Received: 22 August 2022 | Revised: 07 November 2022 | Accepted: 12 November 2022
DOI: 10.37917/ijeee.19.1.11 Vol. 19| Issue 1| June 2023

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

IJEEE University of Basrh College of Engineering

Identifying Discourse Elements in Writing by Longformer for NER Token Classification

Alia Salih Alkabool [1], Sukaina Abdul Hussain Abdullah[2], Sadiq Mahdi Zadeh[2], Hani Mahfooz[2]
[1] University of Basrah, Basrah, Iraq
[2] Islamic Azad University, Isfahan, Iran

Correspondence
*Alia Salih Alkabool
University of Basrah, Basrah, Iraq
Email: aliasalihjali@gmail.com

Abstract
Current automatic writing feedback systems cannot distinguish between different discourse elements in students' writing. This is a problem because, without this ability, the guidance provided by these systems is too general for what students want to achieve on arrival. This is cause for concern because automated writing feedback systems are a great tool for combating student writing declines. According to the National Assessment of Educational Progress, less than 30 percent of high school graduates are gifted writers. If we can improve the automatic writing feedback system, we can improve the quality of student writing and stop the decline of skilled writers among students. Solutions to this problem have been proposed, the most popular being the fine-tuning of bidirectional encoder representations from Transformers models that recognize various utterance elements in student written assignments. However, these methods have their drawbacks. For example, these methods do not compare the strengths and weaknesses of different models, and these solutions encourage training models over sequences (sentences) rather than entire articles. In this article, I'm redesigning the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements corpus so that models can be trained for the entire article, and I've included Transformers, the Long Document Transformer's bidirectional encoder representation, and the Generative Improving a pre trained Transformer 2 model for utterance classification in the context of a named entity recognition token classification problem. Overall, the bi-directional encoder representation of the Transformers model railway using my sequence-merging preprocessing method outperforms the standard model by 17% and 41% in overall accuracy. I also found that the Long Document Transformer model performed the best in utterance classification with an overall f-1 score of 54%. However, the increase in validation loss from 0.54 to 0.79 indicates that the model is overfitting. Some improvements can still be made due to model overfittings, such as B. Implementation of early stopping techniques and further examples of rare utterance elements during training.

KEYWORDS: BERT - Bidirectional Encoder Representations from Transformers, NER - Named Entity Recognition, Longformer – Long Document Transformer, GPT2 - Generative Pre-Trained Transformer 2, NLP - Natural Language Processing, GSU - Georgia State University

I. Introduction

1) The importance of writing

Having the ability to write clearly and concisely is a key skill for all careers. Individuals who are able to express their thoughts and ideas have an advantage when writing business emails, proposals, or opposing or supporting new policies. The Source Expert website notes in their article 43 Why Writing Matters to Students: "There are a variety of ways to communicate with others, but writing will always be part of your daily life." [1]. Although writing is an important human skill, many students lack writing skills. The National Assessment of Educational Progress found that less than 30% of high school graduates are proficient writers. They also showed that this problem is more acute in low-income communities where proficient writing rates are less than 15% [2]. As researchers at Georgia State University have pointed out, this problem is primarily due to many schools, especially those in low-income communities, not having the resources to provide personalized feedback on students' writing [3]. Fortunately, one of the problems can be resolved by automatically writing feedback. Automatic writing feedback systems are programs that can analyze and critique students' writing while the teacher is away. These programs are already popular in many applications, such as Microsoft Outlook's Autosuggest and Grammarly. In fact, Trey from the website “apoven”, at how a writing feedback system like Grammarly can be used to expand one's vocabulary and provide them with instant mini grammar lessons [4]. In response, many agencies have taken steps to improve our current automated feedback system.

2) The current machine learning approach

Two institutions, GSU and the Learning Institutions Laboratory, have investigated a machine learning-based approach to improving automated feedback systems. They argue that machine learning models can be trained to accurately classify discourse elements in written works. This model can then be added to an existing feedback system to help the system provide better and more constructive feedback to students. The Learning Institution Lab took the first steps towards creating this model by creating the corpus the Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE). The PERSUADE corpus is a collection of over 25,000 argumentative papers collected from students in grades 6 to 12 [5]. All articles are annotated by professional English teachers in order to understand the different elements of discourse. After creating this dataset, machine learning researchers have the basic facts they need to start training models. In particular, they hope to optimize existing natural language processing (NLP) models for discourse classification tasks, focusing on Google's Bidirectional Encoder Representation (BERT) model from Transformers. I agree with the current approach to fine-tuning this model for discourse classification; however, I believe some steps are required to make these models more accurate.

3) The Discourse Elements

This list of discourse elements has been compiled by a team of teachers and professional writers from The Learning Agency Lab [6]. They believed that this list contained all the important discourse elements that make up the students' writing, and they used this list as a template for creating the PERSUADE corpus. I will use the same rules when improving my own models:

- Introduction - an introduction that begins with statistics, citations, descriptions, or other means of grabbing the reader's attention and pointing to the paper
- Position - opinion or conclusion on the main issue
- Statement - a statement supporting the position
- Counterclaim - an allegation that refutes another allegation or justifies the position to the contrary
- Rebuttal - rebutting the assertion of the counterclaim
- Evidence - an opinion or example to support a claim, counterclaim or refutation.
- Closing Statements - Closing Statements Reaffirming Statements.

4) My approach & potential outcomes

As with current machine learning approaches, I believe that transformation-based models such as bidirectional encoder representations from Transformers (BERT) [7] can be fine-tuned to successfully address discourse classification problems. However, in this article, I also want to examine other transformer models and compare/contrast the different results. In addition, further improvements can be made to the corpus of Persuasive Essays for Rating, Selecting, and Understanding Argumentative and Discourse Elements (PERSUADE). Current corpora divide articles into sequences, each sequence corresponding to a different type of utterance. However, I will restructure the dataset so that the full paper can be provided to the model during training. In this post, I hope to demonstrate that Transformer-based models should be trained concurrently throughout the post to take full advantage of their architecture. I also hope to demonstrate that it is useful for machine learning researchers to evaluate models other than BERT (Bidirectional Encoder Representations for Transformers) models, and I hope to demonstrate that Long Document Transformer (Longformer) [8] -Model is better in the following cases The model comes to discourse classification.

5) Outline

To justify my approach, I first turn to other projects focusing on discourse classification. I then describe in more detail my approach to the problem of utterance classification and what I have done to implement a transformer-based model bidirectional encoder representation slave transformer (BERT), long document transformer (Longformer), and generative pretrained transformer take Step-2 (GPT-2). Afterwards, I will review some of my promising findings and explain their implications for discourse classification tasks. Finally, I'll cover some improvements that can be made for future fine-tuning attempts.

6) Related Works

Researchers Burstein et al. [7] attempted to use a Bayesian classifier to identify thesis statements in student written work. Their model was able to achieve an average overall accuracy of 43%, but most importantly, they were able to show that the classification of propositional statements was generalizable. That is, the model does not need to be retrained for each new paper prompt, and once trained, the model can recognize paper statements in all paper topics. One drawback, however, is that the training set for the Bayesian classifier is small, with only 100 articles, and the authors admit that their model can hold. Another model to mention is the Longformer model modified by programmer Darek Kleczek [8]. Kleczek solves this problem by optimizing a pre-existing longformer model on the hug face website [9], which was originally trained by machine learning engineers at allenai. By fine-tuning the Longformer model, Kleczek was able to achieve an accuracy of 61.4

Taboada et al. [9] Enters the history of Rhetorical Structure Theory (RST) and its advantages today. They found that RST can be used for a variety of applications (including discourse classification) and is a "robust and well-tested theory". Most importantly, they found relationships between various elements of utterances that we hope our model will capture. Instead of trying to specifically define the relationship between the models or create a working machine learning model, the researchers leave it as an open problem for others to solve.

The machine learning researcher Julian Peller [10] addresses the problem of classifying utterance elements by improving Google's BERT model. He approaches the problem as a tokenized classification problem using Named Entity Recognition (NER), where articles are lists of tokens and utterance elements are distinct classes. He also trained on 10,000 articles from the PERSUADE corpus, and he achieved an overall accuracy of 0.226 on the F-1 score.

Ali Habiby [11] tackled the problem of classifying utterance elements in a rather unique way. Instead of defining the problem as a NER token classification problem, Habiby formulates the problem as a Q&A problem, which allows him to use a Q&A model. The Transformer model Habiby chose to fine-tune is Roberta, a BERT-inspired model from Facebook. Habiby used a maximum length of 448 characters and a stride of 192 for his model and trained his model for 3 epochs. His F-1 overall is 0.453.

Roman et al. [12] used several machine learning techniques in their approach to the problem of classification of discourse elements. The first technique they used was weighted box fusion, which combines the outputs of 10 different models into a single decision. Most of the models used are variants of the Deberta model and the Longformer model. After obtaining the model results, the team used post-processing, such as fixing range predictions and utterance-specific rules, to clean up the model's output after making the predictions. The F-1 total is 0.74, and the model is trained for 5 epochs on Nvidia's V100 32GB GPU and A100 40GB GPU. In this project, machine learning researcher Ali Habiby [13] used a random forest model instead of his previous Q&A model to solve the discourse element classification problem. One advantage of this model is that it is easy to understand and replicate. The train/test split chosen by Habiby for this model is 70% train and 30% test, and the model has an overall f-1 value of 0.25. While this model is easy to replicate and understand, I think the model is too simplistic given the low f-1 value to see how the different utterance elements are related to each other.

Lonnie [14] uses the Keras library to create an LSTM network that can classify utterance elements in student papers. One notable layer included in the Lonnie model is the cushion layer of length 1024. This is important because most other solutions are fine-tuned versions of the BERT model, however, the BERT model can only hold 512 tokens at a time. So Lonnie's model is better able to accommodate larger student papers than most other solutions, but Lonnie still trains on one sequence of data at a time, which I think prevents his model from reaching its full potential. Overall, the f-1 value of the Lonnie model is 0.214.

Drakuttala [15], a machine learning researcher, fine-tuned the RoBERTa base model by addressing the discourse element classification problem. One thing that stands out about Drakuttala's method is that he clearly defined each element during the model training. Instead of using 7 classes like most other researchers, he used Claim, Position, Lead and Counter Claim. Drakuttala organized their data into two parts: B and I. Class I, like its name implies, is for words considered part of an entity. Drakuttala used this principle instead of one Lead class— instead, they created two Lead classes, B-Lead and I-Lead. Drakuttala achieved a 0.54 f-1 score during training on 3 epochs with a 1e-5 learning rate and a 512 token length.

II. Approach (and Technical Correctness)

1) PERSUADE corpus

The training and testing data used to fine-tune my model is the PERSUADE corpus, a dataset created by Learning Agency Lab. I chose this dataset because it is specially designed for the problem of discourse classification. The corpus contains over 25,000 student papers, all annotated by writing professionals [16]. To ensure that the dataset is as accurate as possible, each article is annotated using a double-blind scoring procedure and reviewed by another third-party writing professional [17]. The content of this dataset is very good and very useful for training/testing models; however, I believe some changes to the format of the dataset can be made through data preprocessing.

2) Data preprocessing

To preprocess the data for this model, I decided to reassemble the individual sentence sequences into a joint article. In the PERSUADE corpus, articles are divided into sequences, each sequence representing a different discourse element. I believe this is not the best way to optimize transformer-based models as this use positional encoding. Positional encoding is a technique added to the Transformer architecture because the model is acyclic, which means that "Hello World" and "World Hello" sequences look the same in the Transformer model [18]. By adding positional encoding to the word embedding, the Transformer model can learn that different word positions in the text have different meanings, and I believe this tool can be used for discourse classification purposes. This is because certain discourse elements, such as closing sentences, are highly correlated with their position in the text; merging the sequences before starting fine-tuning gives the model a chance to learn how the position of the sequence in the paper is related to its discourse type.

3) Three different models (BERT, Longformer, and GPT-2)

The three models chosen for fine-tuning this document are the BERT, Longformer, and GPT-2 models. I decided to refine some models because I wanted to see how different model architectures address the problem of discourse classification. I was also interested in whether different models are better at classifying different elements of discourse. We chose the BERT model because it is one of the most popular models for NLP tasks. According to the Hug Face database, the BERT model was downloaded 15.8 million times by researchers in April 2022, making it the second most popular NLP model [19]. I chose to include this model in my own study so that my results could be compared with those of other researchers. Another model that I am improving is the GPT-2 model. This model is a popular model, but I included it in the project mainly because of the model's design. Unlike his BERT model, which stacked the coding layers of the transformer, the GPT-2 architecture stacked the decoding layers of the transformer [20]. In this post, we want to see if this small design change affects the output of speech classification results. The last model that I will improve on, and one that I think is the most promising, is the Longformer model. The Longformer model is an extension of the BERT model designed to handle larger input values without compromising quality [21]. This feature is important for my research because data preprocessing produces long input values and most models forget what they learned at the beginning of the sequence. The longformer model is important for my research because it uses data preprocessing without sacrificing quality. I believe this model shows how far my pretreatment technology can go.

4) Hyper-parameters

For the model hyper-parameters I used:

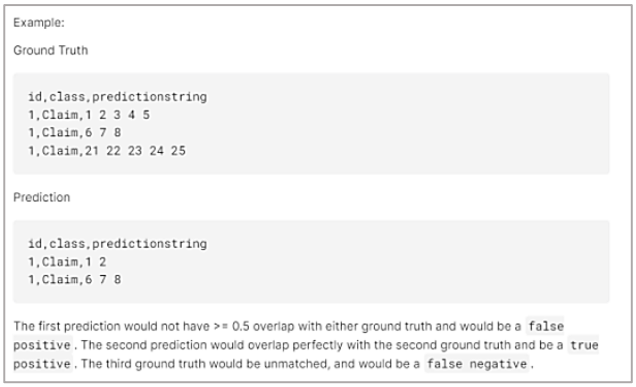
- Batch-size = 1
- Learning-rate = 5e-5
- Epochs = 7
- Warm-up ratio = 0.1
- Gradient-accumulation = 8
- Weight-decay = 0.01

5) F-1 score

To evaluate the model, I will use the f-1 score. To calculate the f-1 score, use the following formula:

$$F - 1 \text{ point} = TP=(TP + 0:5 * (FP + F N))$$

Before we can use this formula, we need to find true positive, false positive and false negative values as defined by GSU researchers. As the GSU team sees in this post, each model evaluation will contain a ground truth and prediction. The ground truth is which utterance class the sequence (phrase) belongs to, and the prediction is which class the model thinks the sequence belongs to. If the predicted sequence overlaps the ground truth sequence by 50% or more, it is considered a true positive. If there is a mismatched predicted sequence then I consider it as a false positive, if there is a mismatched ground truth sequence then I consider it as a false positive. Figure 1 shows examples of these forecast sequences and explains in more detail how they are calculated.



Example:

Ground Truth

```
id,class,predictionstring
1,Claim,1 2 3 4 5
1,Claim,6 7 8
1,Claim,21 22 23 24 25
```

Prediction

```
id,class,predictionstring
1,Claim,1 2
1,Claim,6 7 8
```

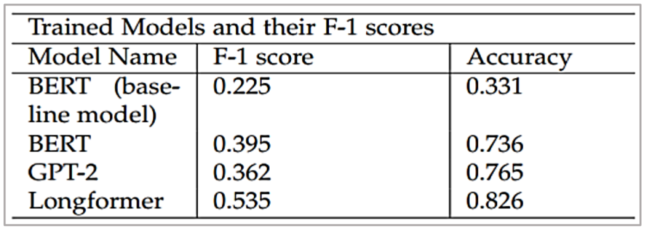
The first prediction would not have >= 0.5 overlap with either ground truth and would be a `false positive`. The second prediction would overlap perfectly with the second ground truth and be a `true positive`. The third ground truth would be unmatched, and would be a `false negative`.

Fig. 1 How TP/TN/FN are calculated.

III. Experimental Results (and Technical Correctness)

1) Data preprocessing and sequence merging

As you can see from the table "Trained models and their F-1 scores", the BERT model trained without my data preprocessing method has a 17% lower f-1 value (Fig. 2), with reduced accuracy 41% (Fig. 3) than using my data preprocessed version of the model. This is because the main and concluding statements almost always appear at the beginning and end of the essay, respectively. My model was able to leverage positional encoding and understand the relationship between introductory and closing statements and their positions in student essays. My work shows that the best way to train a transformer-based discourse classification architecture is to reassemble the sequences into a full article and let the model use their positional encodings to explore relationships between discourse elements.



Trained Models and their F-1 scores

Model Name	F-1 score	Accuracy
BERT (base-line model)	0.225	0.331
BERT	0.395	0.736
GPT-2	0.362	0.765
Longformer	0.535	0.826

Fig. 2: Macro f-1 scores of all models after training

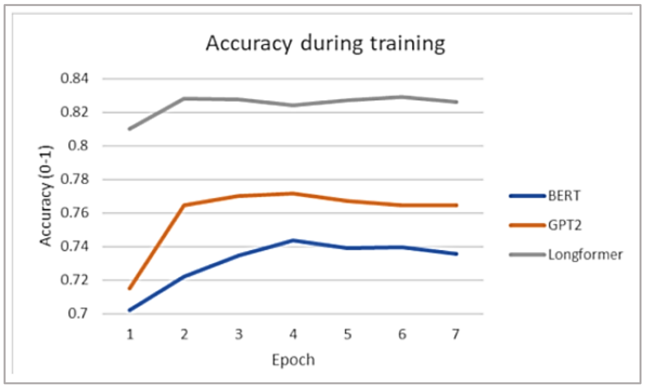


Fig. 3 Accuracy during model training

2) Comparing transformer-based architecture for discourse classification

In my experiments, I distilled 3 Transformer models from the Hugging face library: BERT, Longformer, and GPT2. From the "Trained models and their F-1 scores" table, we can see that of all the fine-tuned models, the Longformer model has an f-1 value of 0.535, which is the best performer. The BERT model ranks second with an f-1 scores of 0.395, and the GPT2 model is the worst with a value of 0.362. My work here shows that the best model for discourse classification is the Longformer model. I believe that the longformer's ability to handle large data inputs without losing important information is why this model has been so successful in my experiments.

3) High Lead/Concluding Statement scores

All models scored relatively high in the main and conclusive claims category, and low in the counterclaim category. As shown in Fig. 4, the average f-1 score for leading and trailing sentences are 0.751 and 0.587, respectively, the two highest among all categories. This goes against conventional wisdom, since cues and conclusive statements are not as common as other categories (such as claims); one would assume that the claim category is the highest because the model has more examples to train on. I believe these results arise because the opening and closing statements are closely related to their position in the paper. That is, the introductory and concluding sentences almost always appear at the beginning and end of the job, respectively, which makes it easier for the model to learn these positionally encoded classes. So, my work here shows that giving a model full paper allows the model to perform well in uncommon categories. However, some categories, such as rebuttals and counterclaims, may require further training examples.

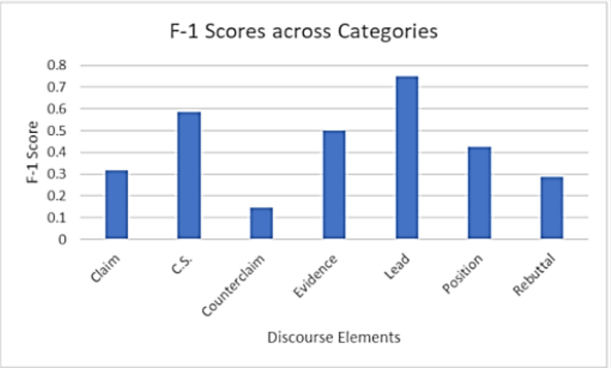


Fig. 4: Average F-1 scores across all models (except Baseline) for each discourse element

4) Increasing validation loss

All models started to show an increase in validation loss after epoch 3, for example, the top-performing longformer model increased its validation loss from 0.54 to 0.79 over epochs 3 to 7 (Fig. 5). According to the Javatpoint article "Overfitting in Machine Learning" [22], a telltale sign of overfitting a model is increased validation error during training, and one way to prevent this is to stop early. Figure 6 show the F-1 scores during model training. As defined by the Elite Data Science website, early stopping is the process of "...stopping the training process before the learner passes that point [point where variance starts to increase] ..."[23]. I believe I should implement early stopping for my model around the 2nd or 3rd epoch because that's when the variance starts to increase. Another approach I could try is to augment the examples during the training phase. According to Xiaoshuang Shi in his article The Problem of Overfitting and How to Resolve It [24], sharing more training examples is a good way to solve the overfitting problem. In particular, I should provide articles with many examples of counterclaims and rebuttals, because that's where my model's performance is weakest. My work here shows that when fine-tuning a model for classifying discourse elements, more emphasis needs to be placed on getting more examples, rather than applying the model to a large number of epochs.

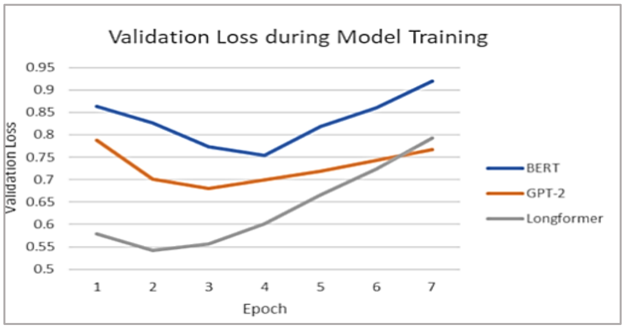


Fig. 5 Validation loss during model training

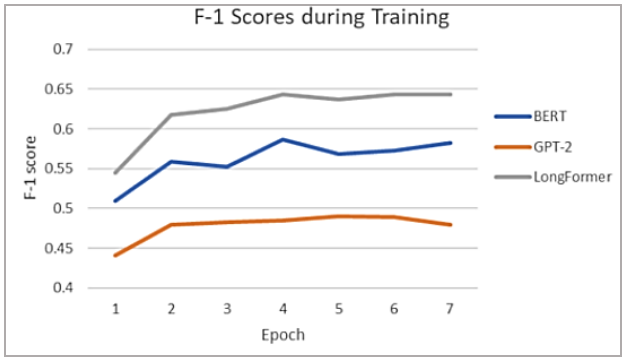


Fig. 6: F-1 scores during model training.

IV. Conclusion

In conclusion, writing is an important skill and it is vital for young people to develop their writing skills. By using an automated writing feedback system, we can help students develop their writing talents by providing a detailed analysis of their writing. One way to improve current automated writing feedback systems is to combine them with machine learning models to differentiate between different writing elements in student essays. In this experiment, I show that the longformer model outperforms the BERT or GPT2 models in discourse classification. I also show how the entire article guides the model during fine-tuning to learn positional relationships between utterance elements, especially for the Lead and Final Statement classes. However, positional encoding alone does not solve the discourse classification problem, and more attention needs to be paid to acquiring more categories of data, such as rebuttals or counterclaims, to improve the overall results.

Conflict of Interest

The authors have no conflict of relevant interest to this article.

References

[1] I. Yulianawati, “Self-Efficacy and Writing : A Case Study at A Senior High School in Indonesian EFL Setting,” *Vis. J. Lang. Foreign Lang. Learn.*, vol. 8, no. 1, pp. 79–101, 2019,

[2] L. Darling-Hammond, “Teacher quality and student achievement: A review of state policy evidence,” *Educ. Policy Anal. Arch.*, vol. 8, no. November 1999, 2000.

[3] Trey, “5 reasons to use grammarly,” Oct 2019. [Online]. Available https://www.apoven.com/grammarly-benefits/

[4] T. N. Fitria, “Grammarly as AI-powered English Writing Assistant: Students’ Alternative for Writing English,” *Metathesis J. English Lang. Lit. Teach.*, vol. 5, no. 1, p. 65, 2021.

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” *NAACL HLT 2019 - 2019 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. - Proc. Conf.*, vol. 1, no. Mlm, pp. 4171–4186, 2019.

[6] I. Beltagy, M. E. Peters, and A. Cohan, “Longformer: The

Long-Document Transformer," arXiv:2004.05150, 2020, [Online]. Available: http://arxiv.org/abs/2004.05150.

[7] J. Burstein, D. Marcu, S. Andreyev, and M. Chodorow, "Towards automatic classification of discourse elements in essays," ACL '01: Proceedings of the 39th Annual Meeting on Association for Computational Linguistics, pp. 98–105, July 2001. https://doi.org/10.3115/1073012.1073026

[8] A. H. Mohammed and A. H. Ali, "Survey of BERT (Bidirectional Encoder Representation Transformer) types," *J. Phys. Conf. Ser.*, vol. 1963, no. 1, 2021, doi: 10.1088/1742-6596/1963/1/012173.

[9] W. C. Mann and M. Taboada, "Rhetorical Structure Theory : looking back and moving ahead," *Discourse Stud.*, vol. 8, no. 3, pp. 423–459, 2006.

[10] J. Peller, "Feedback- baseline sentence classifier [0.226]," Kaggle, Dec 2021. [Online]. Available: https://www.kaggle.com/code/julian3833/feedbackbaseli sentence-classifier-0-226/notebook.

[11] A. Habiby, "Roberta qna model," Kaggle, Jan 2022. [Online]https://www.kaggle.com/code/aliasgherman/roberta-qnamodel-maxlen-448-stride-192

[12] R. Solovyev, W. Wang, and T. Gabruseva, "Weighted boxes fusion: Ensembling boxes from different object detection models," *Image Vis. Comput.*, vol. 107, p. 104-117, 2021, doi: 10.1016/j.imavis.2021.104117.

[13] A. Habiby, "Randomforest only (gradientboostnow)," Kaggle, Jan 2022. [Online]. Available: https://www.kaggle.com/code/aliasgherman/randomforest only-

[14] Lonnie, "Name entity recognition with keras," Kaggle, Dec 2021. [Online] https://www.kaggle.com/code/lonnieqin/namentityrecognition-with-keras

[15] raghaven drakotala, "Fine-tunned on roberta-base as ner problem [0.533]," Kaggle, Dec 2021. [Online]. Avail able: https://www.kaggle.com/code/raghavendrakotala/finetunned-on-roberta-base-as-nerproblem-0-533

[16] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[17] Huggingface, "Models," Apr 2022. [Online]. Available: https://huggingface.co/models

[18] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.

[19] https://huggingface.co/bert-base-uncased

[20] D. Rothman," Transformers for Natural Language Processing: Build Innovative Deep Neural Network Architectures for NLP with Python, PyTorch, TensorFlow, BERT, RoBERTa, and More," Packt Publishing,2021.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter," pp. 2–6, 2019, [Online]. Available: http://arxiv.org/abs/1910.01108.

[22] K Yuki, M. Fujiogi, S. Koutsogiannaki. "COVID-19 pathophysiology: A review". Clin Immunol. 2020;215:108427. doi:10.1016/j.clim.2020.108427.

[23] https://elitedatascience.com/overfitting-in-machine-learning.

[24] X. Shi, Z. Guo, K. Li, Y. Liang, and X. Zhu, "Self-paced Resistance Learning against Overfitting on Noisy Labels," *Pattern Recognit.*, no. II, p. 109080, 2022. doi:10.1016/j.patcog.2022.109080.