Impact of Task Adapting on Transformer Models for Targeted Sentiment Analysis in Croatian Headlines

Sofia Lee[♣] **and Jelke Bloem**[♢♠]

♣ Vrije Universiteit Amsterdam
♢ Institute for Logic, Language and Computation, University of Amsterdam
♠ Data Science Centre, University of Amsterdam
s.m.lee2@student.vu.nl, j.bloem@uva.nl

Abstract

Transformer models, such as BERT, are often taken off-the-shelf and then fine-tuned on a downstream task. Although this is sufficient for many tasks, low-resource settings require special attention. We demonstrate an approach of performing an extra stage of self-supervised task-adaptive pre-training to a number of Croatian-supporting Transformer models. In particular, we focus on approaches to language, domain, and task adaptation. The task in question is targeted sentiment analysis for Croatian news headlines. We produce new state-of-the-art results (F_1= 0.781), but the highest performing model still struggles with irony and implicature. Overall, we find that task-adaptive pre-training benefits massively multilingual models but not Croatian-dominant models.

1. Introduction

Transformers, and Bi-directional Encoder Representations from Transformers (BERT) models in particular, have profoundly shaken the NLP field of research at its core. This can be seen not just in the sheer number of papers produced on this topic, nor the emergence of the sub-field occasionally dubbed 'BERTology' (Rogers et al., 2021), but in the number of papers that begin with a remark on how disruptive and innovative the introduction of BERT really is (Devlin et al., 2019). BERT models typically undergo self-supervised pre-training on massive amounts of data. In addition to boasting state-of-the-art performance across many tasks, BERT models thus serve as a firm base for further domain and task adaptation.

Most approaches to BERT task adaptation involves taking a base model and fine-tuning the model for a specific task. In contrast, Gururangan et al. (2020) find that models may benefit from a continuation of pre-training before the fine-tuning stage. The authors present a novel perspective on domain and task adaption. A language model (LM, or BERT in our case) is usually trained for a general domain and exists within a domain which the authors refer to as the 'LM domain'. Domain adaption occurs when the domain of the language model is brought closer to the target domain. Although domain adaption has a number of different approaches, the authors specifically refer to *domain-adaptive pre-training* (DAPT), which involves continued pre-training with massive amounts of unlabelled in-domain data. The authors find that DAPT before the traditional downstream fine-tuning stage yields improvements in performance compared to just fine-tuning.

Furthermore, the authors contrast the set of data that define domains with that of tasks. Tasks lie within a domain—essentially they are a subdomain with its own associated set of unlabelled task data. Gururangan et al. (2020) find that combining DAPT with an extra task-specific stage of pre-training with unlabelled task data, which the authors refer to as *task-adaptive pre-training* (TAPT), improves performance when conducted before fine-tuning compared to only fine-tuning. The authors find that greatest benefits come when both DAPT and TAPT are performed in sequence before fine-tuning.

In this work, we apply task-adaptive pre-training to the task of targeted sentiment analysis (TSA) for Croatian headlines. TSA is a type of sentiment analysis (or opinion mining) which aims to identify the intention's of an author's sentiment towards a target, usually a named entity (NE), irrespective of the tone (or global sentiment). For example, in the following news headline from Barić et al. (2023), translated from Croatian, contrasting targeted sentiments are exhibited (**bold** indicates targets, SUBSCRIPT indicates sentiment):

> **Norway**$_{\text{POS}}$ is the happiest country on earth; **Croatia**$_{\text{NEG}}$ has fallen three places lower than last year.

Although the tone of the headline is neutral, different sentiments are applied to different NEs in the headline; *Norway* is assigned a positive sentiment, whereas *Croatia* is assigned a negative sentiment. The challenge of this task is to disentangle conflicting global sentiments as well as possibly conflicting local sentiments. In many cases, headlines may contain provocative wording which contain both implicit and explicit sentiment towards different entities being discussed. This will also have to be identified by the model. Although we would

LREC-COLING 2024, pages 8662–8674
20-25 May, 2024.

like for news to be neutral, this is sadly often not the case. Partisianship is common in the news domain also in Croatia, with media exhibiting clear biases. The purpose of building a sentiment analysis model for headlines is to help political scientists in analysing trends in how named entities are covered by the media and how the coverage evolves.

The focus of our work is in the Croatian language, specifically with news from Croatian news portals. Croatian is an under-resourced language, and we expect TAPT to have particular advantages for tasks in such languages. We thus contribute to the field of under-resourced languages, as well as expanding upon Barić et al.'s (2023) approach to targeted sentiment analysis and validating the approach of Gururangan et al. (2020) by applying it to a new language and domain.

2. Background and related work

BERT (Devlin et al., 2019) is a powerful language model, but one that is still in need of further fine-tuning before deployment. Ma et al. (2019) present a two-stage 'curriculum-learning and domain-discriminative data selection' framework for domain adaptation. Lin et al. (2020) explore how domain adaptation may be performed for the purpose of detecting negation in clinical notes, finding that BERT models are resistant to over-fitting due to how broad their pre-training stage is. Two papers which come quite close to our work are by Li et al. (2019); Rietzler et al. (2020), who find that 'coarse-to-fine' domain adaptation yields increases in performance with aspect-based sentiment analysis.

Gururangan et al. (2020) also suggest that different levels of granularity exist with domain- and task-adaptive pre-training (DAPT and TAPT respectively), and that leveraging them together can result in increases in performance. We see our work as a continuation of this work by exploring how TAPT can apply to different Croatian-supporting language models.

2.1. Sentiment analysis

Recent work on sentiment analysis has paid more attention to fine-grained tasks in sentiment analysis, particularly aspect-based sentiment analysis, which involves identifying implicit sentiments towards different aspects of a larger entity (Pavlopoulos, 2014). Targeted sentiment analysis (TSA), the task which we will be exploring, differs from aspect-based sentiment analysis due to its aim of mining the sentiment towards topics or named entities in a piece of text, rather than on features or parts of a named entity with the goal of mining feature-oriented evaluation. Previous work in TSA largely involve identifying the sentiment towards targets in tweets (Jiang et al., 2011; Saif et al., 2013). Such approaches include the use of gated neural networks (Zhang et al., 2016; Jabreel et al., 2018) and BERT models specifically for COVID-19-related tweets (Zhou et al., 2022).

Although sentiment analysis has been a crucial task in NLP, a majority of the work is done on English (Dashtipour et al., 2016). For other languages, the lack of quality annotated data poses a significant challenge. Notable early exceptions concern high-resource languages such as German (Li et al., 2012) and Chinese (Wan, 2008). Aside from an ambitious project that attempts to provide a sentiment lexicon for 136 languages (Chen and Skiena, 2014), much of the non-English research in this field is recent. Salgueiro et al. (2022) provide a polarity data set for Spanish based on political headlines. Basile and Nissim (2013) introduce the first data set of Italian tweets. Few other data sets are available.

The impact of transformer models has also been felt here. Languages for which sentiment analysis transformers models exist include Turkish (Mutlu and Özgür, 2022), Hindi and Bengali (Khan and Shahid, 2022), Swahili (Martin et al., 2021), Spanish (Vásquez et al., 2021), and Russian (Kotelnikov, 2021). Approaches are broad, ranging from applying a multilingual model, pre-training entirely from scratch, transfer learning from a high resource language, or aggregating data from similar or related languages. Makogon and Samokhin (2021) come quite close to our work by performing targeted sentiment analysis on news in Russian and Ukrainian, two Slavic languages.

Our interest in sentiment analysis is situated within the news domain, in which headlines are a notable feature. Tasks related to headlines can include keyword mining (Eiken et al., 2006), generation (Banko et al., 2000) or 'fake news' detection (Liu et al., 2021). Notable data sets for the news domain exist primarily for English and include GoodNewsEveryone (Bostan et al., 2020), which contains a crowd-annotated set of 5000 headlines; SemEval-2004 Task 14 (Strapparava and Mihalcea, 2007), a data set for semantic evaluation; and a Million News Headlines (Kulkarni, 2018), an unlabelled set of a million headlines from Australia. Within the news domain, targeted sentiment analysis is also used for analysing headlines. In contrast to our work, much of this work is performed in financial news headlines (Xiang et al., 2022; Du et al., 2023). Aside from Barić et al. (2023), all work appears to be for English news sources.

2.2. Croatian

Croatian is a South Slavic language spoken primarily in Croatia and neighbouring countries Italy, Austria, Hungary, Serbia, Bosnia and Herzegovina, and Montenegro, plus a large diaspora community worldwide. It is mutually intelligible with Bosnian, Montenegrin, and Serbian (Golubović and Gooskens, 2015). These four languages are often grouped together under the pluri-centric designation Bosnian-Croatian-Montenegrin-Serbian (BCMS) or, formerly, Serbo-Croatian (Brozovid, 1991; Bugarski, 2019). Croatian involvement in machine learning dates back to the 70s. In fact, contrary to popular knowledge, which assumes a much later introduction in the 90s, transfer learning was first described in a paper written in Croatian (Bozinovski, 2020). Despite its role in machine learning history, however, Croatian can presently be considered am under-resourced language (Hedderich et al., 2021). Training data for Croatian is often augmented with data from a similar, related language such as Slovene or from data the very similar but politically distinct neighbouring languages of Bosnian, Montenegrin and Serbian (Ljubešić and Lauc, 2021; Ulčar and Robnik-Šikonja, 2020).

In addition to massively multilingual models such as Multilingual BERT (mBERT, Devlin et al., 2019) or XLM-RoBERTa (Conneau et al., 2020), there also exist language-specific models for BCMS. To our knowledge, only two language-specific BERT models have been created. The first model is CroSloEngual BERT (pronounced 'Crosslingual BERT') or cseBERT[1], produced by Ulčar and Robnik-Šikonja (2020) as part of research on the impacts of transfer learning between related languages on BERT models in comparison to massively multilingual models. Ulčar and Robnik-Šikonja (2020) show that their approach of transfer learning with fewer languages outperforms mBERT in three tasks. Despite its considerably larger proportion of Croatian in its pre-training corpus, cseBERT primarily appears in research for Slovene (Žagar and Robnik-Šikonja, 2022). The second model is BERTić, the current state-of-the-art, BCMS-focused model introduced by Ljubešić and Lauc (2021). BERTić outperforms both cseBERT and mBERT in morphosyntactic tagging, named entity recognition, social media geo-location and commonsense causal reasoning. Its use in Croatian include hate-speech detection (Shekhar et al., 2022) and sentiment analysis of parliament proceedings in Bosnia and Herzegovina, Croatia and Serbia (Mochtak et al., 2022). It has also been used in the closely related but politically distinct Serbian language for sentiment analysis (Batanović and Miličević Petrović, 2022), sentiment-based topic modelling in the context of COVID-19 vaccines (Ljajić et al., 2022), as well as behavioural testing with indeclinable nouns (Lee and Bloem, 2023).

There has been some recent work on Croatian sentiment analysis. Thakkar et al. (2023) provide a sentiment-annotated data set of Croatian film reviews. The work closest to ours is Barić et al.'s (2023) work, which introduces the SToNe data set. This is a data set of Croatian headlines for both tone, or global sentiment of a headline, and targeted sentiment, or sentiment towards a particular named entity within a headline. The authors maintain that there is a statistical relationship between these two aspects of a headline and use a number of approaches, including averaged, mixed or alternative batches, to build a BERTić-based model for targeted sentiment analysis.

3. Data

We use a yet unpublished dataset collected using the TakeLab Retriever (Ćurković et al., 2022), which includes a web scraper that routinely trawls news articles from assorted Croatian web portals, to obtain a dataset of Croatian news headlines and their associated metadata. From this dataset, we extracted 8.34 million headlines, spanning from 1 January 2001 to 26 April, 2023, the day of retrieval. Headlines largely come from Croatian national news sites, though some regional publications and blogs are also included, which occasionally use regionalisms. Of note is the fact that the headlines data set consists of a nearly exhaustive representation of the Croatian online news headline sub-domain at the time of retrieval. This is an unusually rich quantity of data related to our task, thus being particularly suitable for the purpose of task-adaptive pre-training (TAPT).

We tokenised the headlines dataset using the ReLDI tokeniser, a rule-based tokeniser for Croatian provided by the CLASSLA Python package (Ljubešić and Dobrovoljc, 2019). We then applied a fuzzy matching-based de-duplication process, reducing the headline count by 11.91%, to 7.35 million headlines, and pruned 5,027 one-word headlines. We also removed headlines that appear in the SToNe validation and test set. For every tested model except BERTić, we also concatenated all the headlines, tokenised them using each model's respective tokeniser, and then split the headlines into equal-sized chunks of 512 sub-word tokens, the max token limit for each of the models. Subsequently, 99% of the data set was used for training with the remaining 1% used for evaluation.

[1] https://huggingface.co/EMBEDDIA/crosloengual-bert

	Case	NEG	NTR	POS	Total
Hrvatska	NOM	1	1	5	7
Hrvatske	GEN	0	4	2	6
Hrvatskoj	D/L	1	8	2	11
Hrvatsku	ACC	1	9	3	13
Hrvatskom	INS	0	0	1	1
Total		3	22	13	38

Table 1: Distribution of the declension of *Hrvatska* ('Croatia') across different sentiment labels. We merge dative and locative due to the irrecoverability without context.

3.1. SToNe dataset

We also use the as yet unpublished SToNe dataset (Barić et al., 2023), which is is an annotated subdata set of the aforementioned headlines data set containing named entities (NEs) as well as labels for the sentiment towards the NEs and the general tone of the headline. Four NE categories (PER for people, ORG for organisations, LOC for locations, and MISC for everything else) are present. The PER label notably also includes the names of ethnic groups. Targeted sentiment and tone are annotated with a ternary annotation scheme of *negative* (NEG), *neutral* (NTR), and *positive* (POS) labels. Further details on the annotation process are described by Barić et al. (2023). We use the 2,308 headlines annotated with full agreement and did not perform any additional pre-processing.

A peculiarity of this data set is the lack of lemmatization. Although this would have a minimal impact for English due to the lack of inflection, this has a number of consequences for a highly-inflected language like Croatian, where named entities are also inflected. The first consequence is that it creates sparsity. For example, the entity Croatia is not only spread out across different terms, such as 'Croatia', 'Republic of Croatia' and 'HR', but each term is spread out across different declensions, resulting in the term appearing to be much less frequent. In Table 1, we demonstrate how *Hrvatska* ('Croatia', as opposed to 'Republic of Croatia' or any abbreviations) is referred to 38 times in total. However this count is distributed across five different declined forms, each with a different sentiment ratio.

The second consequence is the grammatical function of the entity is partially recoverable by looking at the ending, depending on the type of named entity. In the case of *Hrvatska*, there is a tendency towards positive labels in the nominative case, indicating the agent of an active verb or patient or theme of a passive verb, whereas locative and accusative, both used generally to indicate a location or direction, tend to be neutral. This is, however, dependent on the type of named entity. For example, *Hrvat* appears in six different variations in the data set with 70% positive labels and no clear relationship with case.

Model	Training breakdown
BERTić	Croatian (66.3%), Serbian (23.33%), Bosnian (9.42%), Montenegrin (0.95%)
cseBERT	English (47%), Croatian (31%), Slovene (23%)
mBERT	Includes Croatian, Bosnian, Serbian and Serbo-Croatian
XLM-RoBERTa	Includes Croatian (5.7G), Bosnian (18M) and Serbian (1.5G)

Table 2: Model training data size in Croatian and related languages, if provided.

4. Methodology

We selected five models to examine how additional self-supervised pre-training with unlabelled task data (TAPT) affects targeted sentiment analysis performance. The models selected represent a diverse set of pre-training approaches as well as vary in terms of the number of languages covered, and together encompass a near totality of Croatian language modeling. They are either *Croatian-dominant* or *massively multilingual* models. Table 2 shows comparison of exposure to Croatian training data for each model.

4.1. Croatian-dominant

The first model, BERTić, is a model trained exclusively on corpora derived from Bosnian, Croatian, Montenegrin and Serbian sources (Ljubešić and Lauc, 2021), consisting of 8.39 billion tokens. The headline data set highly overlaps BERTić's vocabulary (94.33%) with only 459 out-of-vocabulary tokens. A notable quirk of BERTić is that it is trained with the ELECTRA objective, or 'Efficiently Learning an Encoder that Classifies Token Replacements Accurately' (Clark et al., 2019). Rather than performing masked language modelling, as done with traditional BERT models (Devlin et al., 2019), ELECTRA models are trained on a *replaced token detection* (RTD) task. This means that the pre-training procedure performed during TAPT must be a continuation of the RTD task. Aside from the ELECTRA objective, BERTić otherwise closely follows the specifications of the base BERT model, with 12 layers and 110M parameters. It was the main focus of previous work done by Barić et al. (2023) on entity-level sentiment analysis in Croatian headlines.

The second model, cseBERT, is a tri-lingual model trained on English, Croatian and Slovene

(Ulčar and Robnik-Šikonja, 2020). It was trained on a corpus of 5.9 billion tokens, predominantly composed of English, with a smaller portion in Croatian. Unlike BERTić but also unlike other BERT models, cseBERT is trained on the *whole word masking* (WWM) task, also known as the Cloze task (Taylor, 1953). WWM masks entire words, requiring the target model to recover the whole word rather than just WordPiece sub-word tokens (Schuster and Nakajima, 2012). The headline data set moderately overlaps cseBERT's vocabulary (68.32%), and it encounters the most out-of-vocabulary tokens out of all the models: 798,487 tokens, or 0.54% of the tokens.

4.2. Massively multilingual

We also tested Bert-Base-Multilingual-Cased, XLM-RoBERTa-Base, and XLM-RoBERTa-Large. All models are pre-trained on 100 or more languages, albeit with varying degrees of Croatian data. The pre-training process objective for these models is *masked language modelling* (MLM), the standard training procedure for BERT models (Devlin et al., 2019).

Multilingual BERT (mBERT) is the original massively-multilingual BERT model (Devlin et al., 2019). We use the updated cased model, bert-based-multilingual-cased. mBERT is pre-trained on a corpus consisting of the top 104 language editions of Wikipedia, including the Croatian, Bosnian, Serbian and Serbo-Croatian editions. mBERT has a high out-of-vocabulary rate on the headlines data set (0.41%), suggesting that mBERT may have the least exposure to the target domain.

The XLM-RoBERTa models (Conneau et al., 2020) differ from mBERT in a few notable ways. They make use of Byte-Pair Encoder (BPE) tokenisation, introduced by Sennrich et al. (2016), instead of the standard BERT WordPiece tokeniser. This approach to tokenisation may be responsible for the considerably low number of out-of-vocabulary tokens for these models ($< 0.001\%$). With 2.5TB of data total, it is by far the model exposed to the most amount of data, although it has only been exposed to 515.23 million tokens of Croatian.

4.3. Training procedure

The training procedure we followed was adapted from Gururangan et al. (2020). We specifically adopted a two-stage approach. The first stage consisted of *task-adaptive pre-training* (TAPT), which adapted the models to the general unlabelled data of the task using their original pre-training objective. For each model, we also produced a version for comparison which omitted this stage. Due to resource constraints, we only trained the models for three epochs. The second stage consisted of fine-tuning on the labelled SToNe data set with loss calculated on the validation portion of the set. We followed the process of the *Target* baseline from Barić et al. (2023). Each model was tested after 10 epochs of fine-tuning, after exploring values between 3 and 50 epochs.

Model	Before	After
BERTić	190,601.81	3,383.42
cseBERT	279.35	12.07
XLM-RoBERTa-Base	218.22	3.64
mBERT	36.55	2.73
XLM-RoBERTa-Large	5.39	2.72

Table 3: Perplexity across pre-training

The final models were tasked on their ability to predict the sentiment labels selected by the annotators. To evaluate the results of the initial TAPT stage, we calculated perplexity on the validation set. The overall evaluation of the task employs F_1-score using macro averages over all classes, including a comparison of gains or losses from TAPT training.

5. Results and analysis

All models showed a drop in perplexity after the task-adaptive pre-training stage, indicating that all models learned from the task. However, each model increased by dramatically different amounts. BERTić dropped from an exceptionally high 190,601.81 to a much lower, but still high 3,383.42. This shows that BERTić still struggles considerably with the replaced token detection task with the pre-training data set. Other models had much lower perplexity values, although, interestingly, the lowest values, both before and after training, all went to the multilingual models. See Table 3 for all pre-training results.

5.1. Model performance

Each model was tested with five seeds with the results from the test set then averaged across the seeds. Our worst-performing models were both versions of mBERT, followed by XLM-RoBERTa-Base without TAPT. The next lowest performing models were both cseBERTs, followed by a tie between XLM-RoBERTa-Base with TAPT and XLM-RoBERTa-Large without TAPT. Although BERTić fared well above most of the competition (F_1 = 0.745), it ultimately lost to XLM-RoBERTa-Large with TAPT, the highest-performing model of the entire set. The results are presented in Table 4.

An entirely different picture is painted when examining the results through the gains (Table

Model	AVG	Neg	Ntr	Pos
BERTić	0.745	0.721	0.770	0.744
+ tapt	0.736	0.733	0.766	0.708
cseBERT	0.718	0.708	0.739	0.706
+ tapt	0.711	0.696	0.752	0.687
mBERT	0.600	0.550	0.688	0.561
+ tapt	0.660	0.634	0.718	0.628
XLM-R-Base	0.669	0.633	0.726	0.648
+ tapt	0.728	0.702	0.763	0.719
XLM-R-Large	0.728	0.723	0.749	0.713
+ tapt	**0.771**	**0.770**	**0.793**	**0.750**

Table 4: Comparison of F_1-scores for all models with and without task-adaptive pre-training (tapt).

Model	AVG	Neg	Ntr	Pos
BERTić	-1.2%	1.7%	-0.5%	-4.8%
cseBERT	-1.0%	-1.7%	1.8%	-2.7%
mBERT	**10.0%**	**15.3%**	4.4%	**11.9%**
XLM-R-B	8.8%	10.9%	5.1%	11.0%
XLM-R-L	5.9%	6.5%	**5.9%**	5.2%

Table 5: The effect of tapt training by percent increase per model per label. A negative number indicates that performance decreased with the inclusion of the tapt stage.

5). All Croatian-dominant models experience decreases in performance with tapt, with BERTić decreasing 1.2% in F_1-score after the added pre-training. cseBERT experiences a similar but slightly smaller decrement, 1.0%. On the other hand, all massively multilingual models experience performance boosts with tapt.

5.2. Error analysis

We perform an error analysis of one run from the highest performing model, XLM-RoBERTa-Large with tapt, henceforth referred to as XLM-LT. Despite its strong performance (F_1= 0.781), there is still considerable room for improvement for XLM-LT. We provide an overview of final scores, Table 6, and a confusion matrix of the results, Table 7.

We break down errors into three categories:

1. **Opposite** errors are errors where the opposite polar label (Neg or Pos) is predicted.

2. **Neutralising** or neutralisation errors occur when Ntr is predicted instead of a polar label, resulting in a polar sentiment being neutralised.

3. **Polarising** or polarisation errors are predictions where Neg or Pos is predicted instead of a Ntr label, resulting in a neutral sentiment being interpreted as a polar one.

	Precision	Recall	F_1-score	N
Neg	0.736	0.800	0.766	115
Ntr	0.834	0.796	0.815	221
Pos	0.762	0.762	0.762	126
Overall	0.777	0.786	0.781	462

Table 6: Results for each label for XLM-RoBERTa-Large.

	Neg	Ntr	Pos	N
Neg	**92**	16	7	115
Ntr	22	**176**	23	221
Pos	11	19	**96**	126

Table 7: Confusion matrix for the labelling performance of XLM-LT. The rows indicate gold labels and the columns indicate predictions.

The logic behind this subdivision is that opposite errors are significantly rarer than neutralising or polarising errors, and are more severe.

Table 7 shows that many of XLM-LT's errors come from polarising errors, that is, by predicting a polar label when the gold label is Ntr. This is also evidenced by the lower precision for both Neg and Pos compared to Ntr. XLM-LT is shown to over-predict Neg labels in particular. That said, XLM-LT very rarely produces opposite errors, making up 18.37% of errors.

5.2.1. Results by named entity types

Table 8 shows a classification chart filtered by NE type. Per is a particularly weak NE type for our classifier, having the lowest F_1-score (0.755). The model appears to to under-predict Ntr labels for this type, producing polarisation errors by assigning Neg or Pos. While only 18.37% of errors overall are opposite errors, 66.67% of such errors are associated with the Per label. Considering that Per makes up 45.89% of the NE types in the test set, something about headlines with Per targets may be difficult to interpret properly. One possible explanation could be irony. Many such headlines mix positive and negative statements for the purpose of creating irony, which is generally indicative of Neg sentiment towards the target.

On the other hand, Org is the strongest NE type for XLM-LT, showing strong performance across the board, except in Org+Pos recall. Aside from that, errors are predominantly neutralisation errors. Abbreviations appear to cause particular difficulty for Org. While abbreviations make up 12.97% of Org in the training set or 5.58% of training overall, they are over-represented in the test set and cause 45% of the errors in Org. Some of these issues may be caused by issues with tokenisation and span labelling in the training and test set of the original dataset, where we observe that inflectional endings of abbreviations are inconsistently included or excluded from spans.

The classifier's performance in Loc, like Org, is almost opposite that of Per. Loc+Ntr is the model's strongest point (precision = 0.855, recall = 0.942, F_1= 0.897). Loc+Pos precision is very high (0.909), indicating that most of its predictions for this label are correct. However, both Loc+Neg and Loc+Pos have low recall. The classifier makes a lot of neutralisation errors, tending to predict a Ntr sentiment with locations.

In Table 9, we further break down the performance of Loc by case. A case-oriented analysis of Loc can be particularly interesting considering that, unlike other NE types, locations tend to be assigned Dative/Locative case but different cases have a different sentiment distribution. For example, Dative/Locative is overwhelmingly Ntr (81.13%) in our test set, but there are statistically more polar sentiments for Nominative and Genitive. Additionally, cases are easily discernible for Loc due to the lack of indeclinable nouns in its semantic class (Lee and Bloem, 2023). Finally, cases can reveal to what extent a classifier can leverage higher level processes such as semantic roles to parse the sentiment of input.

We see that Dat/Loc+Ntr dominates the data set and that the classifier is adept at predicting it, but it is unable to recognise when a location is being portrayed positively or negatively. This is evidenced by its considerably low recall (Neg = 0.500, Pos = 0.600). In the following example, Croatia is in the Dative case, and the author's sentiment towards Croatia is Pos. Although in another location it may be appropriately predicted Ntr, here it is expected that the author intends for the focus on Croatia to be Pos as the actress in question has selected Croatia (i.e., the country of the author) over other places to stay:

> Atraktivna detektivka iz popularne serije boravi u **Hrvatskoj**
>
> The attractive detective from a popular series resides in **Croatia**

		P.	R.	F_1	N
Per	Neg	**0.682**	0.833	0.750	54
	Ntr	0.847	**0.678**	0.753	90
	Pos	**0.730**	0.794	0.761	68
	Avg	0.753	0.768	0.755	212
Org	Neg	0.806	0.806	0.806	36
	Ntr	0.780	0.830	0.804	47
	Pos	0.821	**0.742**	0.780	31
	Avg	0.802	0.792	0.796	114
Loc	Neg	0.778	**0.737**	0.757	19
	Ntr	0.855	0.942	0.897	69
	Pos	0.909	**0.588**	**0.714**	17
	Avg	0.847	0.756	0.789	105
Misc	Neg	0.800	**0.667**	**0.727**	6
	Ntr	0.846	**0.733**	0.786	15
	Pos	**0.692**	0.900	0.783	10
	Avg	0.779	0.767	0.765	31

Table 8: Results for XLM-RoBERTa-Large + tapt distributed across named entity types. Performance below 0.750 is **bolded**.

		P.	R.	F_1	N
Nom	Neg	**0.714**	**0.714**	**0.714**	7
	Ntr	0.818	0.818	0.818	11
	Pos	1.000	1.000	1.000	5
	Avg	0.844	0.844	0.844	23
Gen	Neg	0.800	0.800	0.800	5
	Ntr	0.789	1.000	0.882	15
	Pos	1.000	**0.333**	**0.500**	6
	Avg	0.863	0.711	0.727	26
Dat/Loc	Neg	**0.667**	**0.500**	**0.571**	4
	Ntr	0.902	0.949	0.925	39
	Pos	0.750	**0.600**	**0.667**	5
	Avg	0.773	0.683	0.721	48
Acc	Neg	1.000	1.000	1.000	3
	Ntr	0.750	1.000	0.857	3
	Pos	**0.000**	**0.000**	**0.000**	1
	Avg	0.583	0.667	0.619	7

Table 9: Results for XLM-RoBERTa-Large + tapt, Loc named entity types, divided further by case.

The performance here suggests that the classifier is not picking up on more implicit sentiment, particularly with the Dative/Locative case. On the other hand, Nom+Pos is particularly strong as is Nom overall, with errors being confusion between Ntr and Neg. Meanwhile, Pos is weak across both GEN and ACC, although ACC in particular is weak. These are both associated with lower precision in Ntr, further suggesting that locations are consistently predicted to be Ntr by XLM-LT. Overall, this case-based analysis shows evidence of unintended shortcuts being learned. Particularly, the classifier seems to associate certain semantic roles with certain sentiments.

6. Discussion

Gururangan et al. (2020) tested the tapt approach with different amounts of domain-relevant data, finding that the more domain-relevant, the better the performance. Our work, in contrast, tested the same domain and task-relevant data set, but with different models trained on different languages. We found that tapt indeed yielded benefits to massively multilingual models, but we observed regressions in performance for Croatian-specific models. However, it is worth noting that not all improvements nor regressions were equal. In fact, none of the models showed changes in performance in the same way, not even the two XLM models which had been pre-trained on the same data. Our results suggest that TAPT is a suitable approach if and only if the models being trained have not been exposed to this data already.

We also suspect that the size of the model plays a role in what the model gets out of TAPT. It is possible that XLM-RoBERTa-Large's expanded parameters allows it to pick up on subtleties in NTR that allowed it to see the largest amount of improvement in handling that label. Meanwhile, while mBERT saw the most improvement overall, including a significant improvement in NEG, it still performed the worst out of all models after TAPT; its large improvements only demonstrate the proportion by which it improved from a rather poor-performing model.

Lastly, we propose the existence of languages as 'super-domains', which apply in particular to under-resourced languages. This is an aspect thoroughly unexplored by Gururangan et al. (2020), as the authors are concerned with English language modelling. A language 'super-domain', an order above domains or tasks, would consist of all the data in existence for a language and possibly even related languages. A 'super-domain' would also have its own related pre-training task, *language adaptive pre-training* (LAPT) which we propose occurring before domain-adaptive pre-training.

One of our takeaways from this work is that under-resourced languages require a special kind of attention that high resource languages do not. Obstacles relating to lack of data need to be overcome. This includes more direct obstacles such as encountering little means to extend the training for a particular model because the model has already seen all data for the language. There are also indirect obstacles, such as needing to use models based on alternative but less accessible architectures like ELECTRA or whole word masking, to make use of all available resources. On top of that, under-resourced languages require particular focus on addressing biases, which are amplified by the low resolution of available data. Addressing such biases requires sensitivity not just to explicit but also implicit understanding of text, necessitating particular familiarity with the language in question as well as its surrounding culture, political situation, and history.

While we demonstrate how TAPT improves performance, we strongly underline the fact that a keen understanding of both the necessity and suitability of an approach is key. This means that TAPT should not become a 'must-do' but rather be included as part of a diverse toolbox of approaches domain and task adaptation if seen fit.

6.1. Future work

We observed in Table 9 a potential statistical relationship between semantic role and sentiment. Future work could incorporate this information either by passing case to the classifier, which is predominantly a grammatical function, or by passing semantic label to the classifier, thus indicating the semantic role of the target. The inclusion of case has not yet been researched extensively in sentiment analysis for highly inflective languages such as Croatian.

Although lemmatisation has been shown to have minimal impact on sentiment performance in English (Palomino and Aider, 2022), it is unknown how this will impact highly inflected languages such as Croatian. There still remains an unexplored possibility for the lemma of the named entity to be passed, for example using the reldi-tagger (Ljubešić and Dobrovoljc, 2019), F_1= 98.17.

Gururangan et al. (2020) show not only benefits from TAPT followed by the usual fine-tuning procedure, but also by performing a preliminary stage of in-domain pre-training before both of these training stages. This process of *domain-adaptive pre-training* (DAPT), is one which we have entirely skipped in our study. There is, nevertheless, further potential for improvement by incorporating more in-domain data. In our case, this could be more general unlabelled data from the Croatian news domain. Furthermore, building on our previous discussion of language super-domains, there also exists the potential to perform both LAPT and DAPT. This may be particularly beneficial for models that have not seen the full extent of Croatian training data nor training data from the closely related Bosnian, Montenegrin, and Serbian used for BERTić. Exposure to more general language-related data should result in performance gains by adapting the models away from being language-neutral (in the case of multilingual models) into being more language-specific. Future work can thus compare different combinations of LAPT, DAPT and TAPT with fine-tuning, although we also echo the warnings of Gururangan et al. (2020), that training with specific data first followed by more general data may lead to catastrophic forgetting.

Lastly, there is potential for improvement of the models. It is likely that BERTić has been undertrained and would benefit from significantly more epochs of pre-training, as shown by its high perplexity values in pre-training. Its WordPiece tokeniser may also be less suitable for highly inflected languages. As for XLM-RoBERTa, it is likely that an even larger model be better adept at the task. Despite some previous research warning of 'English influence' when using multilingual models in some tasks (Papadimitriou et al., 2023), we predict that with a task like targeted sentiment analysis, this may not necessarily be an issue. However, we found that performance increases diminished as model size increased, and a larger model may not perform better enough to justify the resources required to train it.

7. Conclusion

We developed sequence classification models for the task of targeted sentiment analysis, trained with and without task-adaptive pre-training (TAPT) using a very large database of unlabelled Croatian headlines to identify the impact on their performance. Our top model, XLM-RoBERTa-Large (F_1= 0.771), outperformed the previous state-of-the-art (Barić et al., 2023). Our findings indicated that TAPT yielded improvements on massively multilingual models, but not under-resourced language-specific models.

We found that the highest-performing model still suffered from linguistic issues such as irony-detection, understanding aspects, and implicature. There were issues associated with span errors and potential over-fitting with semantic roles. These would have to be addressed through further modifications to the training data set and explored in future work. Finally, we found that Croatian's status as an under-resourced language may have had a large impact on how these models changed. Our work may have demonstrated what happens when a model continues pre-training on data it has already seen. Other quirks with our approach may have also been influenced by low resources.

However, we were also able to contribute to research relating to domain adaptation by Gururangan et al. (2020) by exploring how TAPT works in low-resource settings. Languages can be considered a 'super-domain', adding another layer to coarse-to-fine adaptation paradigms. Future work should consider exploring the impact of language adaptive pre-training for multilingual models, especially when the alternative monolingual or near-monolingual models have already seen nearly all data available.

Lastly, we hope that our contributions can expand future work in under-resourced languages and continue to highlight that they require particular types of approaches and thinking. If tasks and domains require familiarity on the NLP researcher's part, then languages and super-domains do as well. Although larger models are indeed beneficial, they require careful application and treatment in order to succeed.

8. Ethical considerations and limitations

Echoing ethical concerns of Rupnik et al. (2023), we would like to acknowledge that, although the bulk of the data we work with comes from Croatian news portals, we cannot be sure of all the perspectives of the authors with regards to the language that is being used. A small minority of articles in the headlines data set, and possibly SToNe, come from Bosnian and Serbian sources. On top of that, it is possible that articles are simply copied over from other languages with little to no modification. However, we justify their inclusion by noting that they constitute a very small amount of our data and are represented in our models only as statistical relationships based on headlines.

Aside from language identity of the source data, we can only attest to our model's performance in Croatian-dominant data only. Although we have observed the similarities between Croatian, Bosnian, Montenegrin and Serbian, we cannot be certain that the performance of our specific model can be generalised beyond Croatian. Considering our observations that there is a correlation between 'Croatia' and positive sentiment in our data set, we note that there may be biases that are related to the cultural or regional domain of the data rather than being of linguistic or lexical significance.

The intention of this model is specifically to track trends and biases in Croatian news. We caution users of such model, whether it is XLM-RoBERTa-Large with TAPT or another one borne from another approach to training, to take the results with a grain of salt. Even if we were to find a model which achieves a perfect average F_1-score on our test set, we cannot be certain that the model is free of biases. While further testing, such as through behavioural testing as discussed above, may be performed to identify where biases exist, this still does not preclude the possibility of bias in the system.

Ultimately, headlines are simply headlines. Much like the adage of how a book should not be judged by its cover, a news article cannot be judged solely by its headline. Headlines may serve as indications of news trends (Bourgonje et al., 2017), but they alone may not capture the full picture of how an entity is being depicted. In fact, headlines may even be intentionally construed to mislead, confuse or shock a reader into reading an article. This is to say nothing of future, as-of yet unrepresented in training trends in headline title styling. Simply put, headlines are not the end all, be all of news analysis but rather only one small, albeit crucial part of a larger system of news media, which includes articles, authors, publications and portals. We urge those who use this tool to be aware that AI language models are another form of statistical analysis that represents a simplification of data, in this case, a rather restricted subset of a domain that is notably fraught with partisanship and misdirection.

Although most BERT models use a masked language modelling (MLM) training objective which masks a certain proportion of sub-word tokens, not every model uses this approach. In our research, both our language-specific models used a non-MLM approach; cseBERT used the similar but still more challenging whole word masking approach, whereas BERTić used the ELECTRA objective of replaced token detection. This limits the generalizability of our findings. In absence of a language-specific MLM model, we are unable to determine the extent to which training objective itself is responsible for both cseBERT and BERTić's declines in performance after TAPT. It is difficult to evaluate the suitability of these tasks for the corpus we had worked with. Although the closeness in training corpus is the most likely culprit, we cannot ignore the fact that the use of considerably different models, pre-trained on considerably different corpora with considerably different objectives may result interfere with how certain we can be about our conclusions with the drawbacks of TAPT. We can only with certainty attest the inverse, that TAPT benefits multilingual MLM models whose pretraining corpus contains the least amount of task-related data.

Time and resource constraints restricted our approach. Model performance was limited by training on a limited number of epochs, although this was the same for all models. A more fair comparison could consist of allowing hyper-parameter tuning of each model according to their respective training objective, as it was clear that some models needed more epochs than others. We predict, however, that fairness aside, the sheer size of XLM-RoBERTa-Large will continue to dominate and that language adaptation is responsible for all gains witnessed.

We encountered a few possible limitations in terms of replicability. The first limitation is the availability of both data sets used in our research. Due to licensing issues, the data sets are not available for public use. Access to the headlines is only possible through pre-approval. Thus, the work here can only be reproduced or expanded given access to this data set.

A second limitation is imposed by the nature of randomness with respect to neural networks. We have attempted to minimise the risk by using seeds whenever possible and noting them in our scripts. However, this still cannot account for all possible differences in performance between systems or GPUs. Even with different runs of the same model with the same seed, we occasionally encountered different results. This was the case including the final evaluation of our best-performing, which changed, albeit minimally, in performance despite using the same seed. We have attempted to mitigate this randomness by averaging the performance across five seeds in our fine-tuning and evaluation stage. Although there still exists the possibility of spurious spikes in performance, we expect that our observations should still hold.

9. Bibliographical References

Michele Banko, Vibhu O Mittal, and Michael J Witbrock. 2000. Headline generation based on statistical translation. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 318–325.

Ana Barić, Laura Majer, David Dukić, Marijana Grbeša-zenzerović, and Jan Šnajder. 2023. Target two birds with one STONE: Entity-Level sentiment and tone analysis in Croatian news headlines. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 78–85.

Valerio Basile and Malvina Nissim. 2013. Sentiment analysis on Italian tweets. In *Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 100–107.

Vuk Batanović and Maja Miličević Petrović. 2022. Cross-level semantic similarity for Serbian newswire texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1691–1699, Marseille, France. European Language Resources Association.

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. GoodNewsEveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1554–1566, Marseille, France. European Language Resources Association.

Peter Bourgonje, Julian Moreno Schneider, and Georg Rehm. 2017. From clickbait to fake news detection: An approach based on detecting the stance of headlines to articles. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 84–89, Copenhagen, Denmark. Association for Computational Linguistics.

Stevo Bozinovski. 2020. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3).

D Brozovid. 1991. Serbo-Croatian as a pluricentric language. *Pericentric languages. Differing norms in different nations*, pages 347–80.

Ranko Bugarski. 2019. Past and current developments involving pluricentric Serbo-Croatian and its official heirs. *Language Variation. A Factor of Increasing Complexity and a Challenge for Language Policy within Europe. Budapest: Research Institute for Linguistics, Hungarian Academy of Sciences*, pages 105–114.

Yanqing Chen and Steven Skiena. 2014. Building sentiment lexicons for all major languages. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2019. Electra: Pre-training text encoders as discriminators rather than generators. In *International Conference on Learning Representations*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Kia Dashtipour, Soujanya Poria, Amir Hussain, Erik Cambria, Ahmad YA Hawalah, Alexander Gelbukh, and Qiang Zhou. 2016. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8:757–771.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Kelvin Du, Frank Xing, and Erik Cambria. 2023. Incorporating multiple knowledge sources for targeted aspect-based financial sentiment analysis. *ACM Transactions on Management Information Systems*.

Unni Cathrine Eiken, Anja Therese Liseth, Hans Friedrich Witschel, Matthias Richter, and Chris Biemann. 2006. Ord i Dag: Mining Norwegian daily newswire. In *Advances in Natural Language Processing*, pages 512–523, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jelena Golubović and Charlotte Gooskens. 2015. Mutual intelligibility between West and South Slavic languages. *Russian linguistics*, pages 351–373.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Michael A Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568.

Mohammed Jabreel, Fadi Hassan, and Antonio Moreno. 2018. Target-dependent sentiment analysis of tweets using bidirectional gated recurrent neural networks. *Advances in Hybridization of Intelligent Methods: Models, Systems and Applications*, pages 39–55.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent Twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160.

Shahrukh Khan and Mahnoor Shahid. 2022. Hindi/bengali sentiment analysis using transfer learning and joint dual input learning with self attention. *arXiv preprint arXiv:2202.05457*.

Evgeny V. Kotelnikov. 2021. Current landscape of the Russian sentiment corpora. In *Komp'juternaja Lingvistika i Intellektual'nye Tehnologii*, pages 433–444.

Rohit Kulkarni. 2018. A Million News Headlines.

Sofia Lee and Jelke Bloem. 2023. Comparing domain-specific and domain-general BERT variants for inferred real-world knowledge through rare grammatical features in Serbian. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 47–60.

Hong Li, Xiwen Cheng, Kristina Adson, Tal Kirshboim, and Feiyu Xu. 2012. Annotating opinions in German political news. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1183–1188, Istanbul, Turkey. European Language Resources Association (ELRA).

Zheng Li, Ying Wei, Yu Zhang, Xiang Zhang, and Xin Li. 2019. Exploiting coarse-to-fine task transfer for aspect-level sentiment classification. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 4253–4260.

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller. 2020. Does BERT need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association*, 27(4):584–591.

Hankun Liu, Daojing He, and Sammy Chan. 2021. Fraudulent news headline detection with attention mechanism. *Computational Intelligence and Neuroscience*, 2021:1–7.

Adela Ljajić, Nikola Prodanović, Darija Medvecki, Bojana Bašaragin, and Jelena Mitrović. 2022. Uncovering the reasons behind COVID-19 vaccine hesitancy in Serbia: Sentiment-based topic modeling. *Journal of Medical Internet Research*, 24(11):e42261.

Nikola Ljubešić and Kaja Dobrovoljc. 2019. What does neural bring? Analysing improvements in morphosyntactic annotation and lemmatisation of Slovenian, Croatian and Serbian. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 29–34, Florence, Italy. Association for Computational Linguistics.

Nikola Ljubešić and Davor Lauc. 2021. BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42, Kiyv, Ukraine. Association for Computational Linguistics.

Xiaofei Ma, Peng Xu, Zhiguo Wang, Ramesh Nallapati, and Bing Xiang. 2019. Domain adaptation with BERT-based domain classification and data selection. In *Proceedings of the 2nd Workshop on Deep Learning Approaches for Low-Resource NLP (DeepLo 2019)*, pages 76–83, Hong Kong, China. Association for Computational Linguistics.

Iuliia Makogon and Igor Samokhin. 2021. Targeted sentiment analysis for Ukrainian and Russian news articles. In *International Conference on Information and Communication Technologies in Education, Research, and Industrial Applications*, pages 538–549. Springer.

Gati L. Martin, Medard E. Mswahili, and Young-Seob Jeong. 2021. Sentiment classification in Swahili language using Multilingual BERT.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešič. 2022. The ParlaSent-BCS dataset of sentiment-annotated parliamentary debates from Bosnia-Herzegovina, Croatia, and Serbia. *arXiv preprint arXiv:2206.00929*.

Mustafa Melih Mutlu and Arzucan Özgür. 2022. A dataset and BERT-based models for targeted sentiment analysis on Turkish texts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 467–472.

Marco A. Palomino and Farida Aider. 2022. Evaluating the effectiveness of text pre-processing in sentiment analysis. *Applied Sciences*, 12(17).

Isabel Papadimitriou, Kezia Lopez, and Dan Jurafsky. 2023. Multilingual BERT has an accent: Evaluating English influences on fluency in multilingual models. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1194–1200.

Ioannis Pavlopoulos. 2014. *Aspect based sentiment analysis*. Ph.D. thesis, Athens University of Economics and Business.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4933–4941.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Peter Rupnik, Taja Kuzman, and Nikola Ljubešić. 2023. BENCHić-lang: A benchmark for discriminating between Bosnian, Croatian, Montenegrin and Serbian. In *Tenth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2023)*, pages 113–120, Dubrovnik, Croatia. Association for Computational Linguistics.

Hassan Saif, Miriam Fernández, Yulan He, and Harith Alani. 2013. Evaluation datasets for Twitter sentiment analysis: a survey and a new dataset, the STS-Gold. In *1st Interantional Workshop on Emotion and Sentiment in Social and Expressive Media: Approaches and Perspectives from AI (ESSEM 2013)*.

Tomás Alves Salgueiro, Emilio Recart Zapata, Damián Furman, Juan Manuel Perez, and Pablo Nicolás Fernández Larrosa. 2022. A Spanish dataset for targeted sentiment analysis of political headlines. *Memorias de las JAIIO*, 8(2):92–97.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and Korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725.

Ravi Shekhar, Mladen Karan, and Matthew Purver. 2022. Coral: a context-aware Croatian abusive language dataset. *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 217–225.

Carlo Strapparava and Rada Mihalcea. 2007. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic. Association for Computational Linguistics.

Wilson L Taylor. 1953. Cloze procedure: A new tool for measuring readability. *Journalism quarterly*, 30(4):415–433.

Gaurish Thakkar, Nives Mikelic Preradovic, and Marko Tadić. 2023. Croatian film review dataset (cro-FiReDa): A sentiment annotated dataset of film reviews. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 25–31, Dubrovnik, Croatia. Association for Computational Linguistics.

Matej Ulčar and Marko Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT. In *Text, Speech, and Dialogue*, pages 104–111, Cham. Springer International Publishing.

Juan Vásquez, Helena Gómez-Adorno, and Gemma Bel-Enguix. 2021. Bert-based approach for sentiment analysis of Spanish reviews from TripAdvisor. In *IberLEF@ SEPLN*, pages 165–170.

Xiaojun Wan. 2008. Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 553–561, Honolulu, Hawaii. Association for Computational Linguistics.

Chunli Xiang, Junchi Zhang, Fei Li, Hao Fei, and Donghong Ji. 2022. A semantic and syntactic enhanced neural model for financial sentiment analysis. *Information Processing & Management*, 59(4):102943.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene superglue benchmark: Translation and evaluation. *arXiv preprint arXiv:2202.04994*.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30.

Peilin Zhou, Zeqiang Wang, Dading Chong, Zhijiang Guo, Yining Hua, Zichang Su, Zhiyang Teng, Jiageng Wu, and Jie Yang. 2022. METS-CoV: A dataset of medical entity and targeted sentiment on COVID-19 related tweets. *Advances in Neural Information Processing Systems*, 35:21916–21932.

Sven Ćurković, David Dukić, Marin Petričević, and Jan Šnajder. 2022. Takelab retriever.