Experiment

Creator **Olusanmi Hundogan** Created **Feb 29, 2024, 20:36** Last updated **Feb 29, 2024, 21:27**

Files

- experiment_linktransformer.py
- analyze_linktransformer.py

Motivation

We recently changed our approach to mapping industry and sector values to a common industry classification scheme. For this purpose, we use a machine learning model-based approach. We use the linktransformer library which uses sentence embedding models to map industries and sectors. In this experiment, we want to explore a number of different models and their impact on the model performance. We also compare the approach to our previous technique of mapping semi-manually with multiple mapping jsons.

The configurations tested are:

- manual
- e5_base
- gte_small
- lt_default_lg
- lt_default

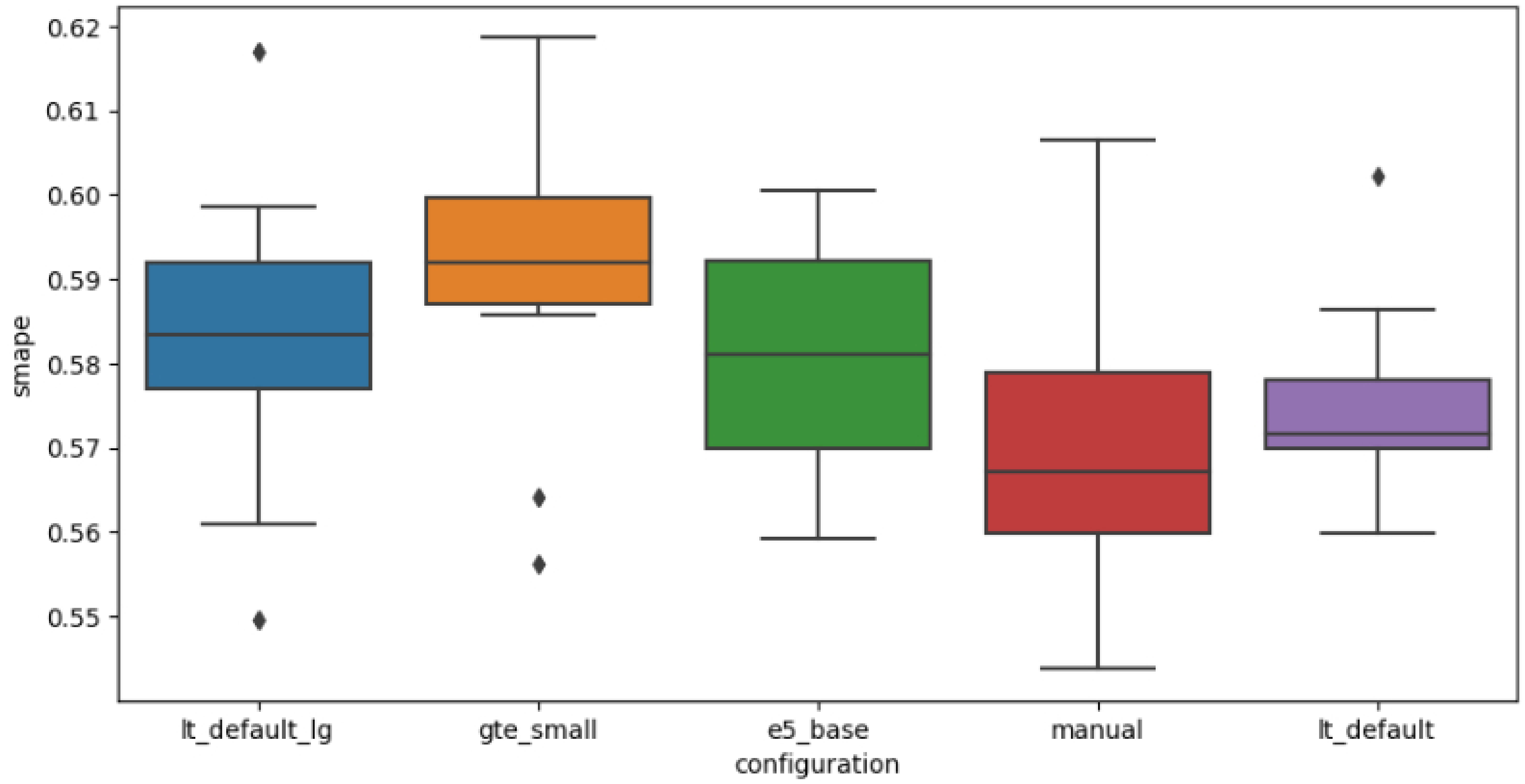
Except manual, all the configurations are models that can be found here: https://huggingface.co/spaces/mteb/leaderboard

Design

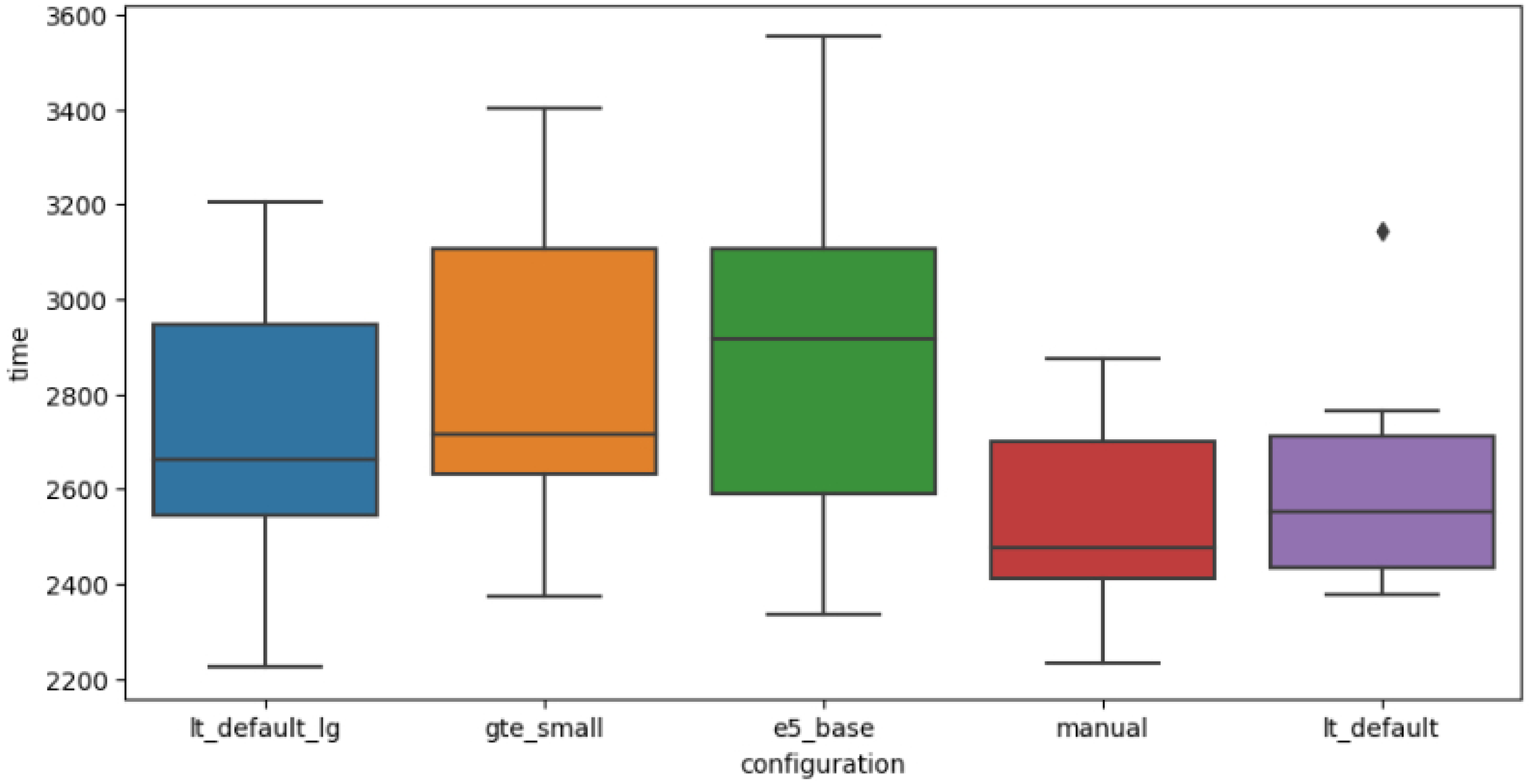
For this experiment, we train an MMA model for each configuration and evaluate the sMAPEvalue on scope 1. We ran the experiment for 10 repetitions for each configuration. We run the experiments without dimensionality reduction.

Results and Insight

The sMAPEresults show that the 'manual' outperforms all the other configurations in terms of median sMAPE This method also reaches lower minimum values but also high maximum values. The second best method appears to be the default link transformer model. This model appears to be more stable than the other methods as well, as the boxplot range is smaller.



With regards to fitting time, both the manual and the lt_default method yield similar speed results. Other models add time to the overall procedure.



Decision

Update 29.02.24 :

The manual approach appears to be better but it is concerning that the model fluctuates strongly. The method can lead to unreliable and unrepeatable results. Hence, it is advisable to use the default linktransformer model. It might be necessaryto repeat the experiment with new feature data coming coming from 12data.