Geolocation Prediction from Jodel and Twitter Messages

Karl Oskar Magnus Holm
Engineering Science and ICT - Department of Geomatics / 2023
koholm@stud.ntnu.no

Abstract

1 Introduction

This project is based on the shared task on Social Media Variety Geolocation (SMG) from VarDial 2020 and 2021 (seventh and eighth edition, respectively), the Workshop on Natural Language Processing (NLP) for Similar Languages, Varieties and Dialects (Chakravarthi et al., 2021; Gaman et al., 2020). The aim of the task differs somewhat from the most common types of NLP VarDial tasks, where the goal typically is to choose from a finite set of variety labels (Scherrer and Ljubešić, 2021, p. 1). Here, the goal is to predict a set of scalars, namely the latitude and longitude from which a social media post was posted. This VarDial task stayed the same from 2020 to 2021, including three language areas: the Bosnian-Croatian-Montenegrin-Serbian language area, the German language area (Germany and Austria in this case), and the German-speaking Switzerland.

This project is limited to the latter of these language areas, that is, the German-speaking Switzerland. Reasons for this include the limited time scope of the task, and having to share the necessary computing resources with fellow students at the department. The goal is to try and recreate the results of Scherrer and Ljubešić (2020) who used a BERT-based classifier, making it a double regression task. I focus on the 2020 dataset because there were a lot more submissions this year as opposed to in 2021, due to the short time between the announcement of the shared task and the submission deadline (Chakravarthi et al., 2021, p. 6).

The reason for picking the task on the German-speaking Switzerland is its similarities to the dialectal landscape of Norway.

2 Background

3 Related Work

4 Model

5 Experiments and Results

5.1 Experimental Setup

All experiments were performed on an NVIDIA GeForce RTX 4090, having 24 GB G6X memory.[1] Computing resources belong to the Department of Geomatics at NTNU and are shared with fellow 5th year geomatics students. I opted for PyTorch[2] in my experiments unlike Scherrer and Ljubešić (2020), who used the high-level SimpleTransformers[3] library.

[1] https://www.nvidia.com/nb-no/geforce/graphics-cards/40-series/rtx-4090/
[2] https://pytorch.org/
[3] https://simpletransformers.ai/

5.1.1 German-speaking Switzerland

Data from VarDial 2020 and 2021 was acquired from a GitHub repository[4] created by Yves Scherrer's, co-author of the winning solution for the SMG task, both years.

5.1.2 Norway

5.2 Experimental Results

6 Evaluation and Discussion

7 Conclusion and Future Work

References

Chakravarthi, B. R., Mihaela, G., Ionescu, R. T., Jauhiainen, H., Jauhiainen, T., Lindén, K., Ljubešić, N., Partanen, N., Priyadharshini, R., Purschke, C., Rajagopal, E., Scherrer, Y., & Zampieri, M. (2021). Findings of the VarDial Evaluation Campaign 2021. *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–11.

Gaman, M., Hovy, D., Ionescu, R. T., Jauhiainen, H., Jauhiainen, T., Lindén, K., Ljubešić, N., Partanen, N., Purschke, C., Scherrer, Y., & Zampieri, M. (2020). A Report on the VarDial Evaluation Campaign 2020. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 1–14.

Scherrer, Y., & Ljubešić, N. (2020). HeLju@VarDial 2020: Social Media Variety Geolocation with BERT Models. *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, 202–211.

Scherrer, Y., & Ljubešić, N. (2021). Social Media Variety Geolocation with geoBERT. *Proceedings of the Eighth Workshop on NLP for Similar Languages, Varieties and Dialects*, 135–140.

[4] https://github.com/yvesscherrer/vardial-shared-tasks