ISSN : 2350-0743

www.ijramr.com

International Journal of Recent Advances in Multidisciplinary Research

Vol. 09, Issue 10, pp.8104-8107, October, 2022

RESEARCH ARTICLE

POINT OF VIEW ON TEXT CLASSIFICATION USING TRANSFORMER MODELS FOR TEXT DATA

Joyinee Dasgupta, Priyanka Kumari Mishra, Arpana Dipak Mahajan, Selvakuberan Karuppasamy and Subhashini Lakshminarayanan

Data and AI Capability, Accenture Technology Centre

ARTICLE INFO

Article History:

Received 28th July, 2022
Received in revised form
29th August, 2022
Accepted 17th September, 2022
Published online 30th October, 2022

Key words:

Text classification; Natural Language Processing; Transformer model; BERT; DistilBert; Roberta; XLNet; T5.

ABSTRACT

Text Classification is one of the most familiar use cases of NLP. The most common type of text classification problem includes spam identification, news text categorization, movie genre category prediction, sentiment analysis, etc. There can be a variety of use cases for every domain. The major disadvantages of the seq2seq model are we lose the dependency information, difficulty remembering the lengthy conversation, exploding gradient problems, etc while transformer-based models pay attention to the sequential words, as well as words far away from each other, their ability of learning, is more rigorous and better than seq2seq models giving higher prediction accuracy. This paper focuses on the multiple transformer pre-trained models that can be leveraged for text classification problems.

INTRODUCTION

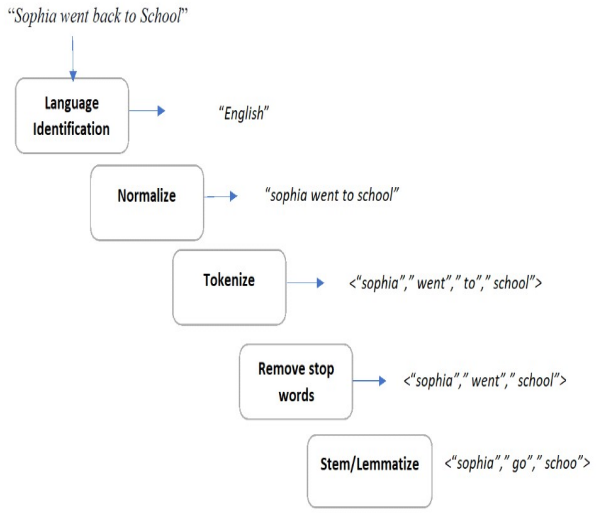
In recent years Natural Language Processing is one of the most evolving fields and have helped us in solving so many kinds of the problem across all industries. It includes text classification, text generation, text summarization, machine translation, chatbots, information retrieval systems, and so on. Text classification among this is a very problem we are dealing with in most industries. The process of text classification or any natural language processing includes data preparation which includes cleaning the data by removing noise from the data /unwanted characters. Another very Important task includes feature extraction to map the features to correct labels. Then the next step is selecting the suitable pretrained model and tuning it according to the data we have collected. In this paper, we have selected a few transformer-based models. There are numerous research studies that demonstrate how BERT model variations have changed and become more effective through time. It has been demonstrated that applying transfer learning, which uses pre-trained language models, and adjusting multiple parameters for a particular task can help predict the likelihood of narratives regarding the implementation of natural language processing applications. A transformer is a neural network that uses self-attention layers and encoder-decoders to convert input data into vector representations. BERT is one of the most well-known models that uses the transformer architecture and was used as a basic model for the development of the transformer model. BERT is an effective model for representing language. In(Rodrawangpai & Daungjaiboon, 2022), By adding layer normalisation and dropout layers to the pre-trained transformer model, Authors proposed a novel text categorization model. According to the authors, the model presented, produces better classification outcomes than utilising a transformer-based language by itself with unbalanced classes.

In (González et al., 2021), authors proposed TWiLBERT, a BERT architectural specialty for Text in Twitter and the Spanish language . The Next Sentence Prediction signal was also modified by the authors to learn the coherence between tweet pairings inside of Twitter chats. Several methods for enhancing Transformer models have recently been put forth. By altering the fundamental architecture of BERT, several techniques may be utilised to boost its performance (J. Li et al., 2022; W. Li et al., 2019). This study suggests a policy evaluation framework to extract optimized feature expression and assess the policy provided by the text more effectively and reliably. Based on the Bert (bidirectional encoder representations from transformers) paradigm, text categorization method. The approach in(Yu et al., 2022), represents the feature vector at the level of the sentence of the policy book using the Bert pre-training language model before feeding the resulting feature vector into the designated classifier for classification. The bidirectional transformer encoder structure serves as the foundation for the Bert pre training language model, which has a significant capacity for semantic expression. The model can learn the traits of policy text by using the policy domains data set gathered in this research for training. The experimental findings demonstrate that the suggested approach may successfully raise the task's classification accuracy for the set of policy textual data. To improve BERT models on the XMC issue, authors from (Chang et al., 2019) X-BERT, the first scalable approach. To develop label representations and create semantic label clusters to better characterise label dependencies, X-BERT specifically makes use of both the label and the input text. A method for fine-tuning BERT models to capture the contextual relationships between input text and the induced label clusters is at the core of X-BERT. Finally, resulted in a cutting-edge XMC technique, is produced by an ensemble of the several BERT models trained on diverse label clusters.

Using a hybrid model called BiGRU-MA, which can capture deep semantic features and address the issue of classifier performance degradation brought on by a lack of semantic information, this research (Jiang & Wang, 2022) combines bidirectional GRU with the self-attention mechanism. This article covers the modelling concepts, utilises text classification-related technologies to model, and discusses the technologies employed before comparing studies with current models to confirm the model's efficacy. Some professionals have blended the benefits of biLSTMs with CNNs(Cai et al., 2019; Hassan & Mahmood, 2017; Sameer & Gupta, 2022; Singh et al., 2019; Zhang & Rao, 2020). Based on the assumption that CNNs can learn feature points from temporal input but cannot learn sequential correlations, this hybrid of neural architectures was created. A biLSTM network succeeded by a CNN layer, however they are typically used in succession and text classification problems can also employ relation classification tasks(Ranjan et al., 2022; Varghese et al., 2020). *Corresponding author: Joyinee Dasgupta, Bachelor of Technology Electrical Engineering, Postgraduate Program of Business Analytics,Bangalore,10 years of Industrial Experience in Machine learning, Deep Learning and Business Analytics.

STEPS FOR MULTICLASS CLASSIFICATION

A multiclass classification includes data collection, data processing, data modelling, model evaluation, deployment.



Transformer Model: Transformer is a neural network-based model. Transformer in NLP is a novel architecture that is introduced to give the solution for sequence -to-sequence tasks once handling long-range dependencies with ease. The Transformer model was introduced in the paper (Vaswani et al., 2017) by A. Vaswani et al.

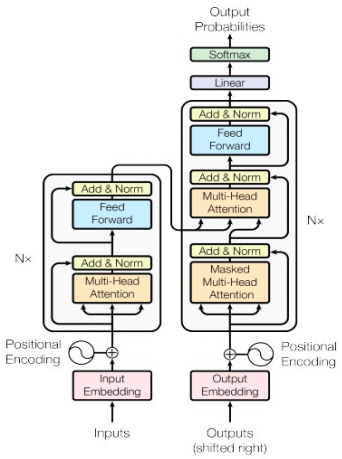


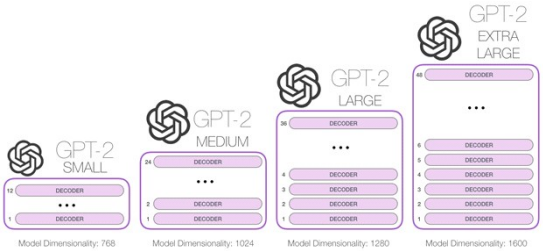
Figure 1. The Transformer – Model Architecture (Source:https://arxiv.org/abs/1706.03762)

Data Modelling: For effective data modelling we have used transformer models.

Figure 2. DataPre-processing Techniques

- GPT-2 Language Model
- DistilBert
- Roberta
- Xlnet
- T5 transformer

GPT-2 Language Model: GPT-2 is essentially a sophisticated Language Model that is based on Transformer Architecture and is trained on 40 Gigabyte of WebText. It was was introduced by (the first author - Alec Radford) in 2019.It helps us in language translation, document summarization, Question-Answering model . It is a collection of multiple decoder units on top of one another with some advanced NLP learning concepts like Residual Connections, Layer Normalization, multiple heads, Masked Self Attention etc. This model tries to essentially predict the next word in the given sequence having seen historical text data. It's a NLP model with position embeddings that is representation of the words with value with respect of its position on the sentence. Here right padding is done here mostly rather than left. It is trained on a very large English corpus with no labels and predicts next token in a sequence. More specifically, the variable to be predicted are in the same sequences as inputs, but with one token (word or character or group of words) moved to the right. Inputs are continuous text sequences of a specific length. To ensure that the predictions for the token I only consider the inputs from 1 to I and not the future tokens, the model internally employs a mask-mechanism. This model works best on the use cases where we need to generate text from prompt text.



(source: http://jalammar.github.io/illustrated-gpt2/)

Figure 3. Different GPT2 models

Distil Bert: It is a smaller, speedy, cheaper, light weighted algorithm. It is a refined version of BERT. It is a transformer-based model created by distilling BERT. It has cut down the parameters by 40% than the *bert-base-uncased* and is generally faster than the BERT base by 60%. It saves approximately over 95% of BERT's performances and it has been evaluated on the language understanding benchmark(GLUE). This model is only pretrained using masked language modelling and we are not using next sentence prediction algorithm. 16 GB of data from the Toronto book corpus plus data from Wikipedia were used to train this model. DistilBert uses a large batch size (400) with gradient accumulation during the training phase, with the accumulation being carried out locally utilising the gradients from several mini batches prior to updating the parameters in each step. Additionally, the training approach omits the learning objectives for segment embeddings and next sentence prediction (NSP). Dynamic masking is employed during inference in place of the static masking used in the BERT basic model (Ahmed & Anand Kumar, 2021). It has 6 encoder blocks. Pooling functionalities of Bert and token type embeddings are not used in BERT. Cosine distance loss method calculates how same are the hidden states of Bert and DistilBert. They use same language model loss. The distillation loss measures resemblance between the outputs of DistilBert and BERT. Below is the model architecture and components.

Table 1. Depicts a Comparison Study of Different Transformer Models

	Model	Parameters	Layers	Attention Heads	Vocabulary size	Main Properties
GPT2	Transformer unsupervised learning method	1.5B parameters	12	12	50k	transformer decoder block
DistilBert	Distilled Bert Model	66M parameters	6	12	30522	smaller, faster, cheaper, and lighter version of BERT
Roberta	Robustly optimized Bert model	125M parameters	6	12	50k	Changed tokenisation method, more epoch, more data, removed next sentence prediction method
Xlnet	autoregressive (AR) language model	340M parameters	24	16	32000	Permutation Language Model to process the sentences
T5	Transformer language model	340M parameters	6	8	32128	Encoder decoder model which is trained on multi task mixture of supervised and unsupervised tasks like machine translation, text summarisation etc .

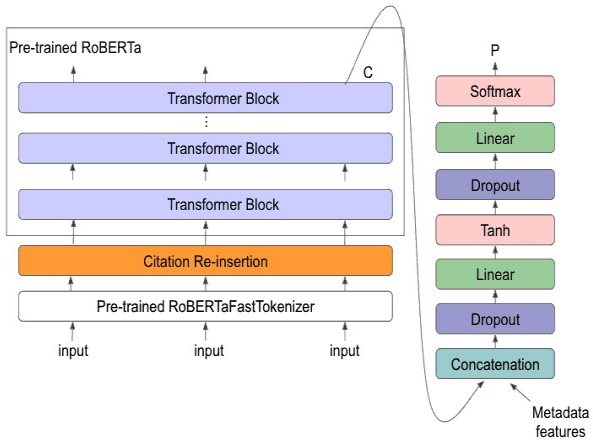


Figure 4. RoBERTa- A Robustly Optimized Architecture Reference 13

BERT Pretraining Approach: This model is a modified version of the BERT model. It included training in much larger mini-batches and learning rates and modified key hyperparameters, removed the next-sentence prediction method of training. It has a much similar architecture just like BERT. This model uses a byte-level BPE as a tokenizer and uses a completely different pretraining approach. Only mask language modelling is present here. It doesn't have token_type_ids. There is no need to indicate which token is associated with which segment. Need to differentiate the segments with the separation token. Batch size increase from 256 to 8000. Vocabulary size increase from 30k to 50k. It has dynamic masking pattern as compared to BERT. This helps in learning different patterns every time the language is fed to the model. Training data increased from 16GB to 160 GB. It is a increase of 10x. It comprises of English Wikipedia plus book corpus ,76 GB from common crawl news data, open source creation of web text dataset -38GB, a portion of Common Crawl data that has been filtered to correspond to Winograd schemas' narrative structure. (Huang et al., 2021)

XLNet: One of most recent model to come out of the rapidly growing field of Natural Language Processing is called XLNet (NLP).This model is proposed by Zhilin Yang et al. in (Yang et al., 2019). This auto-regressive language model, which is built on the transformer architecture with recurrence, outputs the joint probability of a series of tokens. It is a generalised autoregressive approach that takes advantage of the strongest aspects of both AE and AR language modelling while avoiding their drawbacks. First off, unlike traditional AR models, which use a fixed forward or backward factorization order, XLNet maximises the anticipated log likelihood of a sequence relative to all feasible factorization order permutations. The context for each place can include tokens from both the left and the right thanks to the permutation operation. It is anticipated that each position would learn to make use of contextual data from all positions, or to capture bidirectional context. Second, XLNet does not rely on data corruption because it is a generic AR language model. As a result, XLNet is not affected by the pretrain-finetune disparity that BERT is.

The independence assumption established in BERT is removed by the autoregressive objective, which also offers a simple approach to employ the product rule for factorising the joint probability of the projected tokens (Yang et al., 2019). This model is a modification of the transformer model (Extra-large) and is based on auto regressive model. The below steps to use the XLNet model. Perm_mask parameter input helps to control attention pattern during training and testing. Since it is challenging to train a fully auto-regressive model over a range of factorization orders, XLNet is pretrained utilising only a part of the output tokens as the variable to be predicted, which are selected using the target mapping input. This model is one of those rare models that does not have any sequence length limit. To implement XLNet for sequential decoding, use the target mapping inputs as well as perm_mask for controlling the attention span and outputs.

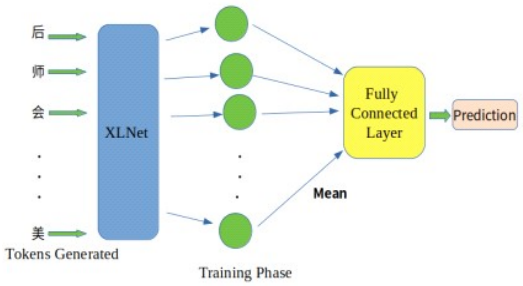
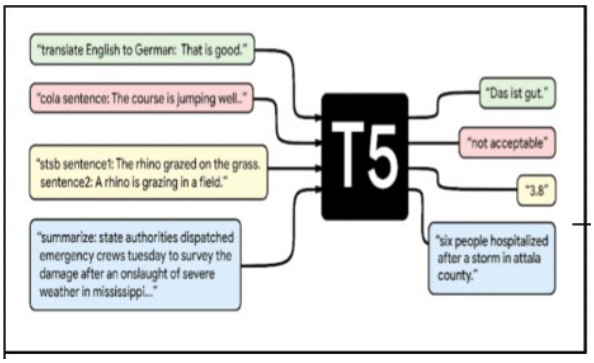


Figure 3. XLNet Architecture Reference 16



1. t5 – small is a language model with 60M parameter.
2. t5 – base model has 220M parameter to tune
3. t5 – large has 770M parameter to tune
4. t5 – 3b has 3billion parameter.
5. t5 – 11b has 11 billion parameters.

Figure 4. T5 Transformer

(Text-To-Text Transfer Transformer) model: T5 transformer is a text-to-text transformer model. We can achieve many tasks with this framework model like machine translation, regression task, classification task, (for example, if we want to predict how similar sentences are in a scale of 1-5), other sequence to sequence tasks like

summarizing a document (for example, summarising all the articles or documents from all the confluence page of an organisation /project .It has an encoder-decoder model architecture .It converts every problem to text ,It takes input of text and again generates text output. This can be used for multitasking because this output can again be input for some other tasks. It uses relative scalar embeddings. Padding can be done on the left side and on the right. This model has achieved SOTA for approximately 20 NLP tasks and researches concludes that it behaves like a normal human. in the below figure every task we take into consideration involves feeding text into a model that has been trained to produce some goal text. This enables us to use the same model, loss function, and hyperparameters to a variety of tasks, such as document summarization(blue), translation(green), linguistic etc.

CONCLUSION

These 5 transformer models discussed are widely used in multiple text classification problem or varied use cases of natural language processing, but it is not limited to only these models. The major solution to any use case is highly dependent on the data, data size and its quality. Also, likewise the accuracy of a text classification model depends on the data preparation and proper model tuning.

REFERENCES

Ahmed, S. S., & Anand Kumar, M. (2021). Classification of Censored Tweets in Chinese Language using XLNet. *NLP4IF 2021 - NLP for Internet Freedom: Censorship, Disinformation, and Propaganda, Proceedings of the 4th Workshop*, *September*, 136–139. https://doi.org/10.18653/v1/2021.nlp4if-1.21

Cai, J., Li, J., Li, W., & Wang, J. (2019). Deeplearning Model Used in Text Classification. *2018 15th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2018*, 123–126. https://doi.org/10.1109/ICCWAMTIP.2018.8632592

Chang, W.-C., Yu, H.-F., Zhong, K., Yang, Y., & Dhillon, I. (2019). *X-BERT: eXtreme Multi-label Text Classification with using Bidirectional Encoder Representations from Transformers*. 1–12. http://arxiv.org/abs/1905.02331

González, J. Á., Hurtado, L. F., & Pla, F. (2021). TWilBert: Pre-trained deep bidirectional transformers for Spanish Twitter. *Neurocomputing*, *426*(xxxx), 58–69. https://doi.org/10. 1016/ j.neucom. 2020.09.078

Hassan, A., & Mahmood, A. (2017). Efficient deep learning model for text classification based on recurrent and convolutional layers. *Proceedings - 16th IEEE International Conference on Machine Learning and Applications, ICMLA 2017*, *2017-Decem*, 1108–1113. https://doi.org/10.1109/ICMLA.2017.00009

Huang, Z., Low, C., Teng, M., Zhang, H., Ho, D. E., Krass, M. S., & Grabmair, M. (2021). Context-aware legal citation recommendation using deep learning. In *Proceedings of the 18th International Conference on Artificial Intelligence and Law, ICAIL 2021* (Vol. 1, Issue 1). Association for Computing Machinery. https://doi.org/10.1145/3462757.3466066

Jiang, T., & Wang, Z. (2022). *Text Classification Using BiGRU with Directional Self-Attention*. 394–397. https://doi.org/10.1109/ictech55460.2022.00085

Li, J., Lin, Y., Zhao, P., Liu, W., Cai, L., Sun, J., Zhao, L., Yang, Z., Song, H., Lv, H., & Wang, Z. (2022). Automatic text classification of actionable radiology reports of tinnitus patients using bidirectional encoder representations from transformer (BERT) and in-domain pre-training (IDPT). *BMC Medical Informatics and Decision Making*, *22*(1), 1–17. https://doi.org/10.1186/s12911-022-01946-y

Li, W., Gao, S., Zhou, H., Huang, Z., Zhang, K., & Li, W. (2019). The automatic text classification method based on bert and feature union. *Proceedings of the International Conference on Parallel and Distributed Systems - ICPADS*, *2019-Decem*, 774–777. https://doi.org/10.1109/ICPADS47876.2019.00114

Ranjan, P., Kumar, A., Kumar, R., Sameer, M., & Gupta, B. (2022). Automated Detection of Blood Pressure using CNN. *2022 IEEE Delhi Section Conference (DELCON)*, 1–5. https://doi.org/10.1109/DELCON54057.2022.9753601

Rodrawangpai, B., & Daungjaiboon, W. (2022). Improving text classification with transformers and layer normalization. *Machine Learning with Applications*, *10*(August), 100403. https://doi.org/10.1016/j.mlwa.2022.100403

Sameer, M., & Gupta, B. (2022). CNN based framework for detection of epileptic seizures. *Multimedia Tools and Applications*, *81*(12), 17057–17070. https://doi.org/10.1007/s11042-022-12702-9

Singh, G., Kumar, B., Gaur, L., & Tyagi, A. (2019). Comparison between Multinomial and Bernoulli Naïve Bayes for Text Classification. *2019 International Conference on Automation, Computational and Technology Management, ICACTM 2019*, 593–596. https://doi.org/10.1109/ICACTM.2019.8776800

Varghese, A., Agyeman-Badu, G., & Cawley, M. (2020). Deep learning in automated text classification: a case study using toxicological abstracts. *Environment Systems and Decisions*, *40*(4), 465–479. https://doi.org/10.1007/s10669-020-09763-2

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/ paper/2017/file/ 3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). XLNet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, *32*(NeurIPS), 1–11.

Yu, B., Deng, C., & Bu, L. (2022). *Policy Text Classification Algorithm Based on Bert*. 488–491. https://doi.org/10.1109/ictech55460.2022.00103

Zhang, Y., & Rao, Z. (2020). N-BiLSTM: BiLSTM with n-gram Features for Text Classification. *Proceedings of 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference, ITOEC 2020*, *Itoec*, 1056–1059. https://doi.org/10.1109/ITOEC49072.2020.9141692

****