

Sequential Sentence Classification in Medical Abstracts Using NLP and Transfer Learning

V. Sunil Tagore¹, K. Satish², K. Mohit³, S. Vishwa Tej⁴, M. Padmaja⁵

^{1,2,3,4}Student, Department of Computer Science Engineering
Gitam, Visakhapatnam, Andhra Pradesh, India

⁵Assistant Professor, Department of Computer Science Engineering
Gitam, Visakhapatnam, Andhra Pradesh, India

Date of Submission: 01-04-2023

Date of Acceptance: 10-04-2023

ABSTRACT:

From an application perspective, researchers need better tools to efficiently skim through the literature. So, automatically classifying each sentence in an abstract would help researchers read abstracts more efficiently, especially in fields where abstracts may be long, such as the medical field. So, we use various NLP techniques and tools to classify the sentences based on the context it belongs to, which also helps the researchers to identify their target thesis or information faster and makes their research more effective. On The PUBMEDdata for sequential sentence classification, Here it constructed numerous Artificial Intelligence (AI) models and used a data filtering strategy. We do a comparative analysis and discuss the results. The models developed and evaluated are Naive Bayes with TF-IDF encoder, Conv1d with token embedding, Tensorflow hub pertained Feature extractor, Conv1d with character embedding, Pretrained token embedding with positional and character embedding.

KEYWORDS: NLP, TensorFlow and TensorFlow Hub, Spacy, Pandas, NumPy, Sklearn etc

I. INTRODUCTION

Research articles are being published at an ever-increasing rate; those without organized abstracts may be challenging to read, which slows down researchers' progress through the literature. Researchers require more effective methods to efficiently scan the literature from an application standpoint. So, automatically categorizing each sentence in an abstract would

make it easier for researchers to read them, especially in subjects where abstracts may be lengthy, like medicine. Hence, we classify phrases according to the context they belong in using a variety of NLP techniques and tools. This aids researchers in quickly identifying their intended thesis or information and improves the efficacy of their research. Research papers are being published at an ever-increasing rate, and those without organized abstracts might be challenging to read. So, to address this issue, we are developing a machine learning model that takes sentences as input from difficult-to-read abstracts, identifies the important data, and classifies them into methods, results, and objectives. The Naive Bayes with TF-IDF encoder, Conv1d with token embedding, Tensorflow hub pertained Feature extractor, Conv1d with character embedding, Pretrained token embedding with positional and character embedding are the models developed and evaluated

II. SYSTEM METHODOLOGY

The offered basic model is made up of various sections. It starts by collecting data from a source and then pre-processing it in various ways. Next, using the methods provided, supply the sentence input. Following the execution of various algorithms, the final prediction and analysis are conducted. Some abstractions are structured, but others are not. Some structured abstracts contain inaccurate predictions. Several ways are used to deal with this problem and transform unstructured concepts to organized ones. Several artificially intelligent algorithms were presented, in which input was transmitted without any filter techniques being applied first, and with filter operations on the

language. When data filtering procedures were utilized vs when they were not, there was a considerable difference in accuracy. The models were developed and tested with five typical methods.

A. Methodology Process Diagram

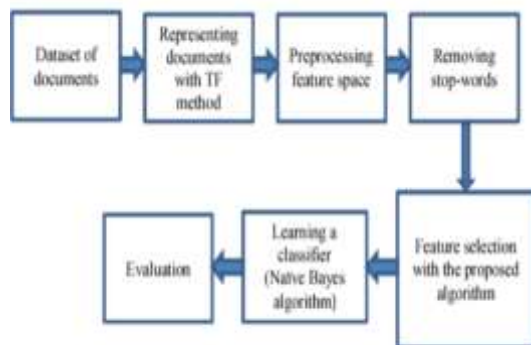


Fig. 1: PROCESS FLOW OF PROJECT

B. Feature Extraction:

Feature extraction refers to the process of transforming raw data into numerical features that can be processed while preserving the information in the original data set. It yields better results than applying machine learning directly to the raw data.

C. Tokenization

Tokenization is a technique used in natural language processing to divide paragraphs and words into smaller elements that may be assigned meaning more readily. The initial stage in the NLP process is to obtain data (a sentence) chunks

D. Data Processing And Reading

The data is processed based on the results of data collection and data cleaning processes to overcome data problems such as data anomalies, missing data values, data redundancy, and inappropriate data. The data is then selected and grouped by type and function to divide it into training and testing data so that it can be applied to the classification algorithm that will be tested

E. One Hot Encoding.

In machine learning, one-hot encoding is the translation of categorical data into a format that can be fed into machine learning algorithms to enhance prediction accuracy. In machine learning, one-hot encoding is a typical approach for dealing with categorical data.

F. Training And Validation

To summarize, the training set is typically the largest subset created out of the original dataset that is used to fit the models. The validation set is then used to evaluate the models in order to perform model selection. So in this process we will be evaluating the model by training and validation field

G. Evaluation

It is the main step in any machine learning model it help us to evaluate the model and to check whether our model predict the accurate output or not if our model haven't predict the accurate output then again we need to train our model with different algorithms and try to make the better predictions basing on our evaluation only our model will be fit to the network.

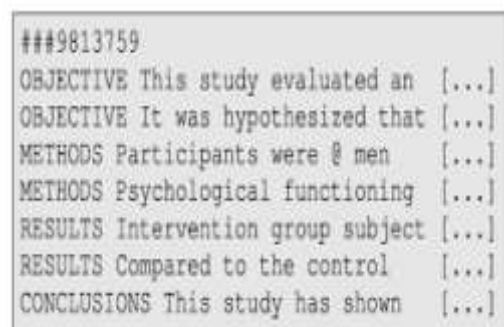
H. Prediction

Here we finally classify the text to one of the 5 classes like background, objective, methods, results and conclusion from medical abstracts.

III. IMPLEMENTATION

A. Dataset

The dataset we will utilize is PubMed 200k RCT, a novel dataset for sequential sentence categorization based on PubMed. The collection contains 2.3 million phrases from around 200,000 abstracts of randomized controlled trials. Each abstract sentence is identified with its role in the abstract by utilizing one of the following classes: background, objective, method, outcome, or conclusion. PubMed contains over 35 million biomedical literature citations from MEDLINE, life science journals, and online books. Automatically identifying each line in an abstract would aid academics in reading abstracts more effectively, particularly in subjects where abstracts might be lengthy, especially in the medical field.



```

###9813759
OBJECTIVE This study evaluated an [...]
OBJECTIVE It was hypothesized that [...]
METHODS Participants were 8 men [...]
METHODS Psychological functioning [...]
RESULTS Intervention group subject [...]
RESULTS Compared to the control [...]
CONCLUSIONS This study has shown [...]
  
```

Fig.2:Pub Med Dataset Text File

B. Naive Bayes With TF-IDF Classifier.

Multinomial Naive Bayes is a probabilistic learning algorithm primarily used in Natural Language Processing. The program guesses the tag of a text, such as an email or a newspaper story, using the Bayes theorem. It computes the likelihood of each tag for a given sample and outputs the tag with the highest likelihood. The Naive Bayes classifier is a collection of numerous methods that all have one basic principle: each feature being categorized is unrelated to any other feature. The presence or absence of one characteristic has no effect on the presence or absence of another.

Multinomial Naive Bayes Classifier

$$C_{MAP} = \underset{C \in C}{\operatorname{argmax}} P(x_1, x_2, \dots, x_n | C) P(C)$$

$$C_{NB} = \underset{C \in C}{\operatorname{argmax}} P(C) \prod_{x \in X} P(x | C)$$

Fig. 3: Multinomial Naive Bayes Formula

C. Random Forest

Random Forest is a famous and widely used method among Data Scientists. Random forest is a type of Supervised Machine Learning Algorithm that is commonly used in classification and regression issues. It constructs decision trees from several samples and uses their majority vote for classification and average for regression.

One of the most essential characteristics of the Random Forest Algorithm is that it can handle data sets with both continuous variables, as in regression, and categorical variables, as in classification. It excels in classification and regression problems. In this lesson, we will learn how random forests function and apply them to a classification job

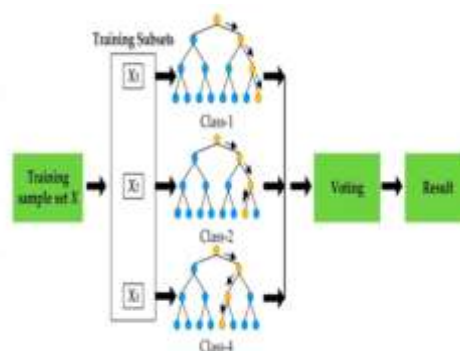


Fig. 4: Random Forest on Sentence Classification

D. Conv_1d with token embedding.

Conv1D is a class. Layer of 1D convolution This layer generates a convolution kernel, which is convolved with the layer input over a single spatial (or temporal) dimension to generate a tensor of outputs. If the use bias option is set to True, a bias vector is generated and appended to the outputs.

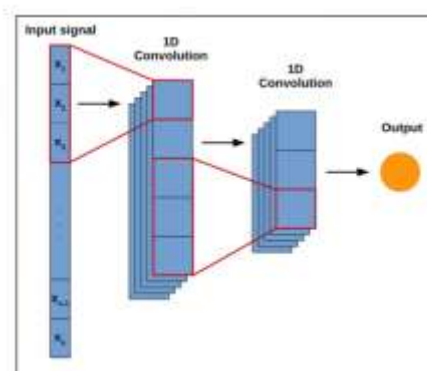


Fig. 5: Convolutional 1D Process flow

E. Transfer Learning (pre-trained token embedding)

Transfer learning is basically applying some other algorithm traits which are well trained with that model, now using that model patterns or traits to our own model or algorithm. This process helps us to save a lot of time as all the tokenization, preprocessing is done already and we can apply our embeddings or to our dataset directly and then we can also deploy fastly.

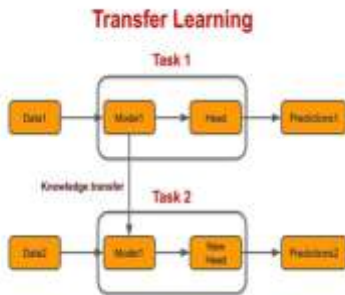


Fig. 6: Transfer Learning Process flow

F. Tribrid_token_char_positional Embedding

Here we basically mix 3 different embeddings first with token embedding, then character embedding and then positional embedding after that we add a bi-directional LSTM and a dense layer on top of it. So using three different embedding increases our model pattern understanding and can improve the model accuracy.

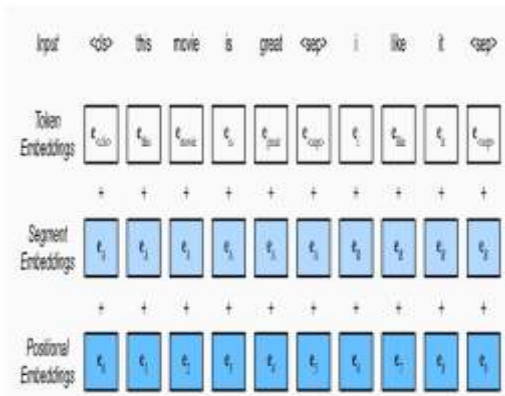


Fig. 7: Tribrid_tok_char_pos embedding flowchart

G. Bi-LSTM(Bi Directional LSTM)

Bi-LSTM is a Two way input taker where it takes one input in the forward direction and the another input in the backward direction using the embedding layer which is after concatenated and flattened to one common dimensional. Also 'Bi' means bi-directional which means it holds both past as well as future data and LSTM is Long short term memory which can hold long term memory and this LSTM is heavily used in the sequential and sequence data.

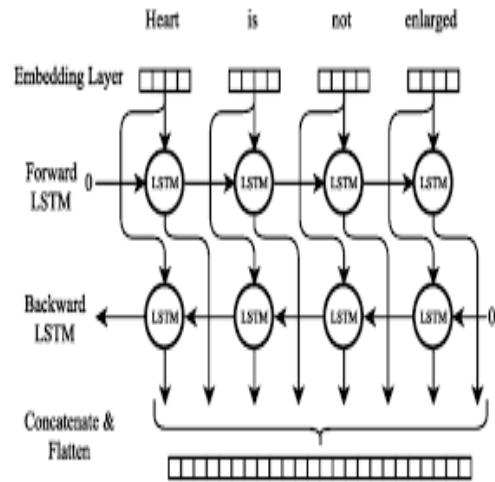


Fig. 8: Bi-directional LSTM Process flow

H. Visualization of our Tribrid model architecture(token+char+positional embeddings)

Transfer Learning with pre trained token embedding +character embedding + positional embedding got an accuracy of 83%



Fig. 9: Tribrid model architecture flow chart

I. TF-IDF(Term Frequency Inverse Document Frequency) Formula

TF-IDF basically helps us to understand the importance and the usage of the text especially in the case of larger texts, Where we can identify that text related information, its words, token and phrases and works as a vocabulary oriented dictionary for all the meanings of the tokens and sentences.

$$w_{i,j} = tf_{i,j} \times \log\left(\frac{N}{df_i}\right)$$

tf_{ij} = number of occurrences of i in j
 df_i = number of documents containing i
 N = total number of documents

Fig. 10: TF-IDF Formula

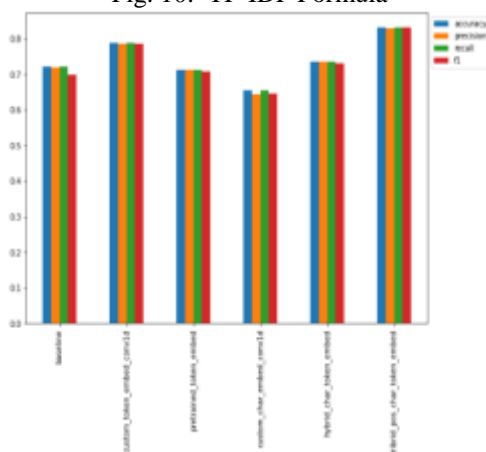


Fig. 11: Models Comparison Graph chart

IV. RESULTS & DISCUSSIONS

ALGORITHMS	ACCURACIES
1. BASELINE(Naive Bayes)	72.1%
2. Random Forest	78.57%
3. Conv_1D With Token Embedded	78.8%
4. Pretrained_Token_Embedded (Transfer Learning)	71.3%
5. Hybrid character with Token Embedded	73.2%
6. Tribrid_Pos_char_Token Embedded(Transfer Learning)	83.09%

Table. 1: Accuracies of various Algorithms

V. CONCLUSION

Sequential sentence classification is essential especially in the medical field, these days medical research is increasing drastically and the search for appropriate medical data is not very clear and easy to read. So In order to solve this problem in this project we used our ML models and trained on different medical abstracts and tried to reduce the complexity of the medical abstracts with the main goal of helping the researchers to choose their related data more easily. Among all our models, the tribal ML model outperformed other models and was able to classify the abstracts according to their respective target labels.

VI. FUTURE SCOPE

In the future, we will improve our model to handle more classes of sentences on diverse datasets. We have also planned to handle real-time abstracts from famous conferences with less computational complexity. Also usage of complete CNN,LSTM,GRU or deep learning techniques can also be tried in future. We are also planning to create an extension which will be further deployed and it would be classifying all of the users' complicated texts or abstracts and making their work easier and faster. More exploration of the classification feature set could possibly lead to better performance. We are also planning to use BERT and sci-BERT for our training samples as an extension of our project.

REFERENCES

- [1]. Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2016. Neural networks for joint sentence classification in medical paper abstracts. European Chapter of the Association for Computational Linguistics (EACL) 2017
- [2]. Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In Human Language Technologies 2016: The Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT.
- [3]. Cross-Domain Multi-Task Learning for Sequential Sentence Classification in Research Papers Arthur Brack TIB – Leibniz Information Centre for Science and Technology & Leibniz University Hannover, Germany Arthur.Brack@tib.eu Anett Hoppe TIB – Leibniz Information Centre for Science and technology