CS 445 Natural Language Processing

Project 3: Text Classification V.2

Due Date: January 3, 23:55

This is the final version of the project. Please make sure to do everything described here.

In this project, you are expected to develop a text classification system for Turkish news articles. You will experiment with the following classification approaches:

- Naïve Bayes
- Logistic Regression
- CNN

For these different models, you will try different hyper-parameters. You will write your findings, results and interpretations into a report and submit that as well.

Dataset:

Dataset provided to you is part of the SUDer [1] data collection. It consists of news articles from Cumhuriyet newspaper. The documents (news articles) were labeled based on their categories as spor, siyaset, ekonomi etc.

You are provided with a train and test split. You are expected to create your own development (validation) split for hyper-parameter tuning. The distribution of the provided dataset with respect to their categories is presented in the following table.

	Train	Test
Türkiye	1641	410
Dünya	1598	403
Spor	1588	379
Video	1588	402
Yazarlar	1585	406

You can download the data from SUCourse.

Implementation Part:

In this project, you are expected to use Google colab since some of the classification algorithms require more computation. You will do your Naïve Bayes and Logistic Regression implementations on ***project03_NBLG.ipynb*** and CNN implementations on ***project03_CNN.ipynb*** files and submit them with the expected outputs.

In your implementations, you will do the followings:

- Preprocessing:

  In this part, you can choose to do some preprocessing steps like lowercasing tokens, removing stopwords etc. These may be useful for some classification approaches.

- Classification:

  You will use Naïve Bayes and Logistic Regression classifiers. They use different hyper-parameters. You need to understand these and fine-tune them. You can use the GridSearchCV function of sklearn for this purpose. There are also different term weighting algorithms, which we covered in the class. You need to try these as well.

- CNN Classification:

  You will develop a CNN classifier. You can start with some shallow ones and go deeper to see the effects of adding additional new layers. In CNN in addition to trying different architectures, you will also try different word embeddings. Initially you will start with random word embeddings and train the embeddings during the general training of the network. Furthermore, you will use pretrained word embeddings to see how pretraining helps. For the pretrained embeddings, please use the followings:
  - Your trained word2vec embedding from Project 02
  - Word2Vec from this link (https://github.com/akoksal/Turkish-Word2Vec)

  When you use a pretrained embedding, you can keep them static or continue to retrain. Please also do this experiment and report all your findings. For CNN implementations, please use Keras. Our TA will share some additional resources with you.

- Evaluation:

  You are going to report the F1 and Accuracy scores. You will also print the confusion matrix in your reports. You will use these to discuss and compare the approaches in your report. Do not just state the obvious please elaborate on the results based on what you have learnt in the class.

Make sure that your *project03_NBLR.ipynb* and *project03_CNN.ipynb* are well commented. After running all the cells, export the html output. You are going to submit both the ipynb and html with outputs.

You can use the popular/standard python packages. In case you are not sure of a particular library, please ask the instructor or TA.

You are expected to implement this project at your own. Your scripts will be analyzed by using state-of-the-art tools for any type of plagiarism.

Report:

In your report, you are going to summarize your approach and your findings. Discuss what is working and what is not.

Especially with the CNN, you need to discuss the effects of architecture and word embeddings on performance in detail.

You are expected to write this report at your own. Your report will be analyzed with Turnitin.

All ipynb and html files should be under the same directory (named as your student ID, only the 5 digit numbers), which you will zip and submit. You will also submit your report. Your report is not going to be in the zip file. You will submit two files: (1) zip file and (2) report in pdf.

Submission Instructions:

- You will submit this homework via SUCourse.
- Please check the slides for the late submission policy.
- You can resubmit your homework (until the deadline) if you need to.
- Please read this document again before submitting.
- After submitting, you should download your submission to a different path to double check whether everything is in order.
- Please do your assignment individually, do not copy from a friend or the Internet. Plagiarized assignments will receive -100.

References

[1] Şen, Mehmet Umut, and Berrin Yanıkoglu. "SuDer Turkce Haber Derlemlerinin Dokuman Sınıflandırması Document Classification of SuDer Turkish News Corpora."