Tasks Accomplished

1. Create my own dataset for text classification. It should contain at least 1000 words in total and at least two categories with at least 100 examples per category. I created it by scraping the web.
2. Split the dataset into training (at least 160 examples) and test (at least 40 examples) sets.
3. Fine-tuned a pretrained language model (Albert-base-v2) capable of generating text with the dataset I created. Report the test accuracy. Discuss what could be done to improve accuracy.
4. Tried a couple of different language models (GPT-J and GPT-Sw3) to gain a better understanding.

Acknowledgement

I would like to extend my gratitude towards Prof. Zoran Tiganj for his guidance and support throughout the project.

References

[1] Geron, A. (2019). Hands-on machine learning with scikit-learn, keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems (2nd ed.). O'Reilly Media.

[2] Machine Learning Notebooks(8,9,10) on GitHub by A. Geron. https://github.com/ageron/handson-ml2 , https://github.com/ageron/handson-ml3

[2] API Reference. (n.d.). Scikit-learn. https://scikit-learn.org/stable/modules/classes.html

[3] API Reference. (n.d.). Pandas. https://pandas.pydata.org/docs/user_guide/index.html

[4] API Reference. (n.d.). NumPy. https://numpy.org/doc/stable/user/index.html#user

[5] Language Model Reference. (n.d.). Hugging Face Transformers. https://huggingface.co/docs/transformers/training