Progress Report 2

Jayveersinh Raj, Makar Shevchenko, Nikolay Pavlenko

March 29th, 2023

1 Current Progress

Progress on a natural language processing (NLP) project (in the context of a computer science course project) is a complex and nuanced process that can be difficult to measure and evaluate, given the wide range of variables involved. One of the primary challenges of measuring progress in NLP is the inherently subjective nature of language itself. Language is dynamic and constantly evolving, with nuances, idioms, and cultural references that can be difficult for machines to comprehend. As such, progress in NLP is not always straightforward, and the interpretation of success can vary widely depending on the perspective of the stakeholders involved. For example, a linguist may view progress in terms of the accuracy of semantic analysis, while a software engineer may focus on the efficiency and speed of the algorithms. Similarly, a user may measure progress in terms of the naturalness and fluency of language output, while a business stakeholder may focus on metrics such as customer engagement or revenue growth. Furthermore, the evaluation of progress in NLP can be impacted by the quality and quantity of training data, the selection of appropriate algorithms, and the availability of resources such as computing power and human expertise. Another factor that can impact the estimation of progress is the iterative nature of software development. Projects often involve multiple iterations, with each one building on the previous work. As such, progress may not always be visible until later stages of the project, making it difficult to track and measure in real-time.

During the current three weeks not many results were reached, as there was a limited amount of time group members could allocate to the project, and most of the work was geared towards the improvement of performance of the pipeline, as previously we couldn't test it on a large dataset due to the extremely long time BERT encoder took. In attempt to solve it we have decided to use lightweight architecture and replaced the BERT encoder with DistilBERT, which is 40% smaller than the original BERT-base model, is 60% faster than it, and retains 97% of its functionality[1].

We have also been able to collect more data to augment the existing selection, focusing on non-English datasets, so we would have the possibility of testing zero-shot cross-lingual transfer in the future.

2 Team Member Contribution

- **Jayveersinh Raj**: metrics research: top 1 accuracy, model research in preparation for deployment;
- **Makar Shevchenko**: testing of DistilBERT, gauging ways to increase the speed of the model;
- **Nikolay Pavlenko**: dataset compilation, verification of linguistical soundness of the data, composition of intermediary report.

3 Plan for the Next Three Weeks

In the future three weeks we intend to finish the work that has been started at this point, and get results from the pipelines that can be visualized and interpreted, making it easier to gauge the progress. So far performance of DistilBERT has been promising, so it might enable us to test the pipeline on a large enough dataset to make the progress more meaningful.

4 Github Link

Link:*https://github.com/SyrexMinus/cross_lingual_nlp*

References

[1] https://towardsdatascience.com/everything-you-need-to-know-about-albert-roberta-and-distilbert-11a74334b2da