Epidemic Intelligence

MAULIK SONI, DAIICT

In this Report I have implemented multi-label text classifiers using machine learning models and statistical techniques as a part of Epidemic Intelligence model. The code is available on https://github.com/MaulikSoni-97/IR_Report

CCS Concepts: • **Natural Language Processing → Text Classification**.

Additional Key Words and Phrases: Hierarchical Attention Network, BERT, Epidemics

1 INTRODUCTION

From the historical experience we can say that the due to epidemic spread the cost, in terms of human loss to economic loss, every country has to be paid is very huge. So to prevent such a situation or at least become aware and take several precautionary steps before any epidemic disaster happens the Epidemic Intelligence based model can become very useful tool.

Epidemic Intelligence(EI) based model in the context of NLP is to use the text data either collected from social media or from different news sources to get the information related to epidemics of interest and then finding risk factor among those text. In general Epidemic Intelligence model can be divide in two parts.First part is collecting the text and classifying based on either it is related or not to the the epidemics of interest.In the second part based on all this classified text the goal is to prediction of the Epidemic in advance or founding the risk factor present in those texts.

So for this project I implemented the first part of this Epidemic Intelligence based model.Now rather than only classifying the text as useful or not for prediction it could be more helpful if we get deeper level information. For e.g, classifying the tweets in multi-label classes such as fever,runny nose, diarrhea in the case of COVID-19, then it might be more useful information rather just simply classifying in either related COVID-19 or not.

Now the report format is something like this. In the second section I explained dataset which I have used. In the third section I explained the implementation. Fourth and Fifth part contains the result and summary respectively. In the last section I have noted down some reference paper from which I took help for my project.

2 DATASET

For this report I have used data from the NTCIR-13 conference having MedWeb[1] task.This task consist of multi-label classification in which they provided 1920 tweets for training and 640 tweets for for testing.Originally tweets were in Japanese language but task organizer also converted them into English and Chinese dataset. Each tweet can be classified in one or many out of 8 given labels. This 8 labels are as following, Influenza, Diarrhea, Hayfever, Cough, Headache, Fever, Runnynose, Cold.

3 IMPLEMENTATION

For the multi label classification I have use three kinds of approaches.For the implementation and comparison of result I have taken help of paper published by Hayate.et.el[2] who are also winner of the MedWEb task.

3.1 Statistical Methods

Before applying any statistical approach first I converted tweets to their vector form for which I have use the TfIdf method.

In Statistical approaches first I tried a naive approach in which I have used one-vs-all method with SVM kernel and I got around 0.77 exact match accuracy. While accuracy is quite comparable to Hayate.et.el[2] but this method may not always performs better because in one-vs-all method at a time the focus is on single class so it means that it doesn't incorporate in the correlation with other label which is very crucial part for this kind of multi label classification. For e.g, someone has fever might have cold as well so at such a time correlation between labels take part very significant role in prediction.

In the second approach in statistical method I used Label Power Set method with logistic regression approach. Basically in this method model converts all possible label combination as an individual class. So if we have 8 labels to classify the each tweet than it makes total 2^8 i.e. 128 total classes and then used given kernel to classify each tweet in one of the class out of 128. So it becomes multi-label classification task to multi-class classification task. Now with this approach I got exact match accuracy around 0.75. Now issue with this approach is its overfitting tendency.

By means of overfitting I mean let's say that there is no single tweet present in training set which has both labels cold and fever but in test cases and in reality both it's possible. So during prediction time model always will never be able to label it entirely correctly.

3.2 Neural Network Based Method

3.2.1 Model architecture.

For this method I have used single step Hierarchical Attention network (HAN) described in this paper[2] while original paper of HAN is published by Zichao Yang et.el.[3]. Single step HAN is use for classification at the sentence level while full HAN is useful for the document classification. As my task was to classification tweets I used single step HAN.

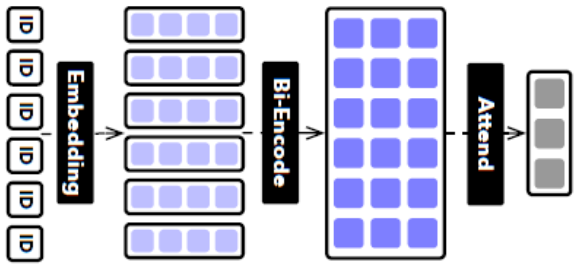

Fig. 1. Single step HAN Model

In HAN model first step was to convert the words into the corresponding vector form which was done by glove embedding[4]. Then these vector applied to Bidirectional Gated Recurrent Unit(GRU) network so it can able extract context from the both forward and reverse direction. Lets $\overrightarrow{x}_w$ represents word to vector representation then sentences

having maximum tokens size T. We will give the sequence of T vectors to GRU in forward and reverse direction. In result we get word representation word in both direction as represented by $\overrightarrow{h}_t$ and $\overleftarrow{h}_t$ in the right and left direction respectively. So to get single vector representation we concatenate the both the vectors.

$$\begin{aligned}\overrightarrow{x}_t &= \text{Glove(word)}, t \in [1, T] \\ \overrightarrow{h}_t &= \overrightarrow{GRU}(\overrightarrow{x}_t), t \in [1, T] \\ \overleftarrow{h}_t &= \overleftarrow{GRU}(\overrightarrow{x}_t), t \in [1, T]\end{aligned}$$

Now to get the sentence represenation we apply the attaention mechanism described in [3].

$$u_t = [W_w h_t + b_w] \tag{1}$$

$$\alpha_t = \frac{\exp(u_t^T u_w)}{\sum_t \exp(u_t^T u_w)} \tag{2}$$

$$s = \sum_t \alpha_t h_t \tag{3}$$

so s is vector representation of sentence now it can be used for the further classification. So a the end it requires single fully connected layer which results vector in R^8. Then based on signum function we can get the output.

$$\hat{y} = W_c s + b_c \in R^8 \tag{4}$$

$$\hat{y} = \begin{cases} 1 & \text{if } \hat{y} \geq 0 \\ \text{-1} & \text{otherwise} \end{cases} \tag{5}$$

3.2.2 *Loss Functions and Idea of Ensemble Model.*

As an ensemble approach I have used three different loss function.

$$Loss_{NLL} = \sum_i^N \sum_{c=1}^8 \ln(1 + \exp(-y_{c,i}\hat{y}_{c,i})) \tag{6}$$

$$Loss_{Hinge} = \sum_i^N \sum_{c=1}^8 \max(0, 1 - (y_{c,i}\hat{y}_{c,i})) \tag{7}$$

$$Loss_{Hinge-sq} = \sum_i^N \sum_{c=1}^8 \max(0, 1 - (y_{c,i}\hat{y}_{c,i}))^2 \tag{8}$$

Apart from the combination of three losses I also tried combination of two losses and for the combination of Hinge and Hinge square loss I get the highest match accuracy.

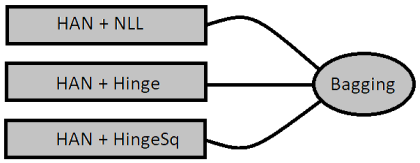

Fig. 2. ensemble idea

3.3 Bert Model

While BERT[5] has the direct API for multi class classification(BertForSequenceClassification) but not for the multi-label classification It is needed to create new class for it. SO for this issue I have used defined class on gitHub repo https://github.com/kaushaltrivedi/bert-toxic-comments-multilabel for this multi-label case.

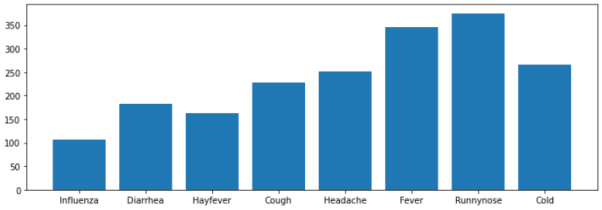

Fig. 3. Class vs Frequency plot(total tweets 1920)

In the BERT model case I experience that the model is overfitting on dataset as the dataset was not large and label occurrence to total number of tweets is quite low as obvious from the bar plot.. The result I got is also very poor is around 0.30 which can be looks very poor. One additional important thing I want to mention that out of 1920 around 500 tweets have not classified in either of class. Secondly during training time I also observed that in a single epoch I got the loss less than 1 while in HAN, even after 15 epoch I barely get minimum loss around 30-35 That's why I suppose that there is overfitting case here.

4 RESULT

As shown in the table I get maximum exact match accuracy for the Han with two loss Hinge and Hinge-Square. Here it can be seen as the comparing to non-ensemble part there is slight improvement over the result. One more important point is that the result which has * as post superscript is score in Reference column is got using multi-language method.Actually data which I had used is in three different language Chinese, Japanese and English. Now in multi-language model every sentence if first converted into their corresponding vector representation and then combination three different language is applied for the classification. In the BERT based model I got very poor accuracy because of overfitting as I discussed in previous section.

Table 1. Results

Method	Exact Match Accuracy	Reference
HAN + NLL	0.761	0.791
HAN + Hinge	0.746	0.795
HAN + HgSqr	0.771	0.786
HAN + NLL + Hinge	0.781	0.836*
HAN + Hinge + HingeSqr	0.793	-
HAN + NLL + Hinge + HingeSqr	0.785	0.842*
TfIdf+Label Power Set(Logistic regression)	0.751	-
TfIdf+Support Vector Classification(One Vs ALL)	0.772	-
Bert	0.30	-

5 CONCLUSION

So from given approaches we can say that from given text we can able to identify at some level the tweets which are more related to particular disease with its symptoms as well. Still there is further scope to the high Accuracy with more data in BERT based model. While multiple combinations of the different models might also be helpful as well.

6 BIBLIOGRAPHY

[1] Shoko Wakamiya, Mizuki Morita, Yoshinobu Kano, Tomoko Ohkuma and Eiji Aramaki: Overview of the NTCIR-13 MedWeb Task, In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies (NTCIR-13), pp. 40-49, 2017.

[2] Iso H, Ruiz C, Murayama T, Taguchi K, Takeuchi R, Yamamoto H,et al. NTCIR-13 MedWeb Task: Multi-label Classification of Tweets using an Ensemble of Neural Networks.

[3] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, pages 1480–1489, 2016.

[4] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation.

[5] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pretraining of deep bidirectional transformers for language understanding.

7 ONLINE RESOURCE

Report Link

https://github.com/MaulikSoni-97/IR_Report