International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 07 Issue: 06 | June - 2023 SJIF Rating: 8.176 ISSN: 2582-3930

News Categorization using Deep Learning Models – BERT and LSTM

Ankita Thombre[1], Vrishali Chakkarwar[2]

[1]M. Tech, Department of Computer Science and Engineering, Government College of Engineering, Aurangabad.
[2]Assistant Professor, Department of Computer Science and Engineering, Government College of Engineering, Aurangabad.

Abstract - In this digital era, there are various platforms that continuously generate news on various content. The news generated, range from environment like global issues to personal matters of celebrities and even that of common people. Platforms like Twitter and Instagram have facilitated in the increase of online content generation by the user, which in turn can be extracted as new trends in market, thus resulting in data essential for businesses. This chaos of news articles over the internet may lead to loss of significant data, reason being sheer amount of information and lack of enough resources to access it. Gaining insight from this chaos requires proper classification of news articles, into appropriate categories. Deep Learning methods tend to understand the structure and semantics of natural language, thus resulting in better classification of news text data. Applying transformer architecture to deep learning techniques can enhance the understanding of languages by machines. This work applies BERT and LSTM models for news text classification task.

Key Words: Text classification, news categorization, Bi-directional Encoder Representations from Transformers (BERT), Long Short–Term Memory (LSTM), deep learning, neural networks.

1.INTRODUCTION

Generation of text data is tremendous in today's age, with the event of social media and various digital platforms. These platforms enable users to share any event to general public. Any post, no matter how trivial can be marketed as news on such platforms. This has given rise to enormous amount of news text. In addition to these, the traditional news sources have created their own digital platform for wider reach and gaining larger user base. In order to gain insights from these texts, it needs to be organized in a proper manner.

News Categorization often referred to as news text classification is a text classification problem in natural language processing (NLP). It refers to assigning respective categories to the news article, depending on the context of the articles. Any news article can be classified into categories such as entertainment, business, sports, international, politics, science & technology, and many more. Such categorization of news articles provides ease of accessing the data. Moreover, deriving useful insights from a well-organized data is much easier and there is no probability of missing out on any article, as in case of an ill organized data.

Automation of text classification tasks like news categorization, however have major hurdles. For machines to accurately classify the news text, they need to understand the context of the text. The language semantics and proper usage of words are to be understood by the machine. Moreover, complex nature of languages present problems like homonymy and polysemy that are difficult to grasp for machines. Many machine learning algorithms have been applied, but the results achieved were barely satisfactory. These algorithms often focus on specific keywords rather than the underlying meaning of the text, thus performing poorly on unseen data.

This paper presents the study about performance of deep learning, neural network based models when applied on NLP tasks such as news categorization. Long Short–Term Memory (LSTM) model and Bi-directional Encoder Representations from Transformers (BERT) model are trained for news categorization task and their performance is compared to determine a better model for news categorization task.

2. Proposed Model

2.1 Related Work

In complications involving the classification of news material, deep learning approaches have made great progress and frequently outperform conventional machine learning methods. Automatic representation learning is a feature of deep learning models like recurrent neural networks (RNNs) and transformer-based models like BERT. They are particularly adept at handling the complex language patterns found in news articles because they can capture complex patterns, semantic relationships, and contextual information in an end-to-end manner without the need for explicit feature engineering. In contrast to standard machine learning techniques, which frequently rely on handmade features that call for domain knowledge and human interaction, they can effectively represent syntactic structures, long-range dependencies, and semantic links, allowing for a more sophisticated comprehension of the text. The representation of text data relies heavily on feature engineering, and the quality of the features has a direct impact on how well the machine learning models perform, frequently failing to capture complex linguistic connections and patterns. They heavily rely on the handcrafted elements' quality and appropriateness to depict the text [1].

Transformers are a sort of deep learning model or architecture developed by author Ashish Vaswani [2] that accentuate the attention mechanism, which enables them to recognize the linkages and dependencies among various words or tokens in a sequence without relying on recurring connections. The Self-attention operates on three matrices: Query, Key, and Value. It calculates attention scores by computing the dot product between the Query and Key matrices. These scores are then scaled and passed through a softmax function to obtain the attention weights. The weighted sum of the Value matrix, using the attention weights, produces the context vector for each word/token.

Positional encodings are incorporated into input words or tokens as learned embeddings. Encoding the relative or absolute position of each word or token in the sequence, embeddings are used to offer positional information to the model. The multi-head self-attention mechanism allows the model to attend to different parts of the sequence simultaneously, enhancing its ability to capture dependencies. The position-wise feed-forward network applies a fully connected feed-forward network to each position independently, allowing the model to capture non-linear relationships [2].

LSTM or Long Short-Term Memory made use of recurrent neural networks (RNNs) and overcame its issue of efficiently capturing and conveying long term dependencies in sequential data. For this, it made use of gating along with a memory cell to store information over long sequences, preventing the loss of important information. The input gate, forget gate, and output gate are gating mechanisms that control the flow of data into, out of, and inside the memory cell. These gates regulate which information is retained or discarded at each time step and are controlled by the sigmoid and tanh activation functions. These features make LSTM specialized in capturing long-term dependencies [3].

BERT or Bi-directional Encoder Representations from Transformers is a bidirectional model that can derive context of the selected token from a token before and after selected tokens. It means that BERT can associate a token or word with its correct context and meaning in one single iteration. BERT acquired this exceptional feature through pre-training tasks like Masked Language Model (MLM) in which it worked to predict the masked token and Next Sentence Prediction (NSP) in which it predicted whether the two sentences given to it were in continuation or not. This pretraining helped the model to learn the underlying context of the words and understand the ever-changing nature of languages [4].

Even though BERT is a pre-trained model, it is only familiar with the semantics of the languages. In order to apply BERT to NLP tasks, we need to incorporate the required domain knowledge to give an efficient performance on required task. This incorporation is often termed as fine-tuning the BERT for required tasks. [5] studied different BERT based model along with base BERT to determine the performance on text classification task. The BERT was incorporated with Sequence classification, bi-directional LSTM and Convolutional neural networks (CNN) to perform text classification on different dataset. The BERT base model outperformed other BERT models with additional features of bi-LSTM and CNN [5].

[6] used the Bayesian network and the BERT model to classify the text. The interdependence between the parameters or tokens in the input text was discovered by the Bayesain network, and this information was then given to the BERT model together with the text data. As a result, BERT developed a solid comprehension of the words, their context or meaning, and how each word was used in the sentence. As a result, this model also developed a high accuracy on unobserved data. The results revealed that when a Bayesian network and the Bert model were used combined, classification accuracy was 94.59%, which is 18.06% greater than when only the Bayesian network and only the Bert model were used [6].

Using deep learning techniques, [7] tackled the multi-label classification problem for classifying texts. On a dataset of toxic remarks, deep learning techniques were used to categorize the comments into the following categories: "toxic," "severely toxic," "obscene," "threat," "insult," and "identity hate." They included Baseline NN, RNN, CNN, LSTM, GRU, and Bidirectional-LSTM deep learning algorithms. The precision and F1-measure in our research outperformed the previously reported experiment in Bidirectional LSTM and GRU models with Glove and fastText pre-trained embedding by more than 8%, and in the other models with at least 5%. Long short-term memory (LSTM) outperforms nearly all other models in the four evaluation criteria, followed by CNN, bidirectional GRU (with fastText and Glove), and simple neural network (NN) models [7].

Deep Learning approaches like Doc2Vec model have also been implemented for categorization of news text [8]. A common architecture used widely is the Convolutional Neural Network (CNN) combined with the Long Short-Term Memory (LSTM) model. In the context of news categorization, the CNN layers can effectively capture local features and patterns within individual words or n-grams, while the LSTM layers can capture the contextual information and long-term dependencies across the entire sequence of words in a news article. By combining the strengths of CNNs and LSTMs, the model can learn informative representations of the news text for accurate categorization. [8] proposed a CNN + Doc2Vec model that achieved an accuracy of 94.17% as compared to the general CNN based architectures and machine learning techniques like Support vector machine (SVM) and such.

BERT was incorporated with Siamese and triplet network structure that can obtain meaningful sentence embeddings to make the model understand the relationships of the tokens and embeddings. This was the Sentence-BERT proposed that was capable of inspecting similarities in the text [9].

[10] applied the knowledge distillation method on BERT LM. It is a method for condensing a large model so that its knowledge can be transferred to a smaller one. To instruct a smaller model, like a student model, to imitate the actions of a larger model, like a teacher model, is an analogy. This training is primarily based on the training loss for the student model DistillBERT, which is a masked modelling loss with an extra cosine embedding loss. This student model was made from the teacher-BERT model by halving the number of layers in the original architecture. As a result, the normalization and linear layers of the BERT model were greatly optimized while alternate layers were taken into account during training. Then, using the IMDb dataset, DistillBERT was applied to similar tasks like question-answering tasks (SQuAD v1.1) and sentiment categorization. It performed as well as, if not better than, BERT's results for the same NLP tasks. It is noteworthy that DistillBERT still achieves a 97% accuracy in language processing despite having 40% fewer parameters than the BERT model. In comparison to $BERT_{BASE}$ and $BERT_{LARGE}$, DistillBERT was shown to be 60% faster [10].

The fine tuning of BERT for any specific NLP task requires a set procedure. [11] carried out extensive experimentation for deriving a general procedure for BERT fine tuning for text classification. It experimented with hyperparameters, catastrophic forgetting, in-domain and cross-domain further pre training and such methods to put forward a general fine-tuning technique [11].

2.2 Methodology

This study aims to find which deep learning model works best for the task of text classification with focus being on news text classification. To this aim, the pretrained BERT model is trained on news dataset to identify features and semantics of news articles accordingly to the categories they belong to. Similar work is done for training LSTM model and their results are then compared to verify an efficient model.

• LSTM based Classifier:

The LSTM model takes the pre-processed training dataset and it is passed to the defined embeddings layers. The model takes input in form of sequence of word embeddings. The semantic content and contextual information of words used in news stories are both captured by these embeddings, to comprehend word relationships and their applicability to the categorization of news task. The multiple hidden stacked layers allow the model to learn and represent complex patterns and relationships within the news articles and learn intricate news representations. The output layer then produces final predictions for news categories. As it is a multiclass classification, a softmax activation function is used to generate probability distribution over the news categories. The output is the category with highest probability.

• BERT based Classifier:

News articles that have been tokenized serve as the BERT model's input. These articles are divided into word-level tokens, and unique tokens are added to indicate the beginning and completion of the input by the original pretrained BERT itself, including [CLS] (classification) and [SEP] (separator). These tokens are then sent to the stack of transformer layers, each of which has a feed-forward neural network and a self-attention mechanism. The feed-forward neural network handles the attention outputs while the self-attention mechanism enables the model to collect contextual links between words. BERT then trains on the input data, using the knowledge gained from the MLM and NSP pre-training tasks. Performance is regulated by the fine-tuning layer, which consists of pooling, dropout, and a dense output layer. Dropout helps avoid overfitting as the pooling layer extracts a fixed-length representation from the BERT output. The recovered representation is mapped to the number of news categories for classification by the dense output layer.

2.3 Experimentations

The dataset used for news categorization is BBC-Text dataset. The dataset consists of five categories – politics, entertainment, business, technology, and sports.

Although the models used here are pretrained, a few preprocessing tasks are to be carried out. The tokenization and truncation task are needed to remove unnecessary characters and symbols, whereas encoding and vectorization methods converts the text data into numerical representations on which the neural networks work.

The dataset is first spilt into train, validation, and test dataset. The train dataset provides the model with the features and relationship of tokens it needs to learn. Once the model is trained, it may be the case that the entries of a certain news category are more in number than the other. This might result in overtraining of model for that category, which can impact the model accuracy in real time data. Thus, to avoid overfitting, the models are again trained on validation set for the purpose of optimization.

The proposed system architecture for news categorization is as shown below in the fig.1.

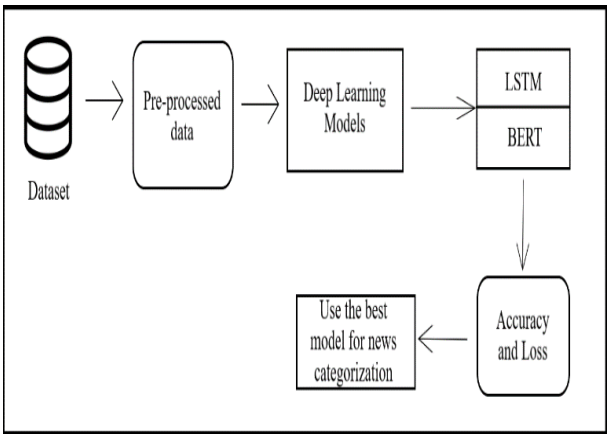

Fig -1: Proposed System Architecture.

Although BERT is a pre-trained model, for a specific task as news categorization, we need to train it in order to enable it to capture news related word and their relationship and applicability. The training dataset is supplied first to LSTM model and then to BERT model separately and efficiency is calculated on the training dataset. Then the validation dataset is used to fine tune the models during training phase by hyperparameter optimization. This increases the efficiency of the model. Finally, the test dataset is supplied to both models to determine which one performs better and gives higher accuracy.

The model with better results on these datasets is used to create a system for news article categorization that can be used by end-users.

2.4 Results

The performance of models is evaluated through accuracy and loss measures on training as well as validation data. The results of the same are depicted in table.

BERT performs exceptionally well as compared to LSTM model. The bi-directionality along with pretraining tasks of masked language model, next sentence prediction and question answering gives BERT an over various natural language processing task.

Table -1: Classification Accuracy acquired by both models.

Accuracy	BERT	LSTM
Training	99.50 %	91.84 %
Validation	99.50 %	88.06 %

Table -2: Classification Loss incurred by both models.

Loss	BERT	LSTM
Training	3.8 %	32.7 %
Validation	3.3 %	22.6 %

As shown in table-1, BERT acquires an accuracy of 99.5% on training as well as on validation dataset. This conveys that BERT will give a near accurate classification on unseen data as well. LSTM, accomplishes good results on training dataset, however performs poorly on validation data. Moreover, the loss incurred by the LSTM model is far higher than that of BERT, which suffered a minimal loss of 3.3%, thus retaining almost all semantics, context, and relationship of data. As a result, BERT preforms excellently on test dataset with an accuracy of 97.8 % whereas LSTM acquires an accuracy of 81.6 % on test dataset.

3. CONCLUSIONS

The study explores how models like BERT and LSTM work with language semantics and context when dealing with NLP tasks. We learn that language representations change as per the NLP tasks, thus news categorization tasks require a focused training of both the models. The results show that BERT is superior than a simple LSTM for the task of news text classification or news categorization. The bi-directional ability of BERT enables it to understand the correct context of the news articles, hence the higher performance. Thus, we conclude that BERT when used on real world unseen data will give better results and classify almost accurately in case of news categorization task.

ACKNOWLEDGEMENT

I would like to express my gratitude to my guide and all other faculty for their suggestions and guidance without which it would not be possible for me to complete this work. I deeply thank all anonymous reviewers for their valuable time they put in for reviewing our manuscript. I would also like to thank all researchers who shared their work publicly that turned out to be very helpful resources for this work.

REFERENCES

1. Suneera C M, Jay Prakash: “Performance Analysis of Machine Learning and Deep Learning Models for Text Classification”,2020 IEEE 17th India Council International Conference (INDICON).
2. Ashish Vaswani, Noam Shazeer,Niki Parmar,Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, Illia Polosukhin: "Attention Is All You Need" arXiv:1706.03762v5 [cs.CL] 6 Dec 2017.
3. Sepp Hochreiter, Jurgen Schmidhuber: “Long Short-Term Memory”, 1997.
4. Jacob Devlin, Ming-Wei Cheng, Kenton Lee, Kristina Tautonova: “BERT: Pre-training of deep bidirectional transformers for language understanding” in Proc. Annu. Conf. North Ameri. Chapter Assoc. Comp. Linguistics, Hum. Lang. Technol., 2019, pp- 41714186.
5. Samin Mohammadi, Mathieu Chapon: “Investigating the Performance of Fine-tuned Text Classification Models Based-on Bert” in 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS).
6. Songsong Liu, Haijun Tao, Shiling Feng: "Text Classification Research Based on Bert Model and Bayesian Network”.
7. Hamza Haruna Mohammed, Erdogan Dodgu, Abdul Kadir Gorur, Roya Choupani: "Multi-Label Classification of Text Documents Using Deep Learning", 2020 IEEE International Conference on Big Data (Big Data).
8. Hasibe Busra Dogru, Sahra Tilki, Akhtar Jamil, Alaa Ali Hameed: "Deep Learning-Based Classification of News Texts Using Doc2Vec Model" 2021 1st International Conference on Artificial Intelligence and Data Analytics (CAIDA).
9. Nelis Reimers, Iryna Gurevych: “SentenceBERT: Sentence Embeddings using Siamese BERT-Networks.”
10. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf: “DistilBERT, a dis-tilled version of BERT: smaller, faster, cheaper and lighter”.
11. C. Sun, X. Qiu, Y. Xu, and X. Huang, “How to fine-tune bert for text classification?” in China National Conference on Chinese Computational Linguistics. Springer, 2019, pp. 194-206