Comparative Analysis of Deep learning-based News Topic Classification Models

Seunghoi Kim
kim.kim.17@ucl.ac.uk

Byoung Hun Min
brian.min.15@ucl.ac.uk

Minjae Kang
minjae.kang.16@ucl.ac.uk

Pavlos Demetriou
pavlos.demetriou.20@ucl.ac.uk

Abstract

The development of technology has allowed for vast amounts of data to be available online and this increased the needs of making these data structured and categorised. Automatic news classification is the task of categorising news article according to their topics. This paper evaluates and compares some of the most widely used deep learning models for classification tasks - vanilla RNN, LSTM, GRU, Bi-GRU, and fine-tuned BERT with a linear classification layer. These models are trained, validated, and tested on the BBC News dataset consisting of five categories: business, entertainment, politics, sport, and technology. The performance of each model is evaluated by analysing five evaluation metrics: confusion matrix, accuracy, precision, recall and F1-Score. The results show that the fine-tuned BERT model significantly outperforms other RNN-based models. BiGRU came close with BERT in terms of performance but has 25 times faster inference times and significantly less number of parameters.

1 Introduction

As technology develops and vast amounts of data are digitalised, there are uncountable amounts of text data from a variety of sources. This volume of text is an inestimable source of information and knowledge that are in unstructured forms (Shah et al., 2020). Therefore, it is important to convert these data into structured form and acquire its meaning. Automatic text classification is a classical NLP problem that involves assigning documents to a class and is used in a wide range of applications such as spam detection, news topic classification, and sentiment analysis. This paper will analyse multi-label text classification techniques under the context of news topic classification which has significant use cases such as personalised news recommendations (De Clercq et al., 2020). The focus will be on deep learning-based classification which involves leveraging the content of the article to gain a deep textual understanding of the news and assign it a topic from a list of pre-defined classes.

1.1 Traditional Machine Learning

Before the rise of deep neural networks, classical approaches existed to deal with the problem of text classification that extracted hand-crafted features using the technique of *Bag of Words*, and then fed the extracted features into a classifier such as a *Support Vector Machine* (SVM) (Shervin Minaee and Gao., 2021). However, these methods require high-level engineering and computation.

Traditional machine learning techniques are generally divided into 2 steps (Chi Sun, 2019):

1. Manual feature extraction
2. Feeding extracted features to a classifier to make a prediction

For the manual feature extraction, a bag of words (BoW) is commonly used where the occurrence of each word is used as a feature for training a classifier. An extension to BoW is n-grams which look at the frequency of a group of n number of words and therefore capture more meaningful features.

As for classifiers, there are several algorithms: Naive Bayes, SVM, gradient boosting trees, and random forest.

Although these traditional approaches have some advantages of needing relatively less training data to gives a decent performance, they possess some limitations (Chi Sun, 2019). Firstly, the manual feature extraction procedure requires tedious and high-level feature engineering and analysis to acquire decent performance. Furthermore, the manual feature extraction is domain-specific and hence difficult to generalise to new tasks. Unlike deep

learning-based feature extraction(pre-trained models), these models cannot utilise large amounts of training data due to pre-defined features. Lastly, in contrast to contextual embedding such as BERT, manual feature extraction such as BoW (any n-grams) are not robust to contexts.

1.2 Pre-trained Embeddings

Some of the most popular pre-trained embeddings are the *word2vec* which was created by Google in 2013 and trained on 6 billion words and GloVe which maps the words into an embedding space and the distance between words represents similarity. The embeddings are responsible to create vector representation of words such that similar words share a similar representation (Qiu, 2020). One disadvantage of the *word2vec* models is that they do not capture any contextual information and this problem was solved with the model *ELMo* which came in 2018. In order to perform even better in the classification task, one intermediate but very useful step is the contextual representation of the document. We introduce BERT which is a pre-trained model created by Google on Masked Language Model Task and Next Sentence Prediction Task using a large corpus. It achieves state-of-the-art performance on various NLP tasks using only some fine-tuning. Another state-of-the-art model that is currently being used for a various NLP tasks is the Generative Pre-trained Transformer (GPT) model and its later improvements GPT-2 and GPT-3 (Shervin Minaee and Gao., 2021).

1.3 Aims

The primary aims of this study is to conduct a comparative analysis of five deep learning-based techniques: RNN, GRU, LSTM, BiGRU, and Fine-tuned BERT for the task of news topic classification. The performance of the different techniques are evaluated and compared to reveal their strengths and weaknesses.

2 Related Work

This section will review similar studies on automatic text classification in the context of news articles and identify the gaps in the current state of research.

There are several comparative studies of news classification with traditional machine learning-based classifiers trained and evaluated on the BBC news dataset. Hussain et al. (Hussain et al., 2020) presented a comparative analysis of Naive Bayes, Support Vector Machine (SVM), Random Forest (RF), and K-Nearest Neighbours (KNN) classifiers by evaluating their performance on the BBC News Dataset with features extracted using TF-IDF. The classifiers were evaluated on 20% of the total dataset (80% used for training) corresponding to 298 news samples using metrics including confusion matrix, accuracy, precision, recall, and F1-score. The study found the SVM model to be the best performer with 98.3% classification accuracy.

Similarly, Shah et al. (Shah et al., 2020) carried out a comparative analysis of Logistic Regression, Random Forest, and K-Nearest Neighbours with TF-IDF feature representation and evaluated the classifiers on the 25% of the BBC News Dataset (85% used for training) with five evaluation metrics namely precision, accuracy, F1-score, support and confusion matrix. The highest performing classifier was found to be the Logistic Regression model with an accuracy of 97%.

Both studies compare and evaluate the performance of traditional machine learning-based classification models using TF-IDF feature representations on the BBC News Dataset and demonstrate high performances in terms of classification accuracy.

Similar studies on traditional machine learning-based models have also been carried out on different news article datasets. Katari and Myneni (Katari and Myneni, 2020) conducted an empirical study by implementing and evaluating Naive Bayes, SVM, and Artificial Neural Network models on Azerbaijani news articles which consists of 150000 labeled articles. Moreover, Suleymanov et al. (Suleymanov et al., 2018) reviewed various machine learning-based classifiers on numerous datasets including Fudan University news (Chinese), Azerbaijani news articles and a corpus of Bahasa Indonesian news portals.

Some previous research (Lindén et al., 2018) (Shahi and Pant, 2018) analysed the performance of deep neural network models on news classification. However, these studies compare DNN classifiers with traditional ML classifiers rather than comparing and evaluating the different DNN classifiers. Moreover, the depth and complexity of the DNN models used was generally low. More importantly, the authors rely on either manual feature extraction techniques such as TF-IDF or non-contextual word

embeddings such as word2vec which are not state of the art approaches.

As shown by reviewing numerous related work, most of the research on news classification focuses on traditional machine learning methods such as Naive Bayes and SVM. Although there exists some comparative analysis research on deep learning-based classifiers, they do not implement state of the art methods. Deep learning approaches have shown a significant improvement in many NLP tasks and are able to achieve state of the art performance due to its contextual feature extraction and therefore it is important to analyse and compare deep learning models in the context of news classification. This paper aims to address the gap in current research by carrying out a comparative analysis of news classification using deep learning-based models in conjunction with state of the art approaches.

3 Methodology

In this section we describe the five deep learning-based models which we implement and evaluate for news classification task.

3.1 RNN

A fundamental and basic neural network that is well known for its use in NLP tasks is *RNN* (Recurrent Neural Network). They are able to detect word dependencies in text and store information about the sequence. Its *cell* can be seen at 2. Unfortunately, RNNs have some disadvantages that can cause problems during training which results in unstable and non-optimal performance. Due to its nature, RNNs are very computationally slow and they suffer from *vanishing gradients* and *exploding gradients* for long data sequence. Due to the chain rule during training with *Back Propagation through time* (BBTT), the longer the data sequence is, the gradient tends to zero which causes the effect of *vanishing gradients*. As a result, the weights of the model will not be updated and the model will not be able to learn long term dependencies (Sherstinsky, 2020).

3.1.1 LSTM

As mentioned previously, vanilla RNN has some limitations so a variant called *LSTM* is used to address these problems. LSTM stands for Long short-term memory which has three gates. First, the *Forget Gate* which i responsible to decide which information to keep from the cell state and which to ”forget” (remove). Next is the *Input Gate* which decides what new information to store inside the cell state. The last gate is the *Output Gate* which controls the output of the hidden state that will be fed inside the next LSTM cell. This variant of LSTM better captures the limitations of vanilla RNN on long dependencies in a text. LSTM using addresses the problem of vanishing/exploding gradients by introducing the three gates stated before and a memory cell to remember information over a long period of time (Shervin Minaee and Gao., 2021). The disadvantage of LSTM models is that they are computationally expensive, due to its high complexity using the three gates states above.

3.1.2 GRU

A more efficient variant of LSTM is the *Gated Recurrent Unit* (GRU). The reason it is more efficient than LSTM is because it uses two gates instead of three, the *update gate* and the *reset gate*. GRU is more computationally efficient and needs less computing power than LSTM and it is still as powerful as an LSTM. In the GRU setting the update gate generalises the purpose of the input and the forget forget gate of an LSTM (Gruber and Jockisch, 2020). However, GRU’s also have some disadvantages such as slow convergence rate and low learning efficiency. This has as a result a longer training time and the possibility of under-fitting (Xin Wang, 2019).

3.2 BiGRU

The clear disadvantage of non-bidirectional RNN is that the information from the early time step gets faded away, resulting loss of information. However, tasks like text classification have to be able to see global features equally. Hence, we used Bidirectional Gated Recurrent unit for GRU. Bidirectional models are very popular because by learning representations from both the past and the future helps the model to capture useful contextual meaning of the input text. In addition, this will eliminate any possible misconceptions that might occur by learning only from one direction (Tran et al., 2019).

3.2.1 BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is a language representation model proposed by Devlin et al. (Devlin et al., 2018). The BERT model uses a multilayer bidirectional Transformer encoder and is comprised of two steps: *pre-training* and *fine-tuning* (González-Carvajal and Garrido-Merchán, 2020).

Figure 1: Model Architecture for RNNs

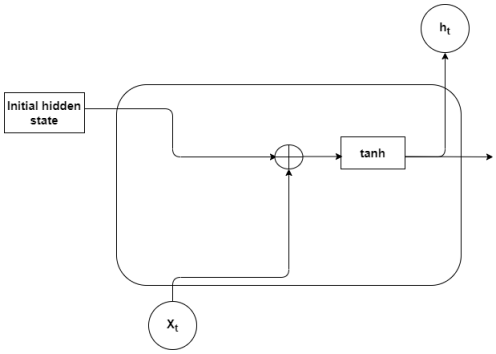

Figure 2: RNN

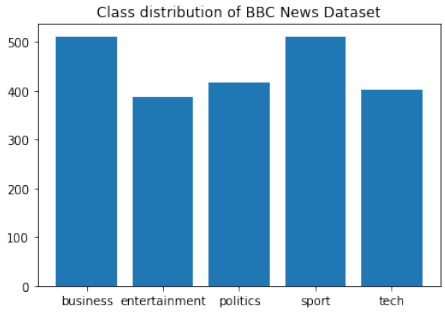

Figure 3: Class distribution of BBC News Dataset

The pre-training step involves training the model on a large unlabelled dataset over different pre-training tasks. The fine-tuning step makes use of transfer learning and initialises the BERT model with the pre-trained parameters. This model is then fine-tuned using task-specific labelled data. With this approach, BERT is able to achieve state of the art performance in many tasks (Devlin et al., 2018) with less data and computation time than traditional models. Finally, a single linear layer is added on top for classification.

4 Experiment

We investigate the different techniques: RNN, LSTM, GRU, BiGRU, BERT for the task of news topic classification.

4.1 Dataset

The dataset we use to train, validate and test the different models is the BBC News Dataset (Greene and Cunningham, 2006), provided for bench marking use in machine learning research, which consists of 2225 documents from the BBC News website from 2004-2005. The documents are labelled under the following five categories: business, entertainment, politics, sport, tech. The class distribution is shown by Figure 3.

4.1.1 Pre-processing

We perform basic pre-processing by concatenating the article title and body text and converting them to lowercase. The text is also truncated to ensure that it does not exceed a maximum of 200 tokens. Moreover, the categories are converted into integer labels numbered from 0 to 4. We have not applied rigorous pre-processing techniques such as stop word removal, stemming, and lemmatisation because the purpose of our study is to compare the different classification models rather than trying to achieve the highest performance.

4.1.2 Data Split

From the full dataset of 2225 articles, 80% is used for training, 10% is used for validation and the remaining 10% which corresponds to 225 samples is used for testing and evaluation.

4.2 Environment

We used Google Colab with K80 GPU for training and inferencing models and used Pytorch, Hugging Face Transformers and torchtext for model implementations, which contain various pre-trained word embeddings and language models such as BERT.

4.3 RNNs

The RNN models are trained with Adam Optimizer with learning late 1e-4. Since it is a classification task, we chose our objective function as cross entropy function. However, we've noted that there was a slight class imbalance in the dataset. To resolve class imbalance issue and train much robustly, we used Focal Loss proposed by (Lin et al., 2017) for BiGRU.

$$\mathcal{L}_{focal} = -\sum_{i=1}^{n} \alpha_i (1 - q_i)^{\gamma} \mathcal{L}_{ce}$$

where $\mathcal{L}_{ce}$ is cross entropy, α and γ are hyper-parameters for class weights and penalty on hard examples respectively. During our experiment, we have set α as [0.8,1.0,1.0,0.8,1.0] and γ as 2. Furthermore, unlike other RNN models, we used GloVe (Global Vectors for Word Representation) pre-trained word embeddings. This provided a boost in classification accuracy as it allowed the model to focus more on classifying difficult examples and minority classes correctly compared to a standard cross entropy loss.

4.4 Fine-tuned BERT

The BERT-base model (Devlin et al., 2018) with 12 transformer blocks, a hidden size of 768, 12 self-attention heads, and a total of 110M parameters was fine-tuned with the Adam Optimizer using the following hyper parameters.

Epochs	Batch Size	Learning Rate
4	16	2e-5

The optimal hyper parameter values were selected based on an exhaustive search from a set of values that were found to work well for fine-tuning the model (Devlin et al., 2018).

5 Results and Discussion

As explained in the methodology section, we evaluated various RNN models including Vanilla RNN, GRU, LSTM, BiGRU and a fine-tuned BERT model with a linear classification layer. Table 1 shows that the fine-tuned BERT model achieved an impressive 98.2% classification accuracy on the unseen test set which clearly demonstrates highest performance on the dataset out of other comparative models. It is also notable to see that BiGRU with Focal loss achieves sub-par results compared to BERT and outperforms other RNNs by large margins.

Table 2 shows the number of parameters for each model. Since GRU and LSTM have more number of gates, they have much larger number of parameters compared to vanilla RNN. Out of all models, BERT has the highest number of parameters, around 25 times larger than other models. On the other hand, BiGRU with Focal loss only leads to small increase in the number of parameters compared to non-bidirectional RNNs while outperforming them significantly.

Figure 9 shows the validation loss for each model. Due to early stopping function, each model has a different number of steps. As can be seen, RNN has the highest loss out of all models and is also overfitted relatively quickly. GRU and LSTM are both non-bidirectional and it shows GRU is slightly better than LSTM, which shows that GRU is more efficient and effective compared to LSTM. Hence, we implemented BiGRU with Focal Loss, which resulted in the lowest validation loss and converged much quicker compared to other RNN models. This proves that bi-direction in RNN gives richer features by emphasizing the latter information too. Although BERT has slightly higher loss compared to BiGRU in the figure, it is not possible to directly compare as BERT was fine-tuned for only 4 epochs.

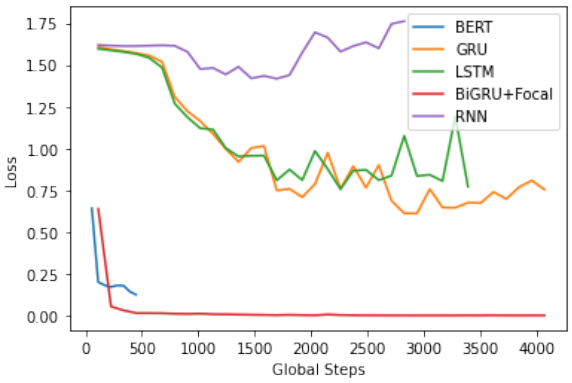

Figure 4: Validation loss

6 Conclusion

In this research report, we introduced various current text classification techniques ranging from vanilla RNN to fine-tuned BERT and provided a qualitative and quantitative analysis of the models on the BBC News dataset. The results have shown that BERT significantly outperforms every other RNN model. While BERT achieved the highest performance, it also had the longest inference time and the largest number of parameters. This may not be ideal in many practical applications that require high performance and real-time processing in low-powered devices. This gives BiGRU an advantage over BERT as it achieved sub-par results with only about $\frac{1}{6175}$ number of parameters and 25 times faster inference compared to BERT. Therefore, BERT may be preferred in situations that require high performance in high-powered devices; whereas, BiGRU may be preferred in low-powered devices. Although we have conducted an intensive comparative analysis

Models	Accuracy	Precision	Recall	F1-score
Vanilla RNN	0.462	0.445	0.446	0.444
GRU	0.839	0.830	0.829	0.829
LSTM	0.798	0.793	0.793	0.800
BiGRU	0.969	0.967	0.966	0.966
Fine-tuned BERT	0.986	0.988	0.984	0.986

Table 1: Test results

Models	No. of parameters
Vanilla RNN	12695
GRU	14743
LSTM	15767
BiGRU	17815
Fine-tuned BERT	110M

Table 2: Number of parameters

Models	Inference time (ms)
Vanilla RNN	4.33
GRU	4.80
LSTM	4.66
BiGRU	7.95
Fine-tuned BERT	195.90

Table 3: Inference time

on various models, our research does not include analysis on current state-of-the-art networks such as XLNet (Yang et al., 2019). It is also arguable that the dataset we used in this research may be considered a relatively easy task. For future work, we could include more state of the art networks and evaluate using a more variety of datasets with varying sizes and difficulties for a more robust evaluation.

References

Yige Xu Xuanjing Huang Chi Sun, Xipeng Qiu. 2019. How to fine-tune bert for text classification? page 2.

Orphée De Clercq, Luna De Bruyne, and Véronique Hoste. 2020. News topic classification as a first step towards diverse news recommendation. *Computational Linguistics in the Netherlands Journal*, 10:37–55.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Santiago González-Carvajal and Eduardo C Garrido-Merchán. 2020. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proc. 23rd International Conference on Machine learning (ICML'06)*, pages 377–384. ACM Press.

Nicole Gruber and Alfred Jockisch. 2020. Are gru cells more specific and lstm cells more sensitive in motive classification of text? *Frontiers in Artificial Intelligence*, 3(40):1–6.

Arif Hussain, Gulsher Ali, Faheem Akhtar, Zahid Hussain Khand, and Asif Ali. 2020. Design and analysis of news category predictor. *Engineering, Technology & Applied Science Research*, 10(5):6380–6385.

Rohan Katari and Madhu Bala Myneni. 2020. A survey on news classification techniques. In *2020 International Conference on Computer Science, Engineering and Applications (ICCSEA)*, pages 1–5. IEEE.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2999–3007.

Johannes Lindén, Stefan Forsström, and Tingting Zhang. 2018. Evaluating combinations of classification algorithms and paragraph vectors for news article classification. In *2018 Federated Conference on Computer Science and Information Systems (FedCSIS)*, pages 489–495. IEEE.

Yige Xu Yunfan Shao Ning Dai Xuanjing Huang Qiu, Tianxiang Sun. 2020. Pre-trained models for natural language processing: A survey. *arXiv preprint arXiv:2003.08271*.

Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. 2020. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5(1):1–16.

Tej Bahadur Shahi and Ashok Kumar Pant. 2018. Nepali news classification using naïve bayes, support vector machines and neural networks. In *2018 International Conference on Communication Information and Computing Technology (ICCICT)*, pages 1–5. IEEE.

Alex Sherstinsky. 2020. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306.

Erik Cambria Narjes Nikzad Meysam Chenaghlu Shervin Minaee, Nal Kalchbrenner and Jianfeng Gao. 2021. Deep learning–based text classification: A comprehensive review. page 40.

Umid Suleymanov, Samir Rustamov, Murad Zulfugarov, Orkhan Orujov, Nadir Musayev, and Azar Alizade. 2018. Empirical study of online news classification using machine learning approaches. In *2018 IEEE 12th International Conference on Application of Information and Communication Technologies (AICT)*, pages 1–6. IEEE.

Trang Uyen Tran, Ha Thanh Thi Hoang, and Hiep Xuan Huynh. 2019. Aspect extraction with bidirectional gru and crf. In *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, pages 1–5.

Wei Shi Jiarui Liu Xin Wang, Jiabing Xu. 2019. An optimized gated recurrent unit neural network. *Journal of Physics Conference Series 1325:012089*.

Z. Yang, Zihang Dai, Yiming Yang, J. Carbonell, R. Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *NeurIPS*.

7 Supplementary Material

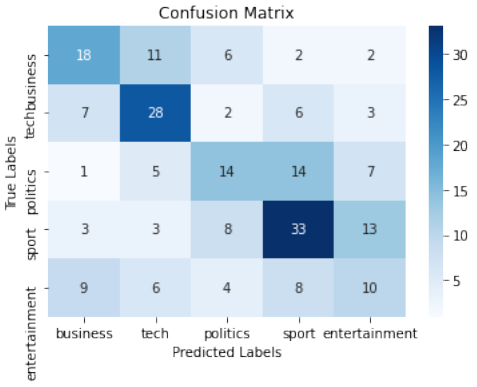

Figure 5: Confusion matrix: RNN

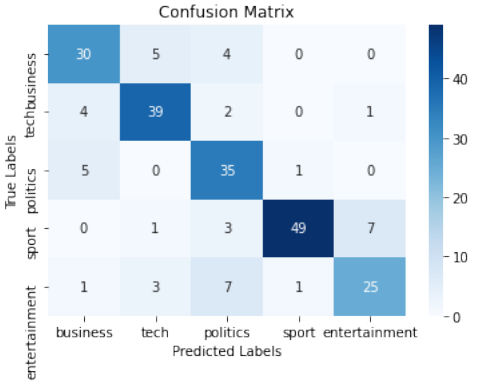

Figure 6: Confusion matrix: LSTM

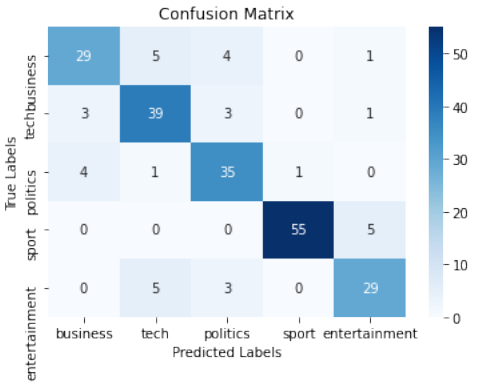

Figure 7: Confusion matrix: GRU

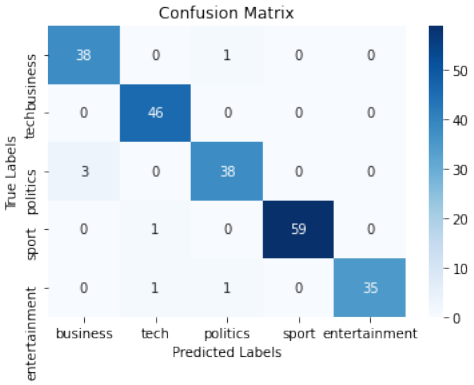

Figure 8: Confusion matrix: BiGRU+Focal

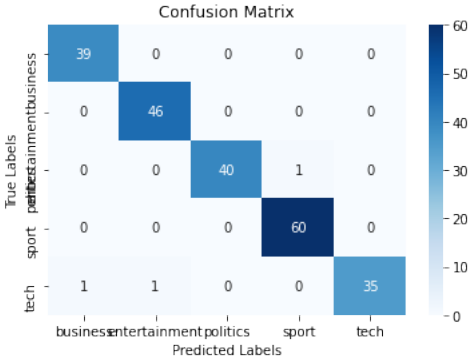

Figure 9: Confusion matrix: Fine-tuned BERT