International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Medical Sentiment Analysis Using Social Media

[1]Spandan Das, [2]Aditya Sharma, [3]Madhavan P

ABSTRACT--*With the increase in the popularity of Internet, many people have created social forums to share and obtain information about medical issues concerning them or their loved ones. These forums contain experiences of people who have dealt with or are dealing with health related problems and may present some insight regarding medicines, diagnosis and treatment or interact with people with similar conditions on the forums. Observing social media forums has excited medical natural language processing analysts to uncover various cathartic anomalies. In this paper, we present a benchmark setup to analyze the sentiment with respect to people's medical conditions considering the information, especially the information available on social medical forums. To pursue this, we requested the data from IIT Patna, who in turn collected the data from the website 'patient.info'. The focus is given on the identification of multiple forms of medical sentiments which can be assessed from people's medical condition and medication.*

Keywords--*Medical Forums, Medical Sentiment, Natural Language Processing, LSTM, Global Max-Pooling.*

I. INTRODUCTION

Sentiment Analysis has been drawing quite a lot of attention recently owing to the growing popularity of social media. The astounding rise in the trend of blogging is noticed in the health communities like the health forums which are filled with millions of people who are looking for health related solutions and support regarding their conditions, seeking advice from professionals online.

Applications of medical sentiment analysis include assessment of clinical records to aid the health care professionals by implementing an automated decision support system.

On the basis of a study by Pew Internet and American Life Project, 80% of online users in USA have explored health related queries on the internet. Moreover, it was found that about 63 percent of users search for information about definitive medical issues on the internet. Also, about 47 percent of people search for medical treatments. Analyzing the huge amount of texts on the medical forums by understanding the sentiments is beneficial as every human forms opinion about any human pursuit. Literature surveys have shown that in recent times medical sentiment analysis is a topic of increasing research importance. Especially classifying the effectiveness of certain drugs and the progress of recovery for patients. In this work, we have looked at how sentiment analysis from online medical forums can be useful to build a patient assisted healthcare system. We have proposed a novel approach to

[1] *Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu*
[2] *Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu*
[3] *Department of CSE, SRM Institute of Science and Technology, Kattankulathur, Tamil Nadu*

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

classify sentiments based on the texts present in the corpus. Our approach explores a deep Bidirectional LSTM network in an attempt to further both, the research and the classification model developed by (Shweta Yadav et al.) in their journal, titled "Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System" in the Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). They had used a Convolutional Neural Network model which established a benchmark for the classification. We try to explore a different approach to the classification problem by leveraging the power of Long Short Term Memory Units (LSTM). Our research focuses on improving the accuracy of classification, to be able to generalize the model better. Instead of using the n-gram model proposed by (Shweta Yadav et al.) 2018, we will tokenize the sentences into word vectors which are converted into a sequence of integer values. These encoded text data are trained on the Word2Vec embedding. Since the dataset has an uneven distribution of the labels, we try to achieve a good F1-Score. We set a benchmark for accuracy in this paper.

II. RELATED WORKS

1.1 Emotion Classification

Sokolova and Bobicev (2013) researched on various emotions from medical web documents. They analyzed the classes of emotions such as optimism (hope, happiness), bewilderment (worry, doubt and concern), appreciation, facts, and facts + optimism. They used naive Bayes classifier with the features derived from lexicon Word-NetAffect. Study conducted by Melziet et al. (2014) on categorization of emotion learned SVM using the feature set containing BoWs, n-grams and specific attributes. Size of the advertisement is also huge because of the addition of hash value in the route entry.

1.2 Medical Sentiment Analysis Using Social Media

In the works presented by Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya belonging to IIT Patna, a benchmark setup for analyzing medical sentiment on information acquired from patient.info has been presented. They have annotated the data based on sentiment analysis. They have used a novel approach of feature mapping and convolutional neural networks to get the benchmark results.

1.3 Adverse Drug Relation

Non-medical social media forums like twitter (Nikfarjam et al., 2015) have been exploited to capture adverse drug effects. Some prominent studies (Leaman et al., 2010; Yates and Goharian, 2013) exploit existing lexicons to identify Adverse Drug Reactions mentioned in user posts. A few of the other popular studies include the works of (Na et al., 2012; Sharif et al., 2014) utilizing machine learning based NLP techniques to identify the ADR.

III. CMS: CORPORA FOR MEDICAL SENTIMENT

Due to the increasing interest of people's self-stated medical posts, the researchers from IIT, Patna went through many online medical forums. They finally settled for 'patient.info' medical forum. 2302 posts were found which contained information about medication. These posts were dated between 25th September 2016 and 15th November 2016. To safeguard user privacy, all user details were deleted.

(Shweta Yadav et al.) 2018 classified the data as **Effective, Ineffective and Serious adverse effects.**

The following points were considered while selecting the source of information for the corpus:

- The site should be extremely reliable and popular with reasonable number of users.
- There should be a fair number of opinions and active discussions pertaining to medication.

Keeping in mind these points, the dataset was created and was also used for the proposed model in this paper. Posts from popular discussion groups such as *Anxiety*, *Depression*, *Asthma* and *Allergy* were selected. The dataset is classified as follows:

Effective: User provides positive sentiment in the form of effectiveness of the treatment or drug.

Ineffective: There is no benefit from the medication or treatment.

Serious adverse effect: The negative comments towards the treatment mainly in the form of adverse drug effect. From the data, we analyze that sentiment in clinical accounts can't always be exhibited in single terms or phrases, instead it mainly depends on the context. The idea of medical sentiment is very convoluted and has many aspects making it very intriguing, but also challenging for automated analysis.

The Data Statistics and Annotation Schemes are presented in the tables and graphs below. As can be seen from the graphs words like *anxiety*, *feeling*, *work*, *help*, etc. occur a lot in the posts. This gives us a general idea about the blog posts and what is prevalent in these posts. Some of the words like *mg*, *xa*, *still*, etc. don't seem to be relevant and thus need to be removed.

Medical Blog	Label
I went off my Lexapro because I was getting bad tremor side effects yesterday. Now I feel I'm getting my same old symptoms. I'll be doing something and my brain will be working properly I'm just afraid I'll forget where I am because everything looks so unfamiliar.	**Serious Adverse Effect**
I have been suffering with General anxiety disorder, panic attacks and depression for a good few years now and it's a battle day to day,, but at the moment I'm feeling quite positive. I'm willing to fight this illness and not run away from it. I'm on citalopram at the moment but don't want to just depend on tablets - I'm also having therapy. Has anyone tried yoga? Mediation? Reiki? Anything along those lines? Or any tips and tricks to remain positive will be much appreciated. Thank you.	**Ineffective**
I've had anxiety for the majority of my life, and now my university degree, is offering a once in a lifetime trip to South Africa - which means a 12hr flight. Considering I will be nervous for the long distance travel without my family,, I have also never been on a plane before..can anyone give me some advice on what i should do? Will it be too much? Would it make me any worse? I am really struggling on how to feel about it and how i should go about it because I have been making progress.	**Effective**

Figure: *Description of Annotation Scheme*

Medication				
Effective	Ineffective	Serious Adverse Effect	Avg # of Sentences	Avg # of Words
462	613	1,226	9	176

Figure: *Data Statistics for Classification*

International Journal of Psychosocial Rehabilitation, Vol. 24, Issue 08, 2020
ISSN: 1475-7192

Figure: *Word-Cloud Representation of the Corpus*

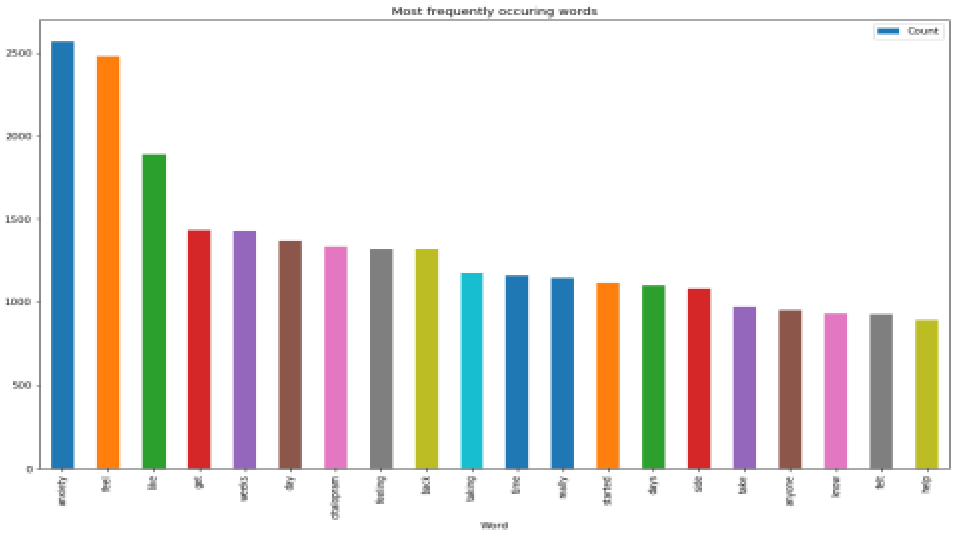

Figure: *Most Frequently Occurring Words in the Corpora*

IV. DATA PREPARATION

The corpus consists of punctuation marks like "*!?*&…*" and words like *am*, *the*, *in*, *I*, etc. which adds little to no value in our use case and thus need to be removed. We utilize stopwords from NLTK and add a few other stop words to the dictionary and remove them from the corpus. Similarly, punctuation marks are also removed. Further, to optimize space and further preprocess similar words, the text is stemmed using Snowball Stemmer. In order to convert the sentences into words, we tokenize each blog post. The tokenized words are then encoded. Each blog post is then converted into a sequence of the same length and thus, the data is prepared to be fed as input to the model. Finally, the corpus is divided into training and testing set.

V. THE NETWORK FOR SENTIMENT ANALYSIS

In this section we propose a method based on a deep Bidirectional LSTM network that extracts the sentiments from the corpus. As presented in the Figure below, the input is fed to an Embedding layer which learns the word embedding. The input fed is a complete blog post in the form of tokens/word embedding. We use global max-pooling over the whole blog post to obtain the global features. This pooled feature is fed to a fully connected

Received: 21 Dec 2019 | Revised: 18 Jan 2020 | Accepted: 05 Feb 2020

neural network. In the output layer we use the softmax classifier to classify the post into 3 output classes. We describe the layers of our proposed model in detail below:

Input Layer**:** Each blog post is tokenized and fed as the input to the model.

Word Embedding **Layer:** This layer turns positive integers into dense vectors of fixed size. Each word in the blog post is looked up in the corresponding word embedding matrix. Word2Vec Embedding has been used in the model. The Embedding layer is essentially a dictionary which maps integer indices to dense vectors. It takes integers as input and looks these up in an internal dictionary and then returns the associated vectors. We have set the weights equivalent to the Word2Vec embedding.

LSTM Layer: The word embedding is fed as the input to the Bidirectional LSTM Layer. LSTM is a type of recurrent neural network which uses different gates to control the memorizing process. LSTM units comprise of various components such as the sigmoid layer, tanh, etc. Sigmoid and tanh play very important roles. The sigmoid layer takes an input and decides which part of the old output should be removed. It decides which part of the new information should be updated or ignored. It can output in the range of 0 to 1. The tanh layer helps in resolving the vanishing gradient problem faced in standard recurrent neural networks. It creates a vector of all the possible values from the new input. In our model, we have used a Bidirectional-LSTM layer to improve the learning process and optimize the speed of the model. The input fed to the Bidirectional-LSTM layer are the word embedding derived from the Word2Vec model.

Dropout Layer: Dropout is proposed by (Hinton et al.) as a regularizer which randomly sets half of the activations to the fully connected layers to zero. It prevents over fitting and results in better generalization of the model.

Global Max-Pooling Layer: This layer outputs 1 response for every feature map. It outputs the maximum feature from the pool.

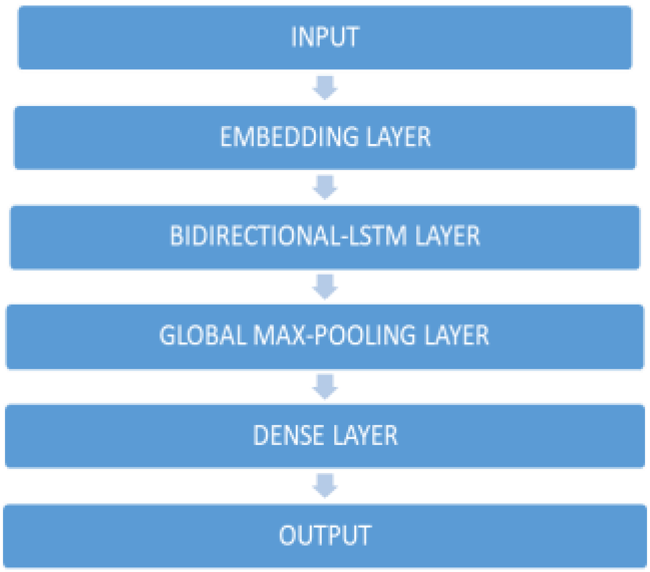

Figure: *Our model for predicting the medical sentiment*

Figure: *Classification Report (1)*

	Classification Models	Precision	Recall	F1-Score
Baselines	SVM	0.74	0.76	0.75
	Random Forest	0.72	0.72	0.72
	MLP	0.74	0.75	0.74
	CNN	0.86	0.77	0.82
	Bidirectional LSTM	0.86	0.86	0.86

Figure: *Classification Report (2)*

VI. CONCLUSION

In this paper we have presented a deep Bidirectional LSTM Neural Network based classification framework to predict the possible medical sentiment category for medication. We are able to obtain significant performance improvements over the baseline models and a higher F1-Score and Recall than the CNN model implemented by (Shweta Yadav et al.) and have succeeded in establishing a benchmark for medical sentiment analysis using social media. In the future, we seek to retrieve drug names from free text and perform sentiment analysis on the blog posts to determine the effectiveness of the respective drugs for the different kinds of medical conditions.

REFERENCES

1. Breiman, L. (2001). Random forests. Machine Learning, 45(1):5–32, Oct.
2. Cortes, C. and Vapnik, V. (1995). Support vector machine. Machine learning, 20(3):273–297.
3. Denecke, K. and Deng, Y. (2015). Sentiment analysis in medical settings: New opportunities and challenges. Artificial intelligence in medicine, 64:17–27.
4. Sharif, H., Abbasi, A., Zafar, F., Zimbra, D. (2014). Detecting adverse drug reactions using a sentiment classification framework. ASE international conference on social computing.
5. Sokolova, M. and Bobicev, V. (2013). What sentiments can be found in medical forums? In RANLP, volume 2013, pages 633–639.

6. Yates, A. and Goharian, N. (2013). Adrtrace: detecting expected and unexpected adverse drug reactions from user reviews on social media sites. In European Conference on Information Retrieval, pages 816–819. Springer.
7. Shweta Yadav, Asif Ekbal, Sriparna Saha, Pushpak Bhattacharyya. (2018). "Medical Sentiment Analysis using Social Media: Towards building a Patient Assisted System". Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).