# Team JUSTR00 at SemEval-2023 Task 3: Transformers for News Articles Classification

**Ahmed Qarqaz** **Malak Abdullah**
Jordan University of Science and Technology
Irbid, Jordan
afqarqaz21@cit.just.edu.jo
mabdullah@just.edu.jo

## Abstract

The SemEval-2023 Task 3 competition offers participants a multi-lingual dataset annotated into three schemes, one for each subtask. The competition challenges participants to construct machine learning systems that can categorize news articles based on their nature and style of writing. We experiment with many state-of-the-art transformer-based language models proposed in the natural language processing literature and report the results of the best ones. Our best performing model is based on a transformer called "Longformer" and has achieved an F1-Micro score of 0.256 on the English version of subtask-1 and an F1-Macro score of 0.442 on subtask-2 on the test data. We also experiment with several state-of-the-art multi-lingual transformer-based models and report the results of the best performing ones.

## 1 Introduction

Transformer-based models have seen immense success in natural language processing (NLP) tasks and the entire machine learning field in general. The self-attention mechanism combined with a deep number of layers has proved to outperform traditional deep learning architectures like the Recurrent Neural Network in the NLP field. Transformer models have also allowed for the process of transfer learning of large models that were pretrained on large corpora. This in turn gave NLP researchers the ability to use the state-of-the-art pretrained models on downstream NLP tasks which has been proven to be a much better method than only training on the downstream task data. In addition to their success in NLP tasks, transformer-based models have also been widely used in computer vision and speech recognition applications. One key advantage of transformer models is their ability to capture long-range dependencies in the input data through self-attention, allowing them to make more informed predictions about the relationship between different elements in a sequence. This has led to breakthroughs in tasks like language translation, sentiment analysis, and question-answering. The pre-trained transformer models like GPT-3 (Brown et al., 2020) and BERT (Devlin et al., 2018) have achieved impressive performance in many NLP benchmarks (Gillioz et al., 2020), often outperforming human-level accuracy (Shlegeris et al., 2022). Another advantage of pretrained transformer models is their ability to adapt to new domains with minimal training data, which greatly reduces the cost and time required for developing new NLP applications. The SemEval-2023 Task 3 competition (Piskorski et al., 2023) is an NLP task in that focuses on the detection of various categories in news articles. Our research aims to explore the performance of different transformer-based models on the Task at hand and provide insights into the strengths and weaknesses of these models for detecting various types of categories in news articles.

## 2 Dataset

The SemEval-2023 Task 3 competition offers datasets of labeled articles from 6 languages. The languages are English, German, Italian, French, Polish, and Russian. Each language has a dataset for training, validation and testing. The competition also offered testing datasets for 3 "surprise" languages (Spanish, Greek and Georgian). In our work, we focus only on the English, German, Italian French and Polish languages. Each dataset was annotated in three schemes one for each sub task. Subtask-1 titled "News Genre Categorization" focuses on categorizing the article based on whether an article is an "opinion", a "report" or a "satire" piece (Multi-class Classification). Subtask-2 is titled "News Framing" and aims to detect the presence of a set of 15 frames that would potentially be used in an article (Multi-label Classification).

*Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1213–1216
July 13-14, 2023 ©2023 Association for Computational Linguistics

Finally subtask-3 is titled "Persuasion Techniques Detection" and aims to detect the presence of a set of 23 pre-defined persuasion techniques in a single paragraph (Multi label classification). Subtasks-1 & 2 work on an article level while subtask-3 works on the paragraph level.

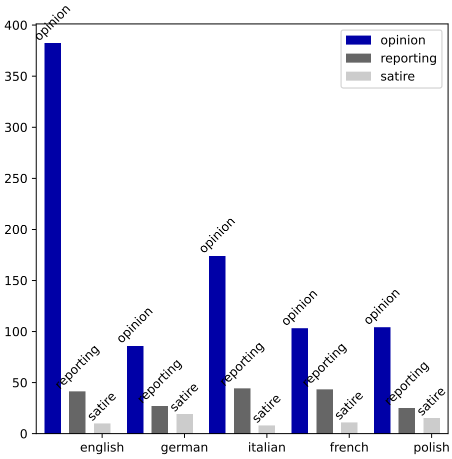

Figure 1: Bar plot of the frequencies of the labels in sub task-1 for different languages.

The datasets are highly imbalanced where certain labels are more frequent than others. For example, the training dataset of subtask-1 for the English language has 382 articles labeled as "opinion", 41 articles as "reporting" and only 10 as "satire". Similarly the labels of subtask-2 for the English dataset are imbalanced with Classes such "Political" and "Morality" having a high number of occurrences (235 & 203) while other classes such as "Public Opinion" and "Economic" have low occurrences (23 & 28). This class imbalance represents one of the major challenges these datasets pose since it would be difficult for the models to learn how to detect the classes with low frequencies in the training data. Figure 1 shows a bar plot of some of the training datasets for subtask-1 showcasing the class-imbalance across all languages. Another major challenge is the inability of the models to take an entire article as input whether in training or inference. The transformer based models use a fixed maximum number of input tokens configuration with 512 tokens being a common one. The articles in the training data typically have a number that exceeds 512 tokens with 75% of the articles having more than 1300 tokens. Sequences that exceed the maximum number of tokens are truncated to only include the first $n$ amount of tokens in an example where $n$ is the maximum input length of a transformer model (e.g. 512).

## 3 Methodology

In this section we describe our methodology where we experiment with a number of different models and report the results. We only focus on the first two subtasks *"News Genre Categorization"* and *"Framing Detection"*. For both subtasks, we stack a fully connected linear layer that takes as input, the pooled output of the transformer model. The pooled output is a vector that represents the input sequence encoded by the transformer. The activation function of the linear layer depends on the subtask. For subtask-1 which is a multi-class classification problem we use a Softmax activation. While for subtask-2 which is a multi-label classification problem we use a Sigmoid activation function. For all our networks we use the Adam optimizer with weight decay and a dropout of 0.1 on the last pooling layer of the transformers.

### 3.1 Pre-trained Models

**BERT** introduced in Devlin et al. (2018) was one of the first groundbreaking transformer-based models that allowed for the fine-tuning of a large pre-trained language model on finer more downstream tasks. Bert utilized the encoder component in the original Transformer architecture (cite "attention is all you need") and was trained using the "Masked Language Modelling" and "Next sentence prediction" objectives on a large dataset. Different versions of BERT were introduced based on the size of the model. The version we use is known commonly as "bert-base-cased". The term "cased" refers to the fact that the model does not lower case the input text and instead maintains the original morphology of a word.

**RoBERTa** (Liu et al., 2019) was introduced as an optimized version of bert where the authors improved on the original bert by (1) using Character-level Byte-pair encoding for the tokenizer (2) Removing the "Next-sentence prediction" objective (3) Increasing the batch-size and finally (4) increasing the training data. Using all of these improvements would lead to a better language model as the authors claim.

**BigBird** was proposed by Zaheer et al. (2020) with an alternative attention mechanism. The orig-

Table 1: Table shows the results of the best performing models on the validation data for the English dataset.

| Model | | | | Subtask-1 | | Subtask-2 | |
|---|---|---|---|---|---|---|---|
| **Model Name** | **Max length** | **Learning Rate** | **Batch-size** | **F1-Micro** | **F1-Macro** | **F1-Micro** | **F1-Macro** |
| bert-base-cased | 512 | 3e-4 | 32 | 0.421 | 0.275 | 0.594 | 0.316 |
| roberta-base | 512 | 1e-5 | 32 | 0.421 | 0.284 | 0.635 | 0.311 |
| bigbird-base | 1400 | 1e-5 | 16 | 0.397 | 0.272 | 0.661 | 0.375 |
| Longformer | 1400 | 1e-5 | 14 | 0.445 | 0.305 | 0.632 | 0.387 |

inal BERT model uses self-attention to calculate the relation of each token in the input sequence to every other token resulting in a complexity of $O(n^2)$ where $n$ is the number of tokens in the input sequence. This in-efficient method is infeasible for longer sequences. BigBird uses a sparse implementation of self-attention to make a more efficient model that can accept a larger input size. Bert's max input length is 512 while BigBird is 4096 tokens. It is worth noting that BirdBird uses the more optimized version of BERT (RoBERTa).

**Longformer** is uses another alternative approach of modifying the self-attention mechanism in the original to achieve a more efficient implementation that allows for reading examples with longer sequences. The authors (Beltagy et al., 2020) highlight three essential concepts of calculating the attention pattern. A *"Sliding Window"* which "slides" accross the input sequence to calculate the self-attention in a local context within the input sequence. A *"Dilated Sliding Window"* which is similar to dilated convolutions in Convolutional Neural Networks is used to increase the receptive field. And finally *"Global Attention"* in which, selected tokens within the input sequence will have global attention where the selected token attends to all input tokens and all input tokens attends to them. For example, the special token `[CLS]` which is commonly used for classification in transformer models can be used as a way of encoding all the information from the entire input sequence by marking it as a selected token for global attention.

**mBERT** means "multilingual BERT" and is a multilingual version of the original BERT (Devlin et al., 2018) with the same number of parameters and architecture. However, the model was trained on a dataset that contains more than just English.

**XLM-RoBERTa** (Conneau et al., 2019) "XLM" is a short-hand that means Cross-lingual Language Model. (cite XLM-roberta paper) proposes to use the RoBERTa model proposed in (cite roberta paper) and train it on a multi-lingual dataset. The model is trained on a 100 languages including English, Italian, French, and German.

### 3.2 Results and Discussion

We perform multiple experiments with each model and report the results. For BERT, RoBERTa, BigBird and Longformer, we train these models only on the English dataset. The multilingual models (mBERT, XLM-RoBERTa) are trained on each language individually. We train each model for 50 epochs and apply an early stopping condition with a patience of 5 epochs. This does mean that for some models, if their performance does not increase after 5 epochs the code halts model training. Table 1 shows the results of training the English only models. We show that training models with a higher maximum input length does indeed lead to better performance. However, a higher input length does require more computational resources and as such it is necessary to use a lower batch size for memory limitations. This issue also prevents us from training language models with a larger size. For example, for BERT we use only the "base" version which has fewer parameters than the "large" version of BERT. We also set a max length of 1400 for BigBird and Longformer even though those two models can accept up to 4096 input tokens. It is worth mentioning that we use a machine equipped with an RTX-A6000 GPU which has 48 Gigabytes of on-memory storage which is considered to be relatively expensive compute and not available to many researchers. Longformer was the best performing model in our experiments and is the model we use for submitting predictions for subtask-1 & 2 for the English language. Longformer achieved an F1-micro of 0.370 and F1-Macro of 0.256 on subtask-1 and subtask-2 respectively on the test data. We do experiments with models capable of large context sizes only on the English data due to the fact that LongFormer and BigBird are only

Table 2: Shows results of training multi-lingual models on each language individually. All models were trained with a batch size of 32 and a maximum input sequence length of 512 tokens.

| Model | | Subtask-1 (F1-Micro) | | | | Subtask-2 (F1-Macro) | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model Name** | **Learning Rate** | **English** | **German** | **French** | **Italian** | **English** | **German** | **French** | **Italian** |
| mbert-base-cased | 3e-4 | 0.530 | 0.644 | 0.666 | 0.779 | 0.348 | 0.323 | 0.258 | 0.305 |
| XLM-roberta-base | 1e-5 | 0.361 | 0.644 | 0.685 | 0.766 | 0.630 | 0.300 | 0.252 | 0.3211 |

pre-trained on English data. We do note however that an open-source multi-lingual large language model called "BLOOM" (Scao et al., 2022) which was released recently. BLOOM has a context size of 2048 tokens and hence it is a promising model for this task. However, we note that simply using models with a large context size input doesn't not entirely solve the problem. According to the annotation guidelines posted by the SemEval-2023 task 3 organizers for the competition dataset.[1], the authors note that for some of the labels/categories in the sub-tasks at hand appear in a form of a fragment in the whole article. Meaning that, a small sentence or even a few words are enough to grant the entire article a label/category. This serves as a hint to the fact that systems with fine-grained analysis of small sequences of tokens present in the articles might prove more promising than systems that process the entire article or even a paragraph all at once.
We present our system's performances on the test data in Table 2.

## 4 Conclusion

We have discussed the transformer-based models that we experimented with. Our best model that we fine-tuned was Longformer and had surpassed the other models due to its ability to read higher sequence lengths. While it is intuitive that a larger max input length leads to better performance, it is possible that adding a preprocessing pipeline instead of naively training on the news articles with no pre-processing would yield better results. The preprocessing pipeline could consist of either handling the class imbalance or focusing more on fine-grained analysis rather than processing the entire all at once articles.

## References

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Anthony Gillioz, Jacky Casas, Elena Mugellini, and Omar Abou Khaled. 2020. Overview of the transformer-based models for nlp tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183. IEEE.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. Semeval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval 2023, Toronto, Canada.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Buck Shlegeris, Fabien Roger, Lawrence Chan, and Euan McLean. 2022. Language models are better than humans at next-token prediction. *arXiv preprint arXiv:2212.11281*.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, et al. 2020. Big bird: Transformers for longer sequences. *Advances in neural information processing systems*, 33:17283–17297.

[1] annotation-guidelines