

This work deals with information extraction from Czech News Stories. We focus on four tasks: Publishing server, Article category, Author’s textual gender and Publication day of week. Due to the absence of a suitable dataset for the tasks, we present CZEch NEws Classification dataset (CZE-NEC), one of the most extensive Czech classification datasets, composed of news articles from various sources, spanning over twenty years. Tasks are solved using Logistic Regression and pre-trained Transformer encoders. Emphasis is put on fine-tuning methods of the Transformer models, which are evaluated in detail. The models are compared to human evaluators, revealing significant superiority over humans on all tasks. Furthermore, the models are pitted against the commercial large language model GPT-3, outperforming it on half of the tasks, despite GPT-3 being significantly larger. Our work sets strong baseline results on CZE-NEC allowing for further research in the field.