The goal of this assignment is to develop a text classification application on TTC4900 News dataset [1]. The preprocessing scheme should involve tokenization and lemmatization phases as described in [2]. The feature extraction part of the project will utilize Bag of Words (BoW), TF-IDF, and Word2Vec, as detailed in [2]. Then, you will implement at least four traditional machine learning algorithms for classification (At least one ensemble learning method, such as Random Forest and at least two boosting algorithms, such as XGBoost and LightGBM), along with an artificial neural network (ANN) similar to [3].

In the second phase, you will implement a complex neural network architecture (ensemble learning) that combines CNN, LSTM, BiLSTM and/or GRU, following the implementation in [3], using the randomly initialized embeddings in the sample code and Word2Vec word embeddings.

Lastly, the transformer part will cover using the multilingual [4] and fine-tuned BERT model with the TTC4900 dataset [5]. Technical details of the fine-tuned model can be accessed from [6] if needed.

The output of each model will be presented in terms of accuracy, precision, recall, F1-score and AUC (ROC curve) results as well as presenting confusion matrix by expressing the accuracy and loss values. Finally, you will identify the model with the best accuracy, precision, and recall values and report your findings and insights in a detailed text document.

[1] https://www.kaggle.com/datasets/savasy/ttc4900

[2] https://www.kaggle.com/code/alperenclk/for-beginner-nlp-and-word2vec

[3] https://www.kaggle.com/code/erdal002/turkish-text-classification

[4] https://www.kaggle.com/code/ayhanc/bert-multilingual-for-turkish-text-classification

[5] https://huggingface.co/savasy/bert-turkish-text-classification

[6] https://arxiv.org/pdf/2401.17396

Hermodel için bu çıktıları da vermelisin
CNN + LSTM Modeli
------with randomly nitialized embeddings
Classifiaction report
accuracy, precision, recall, F1-score
Conf heatmap matrix
AUC/Roc Grafiği
Train-test loss grafiği
train -tes Acc grafiği

-------- gibi her model için böyle bir çıktı istiyorum
*En altta da her modelin F1 ve aCc değeri bunların grafikleri olmalı

1. **Random Vectorizer uygulanmış**:
   - 1. CNN (Random Vectorizer)
   - 2. LSTM (Random Vectorizer)
   - 3. BiLSTM (Random Vectorizer)
   - 4. GRU (Random Vectorizer)
   - 5. Ensemble Model (Random Vectorizer)
   - 6. Multilingual Transformer Model (Random Vectorizer)
   - 7. Fine-Tuned BERT Model (Random Vectorizer)

2. **Word2Vec uygulanmış**:
   - 8. CNN (Word2Vec)
   - 9. LSTM (Word2Vec)
   - 10. BiLSTM (Word2Vec)
   - 11. GRU (Word2Vec)
   - 12. Ensemble Model (Word2Vec)
   - 13. Multilingual Transformer Model (Word2Vec)
   - 14. Fine-Tuned BERT Model (Word2Vec)

3. **Bag of Words**:
   - 15. Random Forest (Bag of Words)
   - 16. SVM (Bag of Words)
   - 17. LightGBM (Bag of Words)
   - 18. XGBoost (Bag of Words)
   - 19. ANN (Bag of Words)

4. **TF-IDF**:
   - 20. Random Forest (TF-IDF)
   - 21. SVM (TF-IDF)
   - 22. LightGBM (TF-IDF)
   - 23. XGBoost (TF-IDF)
   - 24. ANN (TF-IDF)

5. **Word2Vec**:
   - 25. Random Forest (Word2Vec)
   - 26. SVM (Word2Vec)
   - 27. LightGBM (Word2Vec)
   - 28. XGBoost (Word2Vec)
   - 29. ANN (Word2Vec)

Toplam Model Sayısı: 29

Tekli Modeller:
CNN (Convolutional Neural Network)
LSTM (Long Short-Term Memory)
BiLSTM (Bidirectional LSTM)
GRU (Gated Recurrent Unit)
Çiftli Kombinasyonlar:
CNN + LSTM
CNN + BiLSTM
CNN + GRU
LSTM + BiLSTM
LSTM + GRU
BiLSTM + GRU
Üçlü Kombinasyonlar:
CNN + LSTM + BiLSTM
CNN + LSTM + GRU
CNN + BiLSTM + GRU
LSTM + BiLSTM + GRU
Dörtlü Kombinasyon:
CNN + LSTM + BiLSTM + GRU