International Journal of Research in Advent Technology, Vol.7, No.1, January 2019
E-ISSN: 2321-9637
Available online at www.ijrat.org

Retrieval of Disease-Treatment Information from Online Forums Using NLP and Convoluted Neural Network Pipeline

Mamatha Balipa[1], Dr. Balasubramani R[2]
Department of MCA[1], Department of Information Science & Engineering[2], NMAM Institute of Technology, Nitte
Email: mamathabalipa@nitte.edu.in[1], balasubramani.r@nitte.edu.in[2]

Abstract—This paper describes how the authors have applied Convoluted Neural Network to classify messages extracted from online healthcare forums as solutions or non-solutions for the disease Psoriasis. In this work, the comments gathered from online healthcare forums regarding the disease Psoriasis are classified as either giving information about the treatments that have worked and treatments that have not worked or comments that pose more questions. To do this a combination of NLP and Convoluted Neural Network is used.

Index Terms—Keywords: Convoluted Neural Network, NLP

1. INTRODUCTION

The disease Psoriasis is an auto immune disease. The solution for the disease is very rare. Many people have tried various medications and treatments. The treatments people have undergone are Allopathic, Homeopathy, Ayurveda, etc., People who have undergone different types of treatments have shared their experiences online on the web on various health-care forums. People who are interested in finding different types of treatments for the disease and their success rates online, have to wade through different types of comments. This work extracts the messages about successful treatments for the disease Psoriasis posted on online health forums. This work by no means recommends any medication or treatment. The work only sifts through the different types of treatments undergone by people for the disease and gives a consolidated output containing only the comments depicting the treatments that have worked or whose success rate is high. The messages or comments posted on healthcare forums are in English, hence to analyse the text, Natural Language Processing techniques are applied. To classify the comments as solutions or non-solutions, ma-chine learning algorithms like Naive Bayes, Decision Trees, Support Vector Machine, and Logistic Regression were ex-plored. Nave Bayes provided an accuracy of 94/%, Decision Tree provided an accuracy of 88%, SVM provided an accuracy of 88%. In the current work carried out, Convoluted Neural Network was used to categorize the comments and a mean score of 84% has been obtained.

2. LITERATURE

When we think regarding Convolutional Neural Network (CNNs), we tend to generally consider Computer Vision. The use of CNNs were mostly used for major research in Image Classification and play a major role in Computer Vision systems. Nowadays, CNNs are used from Facebooks automatic icon tagging to self-driving cars. More recently researchers have begun to use CNNs to issues in Natural Language Processing and achieved good results.

2.1. Convolution Neural Networks:

CNN is a fully connected layer. Convolution is similar to a sliding window task that works on a matrix. The algorithm has multiple layers having convolutions which use activation functions that are non linear in nature. The functions may be ReLU or tanh. These functions are used on the output. The old neural network tends to connect every neuron in the input layer to every neuron got as output in the subsequent layer. Convolutions are used by CNNs on the layer which is the input to generate the result. Thus there will be local connections where every region in the input layer will be connected to a neuron in the output layer.There will be various filters applied in each layer to get the results. Based on the job to be performed, CNN automatically gets the values of its filters during the training phase.

2.2. CNN and NLP:

The text is uasually represented as a matrix to be used in NLP jobs. Each row in the matrix represent a word or a token.It could be character too.

Thus each row in the matrix is a word embedding. They may be word2vec or GloVe. The entire row belonging to the matrix is used by the filters. The matrix width will be the same as the width of the filters. The height will be normally two to five words. In the above figure, we show three filter region of sizes two, three and four. Each region has two filters. Convolutions are performed on the matrix of sentences by the filters and feature maps of different length are generated. Each map undergoes a max pooling. Thus all six maps are used to generate a feature vector. The resultant features are concatenated to create a feature vector for the final layer. In the final stage, the feature vector is received as input by the softmax layer to classify the sentence. [14]. CNNs are very fast. Convolutions are important in computer graphics. They are implemented by GPUs. Computing 3-grams and above in a large vocabulary is quite expensive. Convolutions filters automatically learn the vocabulary representations without the need to represent the whole vocabulary.

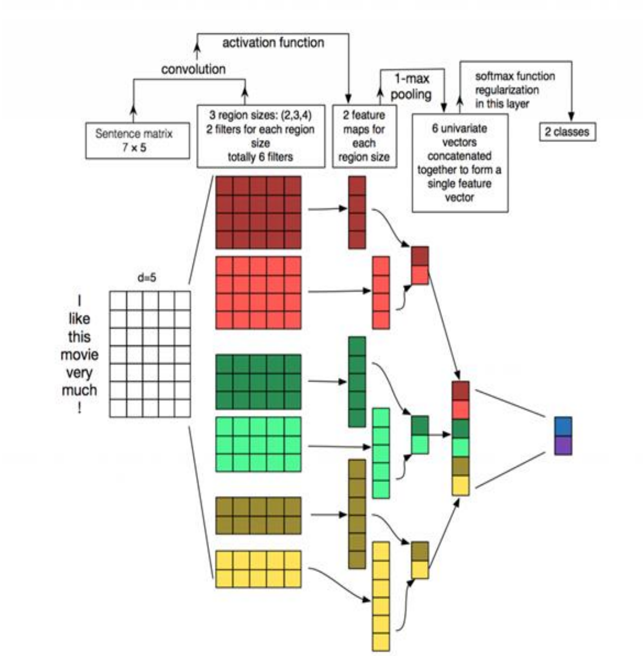

activation function
convolution
1-max pooling
softmax function regularization in this layer
Sentence matrix 7 × 5
3 region sizes: (2,3,4) 2 filters for each region size totally 6 filters
2 feature maps for each region size
6 univariate vectors concatenated together to form a single feature vector
2 classes
d=5
I
like
this
movie
very
much
!

Fig. 1. A Convoluted Neural Network [9]

2.3. CNN Hyperparameters

2.3.1. Narrow vs. Wide convolution:

You can use zero-padding also called wide convolution, if there are no neighbouring elements when the filter is applied to the first element in the matrix. So you can get a larger output. If its not zero padding, then it is narrow padding.

2.3.2. Size of the Stride:

Stride size is by what amount the filter should be shifted at every step. A stride size of 1 is used most of the time. A bigger stride size will create a model similar to Recursive Neural Network which is found to be more efficient than CNN in tasks related to NLP.

2.3.3. Pooling Layers:

After the convolution layers are used, the pooling layers will be used. Pooling layers usually do subsampling. In pooling most of the time a max function is applied to the result of each filter. Pooling can be done over a window too. Pooling provides fixed size output matrix which can be used for classification. So variable size sentences and filters can be used but the dimensions shoukd be the same to be fed into the classifier.

2.3.4. Channels:

Channels are different ways in which you look at the input data. Channels can be either different types of word embeddings or same sentence represented in a different language and so on.

2.4. Convolutional Neural Networks applied to NLP

Convolution Neural Network for Natural Language processing are used for classification, spam detection, Sentiment Analysis, Topic Categorization, etc.,. Pooling and convolutions lose some information about words. So operations like entity recognition and POS tagging are a bit difficult. Some of the applications where CNNs are used in NLP are mentioned below.[8] Kim, Y in his work performs sentiment analysis and topic categorization jobs using CNN architecture to evaluate different datasets. The CNN architecture works very efficiently across datasets. A sentence in the form of word2Vec is given to the input layer. A multi filtered convolutional layer is applied, which is followed by a max-pooling layer ending with a softmax classifier. A complex but similar architecture was earlier proposed by Kalchbrenner et al [7]. [13] Wang et al, added another layer to the architecture that implements semantic clustering. [5] Johnson, R. and Zhang in their work, do not use word2Vec or GloVe and train the convoluted neural network from the begining. They apply convolutions straightaway on one-hot vectors. They use as input, bag of word representation of the sentences. [6] Johnson, R., and Zhang, T, in their work use region embedding to enhance the model trained with a CNN to predict the context of the region. [11] Nguyen, T. H., and Grishman, R. use CNNs for classifying and extracting relations between text. [2][12] Nguyen, T. H., and Grishman, R. et al and Shen. Y et al have used CNN for information retrieval.

3. ABOUT THE WORK

In this work, messages extracted from health forums are extracted and classified as treatments that have worked and comments that do not mention successful treatments. So the issue is to group a remark as an answer for the sickness Psoriasis or as not an answer. In this work, Convoluted Neural Network algorithm is applied for classification.

3.1. Information retrieval:

Information retrieval was performed using crawlers developed by the authors to extract messages from sources like psoriasis-association.org.uk, healingwell.com, MedHelp.org [4] and HealthBoards.com [4]. The search engine to do the same was developed using JSoup API[3], a Java HTML parser library and Apache Lucene[10]. About 2000 posts were collected from psoriasis-association.org.uk, healingwell.com, MedHelp.org [4] and HealthBoards.com [4]. The comments are first divided into training and test text. The training text are manually labelled as belonging to ’solution’ and ’non-solution’ classes. The comments labelled as ’solution’ are comments that specify a treatment that has worked. Comments labelled as ’non-solution’ are those comments that may either be another query regarding the disease or a discussion about a treatment or medicine that has not worked. Healingwell.com, MedHelp.org [4] and HealthBoards.com[4] are health forums having multiple threads discussing issues regarding Psoriasis. Users discuss treatments they have undergone that have not worked, treatments that have worked, post questions, food that aggravate the symptoms or are the cause for the disease, food that give relief from the symptoms and all the issues pertaining to the disease. A post may have a single sentence or more than one sentence. The messages are free flow of text in English, hence the text needs to be transformed into a form which can be processed.

Since the data used is extracted from online healthcare forums, the system utilizes the features of Big Data. Healthcare message boards available online provide huge volume of latest and raw data which can be used to mine useful information. In the first step, that is information retrieval, a search engine was developed by the authors that will search and download all the pages from the web pertaining to the disease Psoriasis. To check the relevance of the page, a threshold value for the count of the occurrence of the token Psoriasis in the document is maintained. The search engine was developed using Apache Lucene and Jsoup API. Using JSoup API, individual comments from the online users in the page are extracted and a corpus of text containing the comments is created.

4.

Topic detection: It is ensured that the topic of discussion in the page is about Psoriasis by using Latent Dirichlet Allocation (LDA) model. For this, first the text is normalized by eliminating stop words, punctuation symbols and lemmatizing the text. A term dictionary of the text and Document Term Matrix is created. Finally the topic of discussion in the text is arrived at by applying the LDA model on the document term matrix.

3.2. Experiments and Results:

The objective of the work is to extract only those messages from online healthcare forums that depict solutions or medica-tions or treatments that have worked for the disease Psoriasis. In this work Convoluted Neural Network is used to classify messages as solutions for the disease Psoriasis or messages that do not depict any solutions or treatments. The Convoluted Neural Network is implemented using Tensorflow API.

3.2.1. Training the model:

First some data loading parameters are initialized. The percentage of the training data to use for validation is initialized to as 0.1%. The model’s Hyper parameters are initialized. The character embedding dimension is initialized to 128. Comma separated filter sizes are assigned as 3, 4, 5. For each filter size the count of filters is initialized to 128. Dropout keep probability is initialized to 0.5. L2 regularization lambda is assigned as 0.0. Training parameters are initialized. Batch size is initialized to 64. Number of training epochs is set to 200. Evaluate model on dev set after this many steps is set to 100. The number of steps after which the model should be saved is set to 100. Number of checkpoints to store is initialized to 5. Allow device soft device placement is set to true. Log placement of ops on devices is set to false.

3.2.2. Data Preparation:

In the first step the messages from healthcare forums that are stored in files are read and the messages are split into words and the solutions and non-solution labels are generated. Here the split sentences and labels are generated. The steps involved in this process are: First the text that have messages that depict treatments that have worked and messages that do not denote treatments that have worked are read and loaded from files. The messages are then split into list of words. The list of words as representing solutions or nonsolutions. In the second step of data preparation, we shuffle the data at each epoch. In this step, a batch iterator is generated for a dataset. The training data is divided into batches. For

each epoch the count of the number of batches is calculated and the data is shuffled at each epoch. After preparing the data, the next step is to buid the vocabulary. Here the words present in the messages and their frequencies are prepared. That is the word vector is prepared for the messages. Then the data is randomly shuffled. The dataset is split into test and training data. Next step is where the training is done.

3.2.3. The training process:

For this a tensorflow graph is used. The session configuration is set with the training parameters mentioned above. The Allow device soft device placement is set to Boolean value True indicating that if

a GPU is not available, then the CPU should be used. Log placement of ops on devices is set to Boolean value Flase. A convolutional neural network that works on text called TextCNN which a deep learning algorithm to classify sentences and perform jobs like sentiment analysis and question classification is used. TextCNN object is initialized with the dimension of the training data, number of classes that is 2 in our case, vocabulary size, Dimensionality of character embedding which is 128, filter size which are 3,4,5, number of filters which is 128 and L2 regularization lambda(default:0.0) is 0.0. The training procedure is defined as follows. A variable that maintains state in the graph across calls to run function is created. The initial value is assigned the value 0, name is assigned a global step and trainable is assigned False. A training optimizer called the Adam optimizer is used. The gradients of loss for the variables in var list is computed. The gradients are applied to variables. An Operation that applies gradients is generated. The Output directory for models and summaries is specified. The Summaries for loss and accuracy is generated. The Training summaries, Dev summaries checkpoints and vocabulary are written into files. All the variables are initialized. The training step where you train the model for a single batch is defined. The step for evaluating model on the dev set is also defined. Shuffled batches are generated and training is done using each batch.

3.2.4. Evaluation of the Model:

The data parameters mentioned above are defined. The labelled messages are loaded. The evaluation parameters are defined. They are the Batch size that is set to 64, the checkpoint directory to store the

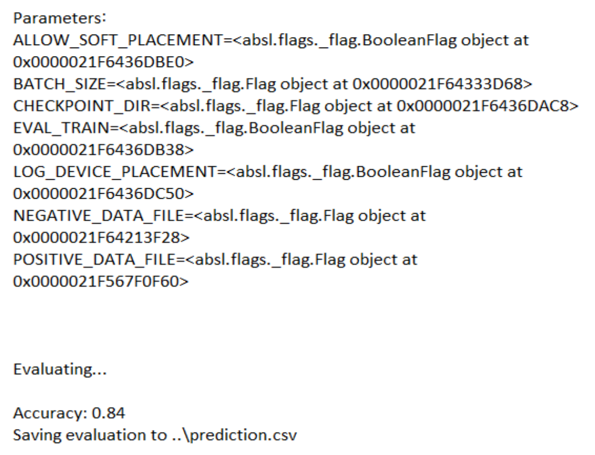
summaries generated during the training run, evaluation of all training data is set to true. Also parameters like Allow device soft device placement is specified as True and Log placement of ops on devices is specified as False The labelled text and their labels from the files are loaded. The data is then mapped into the vocabulary. The next process is the evaluation of the model. Here first the TensorFlow dataflow graph is created for the computations. Then the Tensorflow session is created for which the operations will be defined. The saved meta graph is loaded and variables from the checkpoint files are loaded.

The placeholders from the graph are got by name. The dropout keep probability to avoid over fitting is got. The tensors that need to be evaluated are got. The batches for one epoch are generated. The predictions are collected. For each batch, the tensorflow session is run and the predictions got. The model is run using the test data and the accuracy is printed. The evaluation is saved into a .csv file.

The TextCNN object that implements Convoluted Neural Network for text classification works as follows. The TextCNN uses embedding layer,the convolution layer, max-pooling and at the end the softmax layer. The input, output and dropout placeholders are created. To create the embedding layer, ramdom values from uniform distribution are generated. Parallel lookups on the list of tensors in params are performed. A dimension of 1 is inserted into the tensors shape.

Next a convolution and maxpool layer for each filter size is created. Here the convolution layer using embedded layer and strides as 1,1,1,1 are created. Nonlinearity using relu is applied. Maxpooling over the outputs are performed. All the pooled features are combined. Dropout is added. Final unnormalised scores and predictions are got. The meancrossentropy loss is calculated. Then the accuracy is calculated.

3.3. Output:



```
Parameters:
ALLOW_SOFT_PLACEMENT=<absl.flags._flag.BooleanFlag object at
0x0000021F6436DBE0>
BATCH_SIZE=<absl.flags._flag.Flag object at 0x0000021F64333D68>
CHECKPOINT_DIR=<absl.flags._flag.Flag object at 0x0000021F6436DAC8>
EVAL_TRAIN=<absl.flags._flag.BooleanFlag object at
0x0000021F6436DB38>
LOG_DEVICE_PLACEMENT=<absl.flags._flag.BooleanFlag object at
0x0000021F6436DC50>
NEGATIVE_DATA_FILE=<absl.flags._flag.Flag object at
0x0000021F64213F28>
POSITIVE_DATA_FILE=<absl.flags._flag.Flag object at
0x0000021F567F0F60>



Evaluating...

Accuracy: 0.84
Saving evaluation to ..\prediction.csv
```

Fig. 2. Output of the Model

3.4. Discussion:

In the above experiment Convoluted Neural Network was used to classify the messages from online healthcare forums on the treatments available for the disease Psoriasis as solutions or non-solutions. An accuracy of 84% was achieved. This is lower compared to the other machine learning algorithms applied on the same data. It may be due to the quantity of data required by CNN for analysis. CNN requires millions of messages for analysing the data. The data provided here consisted of few thousands of messages. Classification on the same data was performed using Artificial Neural Network and an average accuracy of 99% was got. Nave Bayes classifier on the same data provided an accuracy of 94%. Decision Tree provided an accuracy of 88%. SVM provided an accuracy of 88% [1].

The above empirical study shows that among the techniques used to classify messages as opinions stating successful treatments or other types of messages, the artificial neural network with an accuracy of 99% provides the better accuracy. The performance of CNN is found to be less in this experiment due to the volume of data.

4. LIMITATIONS

The limitations of this approach is that, since the messages are posted by average users, there may be noise, inaccurate and exaggerated information with spelling mistakes. Mining such information may lead to false positives. But the volume of the data may help in solving this problem. Repeatedly occurring treatments can be considered as true positives. Some of the messages that are posted may be to promote certain drugs or products. So further investigation by medical experts may be required. But since the volume of data required by a CNN algorithm to work effectively is in millions of text messages, its performance with the current dataset is less.

5. CONCLUSION

In this paper the author has extracted messages from health forums and classified them as treatments that have worked and treatments that have not worked for the disease Psoriasis. For classification Convoluted Neural Network(CNN) algorithm was applied. An accuracy of 84% was achieved.

6. CONFLICT OF INTEREST

The corresponding author, on behalf of all authors, states that there is no conflict of interest.

REFERENCES

[1] Mamatha Balipa. Opinion analysis of treatments from online healthcare forums for the disease psoriasis using nlp and artificial neural network. Development, 5(07), 2018.

[2] Jianfeng Gao, Patrick Pantel, Michael Gamon, Xiaodong He, and Li Deng. Modeling interestingness with deep neural networks. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 2–13, 2014.

[3] Jonathan Hedley et al. jsoup: Java html parser, 2015. Website (https://jsoup. org/).

[4] Hyeju Jang, Sa-Kwang Song, and Sung-Hyon Myaeng. Text mining for medical documents using a hidden markov model. In AIRS, pages 553–559. Springer, 2006.

[5] Rie Johnson and Tong Zhang. Effective use of word order for text categorization with convolutional neural networks. arXiv preprint arXiv:1412.1058, 2014.

[6] Rie Johnson and Tong Zhang. Semi-supervised convolutional neural networks for text categorization via region embedding. In Advances in neural information processing systems, pages 919–927, 2015.

[7] Nal Kalchbrenner, Edward Grefenstette, and Phil Blunsom. A convolutional neural network for modelling sentences. arXiv preprint arXiv:1404.2188, 2014.

[8] Yoon Kim. Convolutional neural networks for sentence classification. CoRR, abs/1408.5882, 2014.

[9] Machine Learnings. Text classification using neural networks, 1999.

[10] Apache Lucene. Apache lucene core, 2013.

[11] Thien Huu Nguyen and Ralph Grishman. Relation extraction: Perspective from convolutional neural networks. In Proceedings of the 1st Workshop on Vector Space Modeling for Natural Language Processing, pages 39–48, 2015.

[12] Yelong Shen, Xiaodong He, Jianfeng Gao, Li Deng, and Gr´egoire Mesnil. A latent semantic model with convolutional-pooling structure for information retrieval. In Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, pages 101–110. ACM, 2014.

[13] Peng Wang, Jiaming Xu, Bo Xu, Chenglin Liu, Heng Zhang, Fangyuan Wang, and Hongwei Hao. Semantic clustering and convolutional neural network for short text categorization. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), volume 2, pages 352–357, 2015.

[14] Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. arXiv preprint arXiv:1510.03820, 2015.