamura, H., & Okumura, M. (2005). Sentiment classification using word sub-sequences and dependency sub-trees. Pacific-Asia conference on knowledge discovery and data mining,

Mohan, A. T., & Gaitonde, D. V. (2018). A deep learning based approach to reduced order modeling for turbulent flow control using LSTM neural networks. *arXiv preprint arXiv:1804.09269*.

Mughees, N., Mohsin, S. A., Mughees, A., & Mughees, A. (2021). Deep sequence to sequence Bi-LSTM neural networks for day-ahead peak load forecasting. *Expert Systems with Applications*, *175*, 114844.

Oleynik, M., Kugic, A., Kasáč, Z., & Kreuzthaler, M. (2019). Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. *Journal of the American Medical Informatics Association*, *26*(11), 1247-1254.

Popescu, A.-M., & Etzioni, O. (2007). Extracting product features and opinions from reviews. In *Natural language processing and text mining* (pp. 9-28). Springer.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., Shyu, M.-L., Chen, S.-C., & Iyengar, S. S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, *51*(5), 1-36.

Rivas, R., Montazeri, N., Le, N. X., & Hristidis, V. (2018). Automatic classification of online doctor reviews: evaluation of text classifier algorithms. *Journal of medical Internet research*, *20*(11), e11141.

Rothenfluh, F., & Schulz, P. J. (2018). Content, quality, and assessment tools of physician-rating websites in 12 countries: quantitative analysis. *Journal of medical Internet research*, *20*(6), e9105.

Shrestha, A., & Mahmood, A. (2019). Review of deep learning algorithms and architectures. *IEEE Access*, *7*, 53040-53065.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. Thirteenth annual conference of the international speech communication association,

Tavakol, M., Torabi, S., & Akbar Zeinaloo, A. (2006). Grounded theory in medical education research. *Medical Education Online*, *11*(1), 4607.

Toubia, O., Iyengar, G., Bunnell, R., & Lemaire, A. (2019). Extracting features of entertainment products: A guided latent dirichlet allocation approach informed by the psychology of media consumption. *Journal of Marketing Research*, *56*(1), 18-36.

Wallace, B. C., Paul, M. J., Sarkar, U., Trikalinos, T. A., & Dredze, M. (2014). A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews. *Journal of the American Medical Informatics Association*, *21*(6), 1098-1103.

Wawer, A. (2015). Towards domain-independent opinion target extraction. 2015 IEEE International Conference on Data Mining Workshop (ICDMW),

Zhang, B., Zhang, H., Zhao, G., & Lian, J. (2020). Constructing a PM2. 5 concentration prediction model by combining auto-encoder with Bi-LSTM neural networks. *Environmental Modelling & Software*, *124*, 104600.

Zhang, Y., & Wallace, B. (2015). A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*.

Zhu, X., & Goldberg, A. B. (2009). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130.

Zhuo, L., Zheng, J., Li, X., Wang, F., Ai, B., & Qian, J. (2008). A genetic algorithm based wrapper feature selection method for classification of hyperspectral images using support vector machine. Geoinformatics 2008 and Joint Conference on GIS and Built Environment: Classification of Remote Sensing Images,