Master Thesis

INCIDENT IN ZAGREB, SELF-SUPERVISED TASK ADAPTATION PERFORMED: Impact of Task Adapting on Transformer Models for Targeted Sentiment Analysis in Croatian Headlines

Sofia Lee

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Dr. Isa Maks
2^{nd} reader: Dr. Hennie van der Vliet

Submitted: June 30, 2023

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. This thesis, produced as part of an internship at TakeLab FER, concerns 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.

Declaration of Authorship

I, Sofia Marie Lee, declare that this thesis, titled *INCIDENT IN ZAGREB, SELF-SUPERVISED TASK ADAPTATION PERFORMED: Impact of Task Adapting on Transformer Models for Targeted Sentiment Analysis in Croatian Headlines* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: June 30, 2023

Signed: Sofia Marie Lee

Acknowledgments

First and foremost, I would like to thank my mentor and thesis supervisor, Dr. Isa Maks. Thank you for your encouragement and support throughout this time from a distance, over a thousand kilometres away. I would like to thank the team at TakeLab FER, especially Dr. Jan Šnajder and Ivan Krišto for taking a chance on me and having me on board the Retriever project. I would like to thank David Dukić in particular for his patience and guidance as my supervisor. I would also like to extend particular thanks to colleagues at TakeLab FER, Laura Majer and Fran Jelenić.

I would like to thank Dr. Nikola Ljubešić and Dr. Tanja Samardžić for their inspiring work and for believing in me. I would like to thank I would like to thank Mladen Fernežir, Dr. Davor Lauc, and my colleague Ivan Rep for help with working with BERTić.

I would also like to extend my gratitude to those who have served as academic mentor figures in the past. I would like to acknowledge Dr. Tijmen Pronk for his devotion to Croatian dialectology as inspiration for me to continue with my studies in BCMS. I would also like to acknowledge my bachelor's thesis supervisor, Dr. Jelke Bloem, who helped me get into my first conference and continued support me even long after the thesis was over. Additionally, I would like to thank more personal mentor figures, Lily Lynch and Lidija Andonov, for being there for me and helping me make sense of everything.

I would like to thank my roommate Ela for giving me support and keeping me sane throughout my time here. I would like to thank Sven for his company and hospitality in having me at his workspace, along with Tin, Andro, Lana, Andrea, Sara, Luana, and Hrvoje. Also, I would like to thank Sara and Gabi for co-working with me. Thank you to Stefan and his family for giving me a warm bed that one night and a plate full of *prženica* for breakfast in Frankfurt.

I would also like thank my countless friends, old and new, which I saw during my time here: Tea, David, Ivana, Svetozar, Dora, Iva, Matea, Marika, Tomislav, Jasmin, Marko, and Zrinka. Thank you to Brin and Urška for taking me to Rijeka. Thank you to Ana Br. for the unforgettable tour of Dubrovnik. Thank you to Petar and Miloš (in Serbia), who were always there for me whenever I was up late with something on my mind.

And I would like to especially thank Iva M, Sara P, Martin, Grgur, Selma, Xenia, and Ameena. I am eternally grateful for your support.

And finally, I cannot forget Srna and Ursula, the office dogs, and the countless cats that I have met in Croatia. For everyone else I have left out of this—I apologise—you are not forgotten; I could write another 15,000 words about

all the people who have left an impact on me, but I will have to stop here.

Thank you everyone for welcoming me into your world and being with me on this journey so far. Thank you for your warmth and kindness.

Od srca, hvala.

List of Tables

Contents

Chapter 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., 2020), 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., 2018). 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. Their flexibility and adaptability is truly a part of the interest it has generated. The question, then, is for the best way to approach 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 sub-domain with its own associated set of unlabelled task data. See Table 1.1 for an example breakdown of domains and sub-domains related to news headlines. 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.

Types of data		Description	Examples
Unlabelled	LM domain	Everything that a LM has been pre-trained on; usually domain-agnostic	Massive collections of raw text in one or more languages
	Domain	All data that pertains to a particular field	Large collections of news text, including headlines
	Task	Data specific to a task within a domain, essentially a 'subdomain'	All news headlines, but no body text
Labelled	Task	Annotated data used to fine-tune a model for a task	A selection of polarising headlines with sentiment labelled

Table 1.1: Different types of data, illustrating the differences between domains and subdomains, ordered from largest to smallest. Each level overlaps with the previous to some extent. Unlabelled data can be used for pre-training while labelled data is used for fine-tuning.

Our task in particular is targeted sentiment analysis (TSA), 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). We perform this task on Croatian headlines. 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, sentiment may also be explicitly or implicitly expressed, which will also have to be identified by the model.

The focus of our work is in the Croatian language, specifically with news from Croatian news portals. Croatian is a low-resource language, meaning that there is limited amount of research being performed on the language as well as limited number of data sets and relevant tools available. Although there has been work recently performed on targeted sentiment analysis in Croatian news (Barić et al., 2023; Thakkar et al., 2023; Babić et al., 2021), we will be the first to our knowledge to attempt this with the inclusion of TAPT as proposed by Gururangan et al. (2020).

1.1 Motivation

This work is part of an internship at TakeLab FER (Faculty of Electrical Engineeering and Computing, in Croatian *Fakultet za elektrotehniku i računarstvo*). TakeLab FER is a text analysis and knowledge engineering research group based in Zagreb, Croatia. It is affiliated with the University of Zagreb and focuses on natural language processing, machine learning and text analysis. With a focus on Croatian and the neighbouring related languages, TakeLab is one of the leading research groups in the sub-field of Slavic NLP.

One of the main projects of TakeLab is Retriever, a platform that performs real-time text mining on Croatian news (Ćurković et al., 2022). At the time of this writing, Retriever has analysed over eight million articles and continues to analyse a large quantity of new articles daily. With the data collected, Retriever provides an interface for data scientists to study how the news cycle covers certain topics or phrases. The current implementation allows users to search for named entities or phrases and then returns a graph of how many articles match the query over the period of time covered. There are currently over 40 users of Retriever, and access to the platform is given out on a per-request basis.

Our goal is to research approaches to targeted sentiment analysis for Croatian headlines with the aim of implementing a classifier in the TakeLab Retriever article text mining pipeline. Such a classifier would be used to produce data about trends in Croatian news for the purpose of analysis in social and political science. Although TakeLab has produced such a model already (Barić et al., 2023), they are interested in employing their massive amounts of task-relevant data to perform TAPT to improve performance.

1.2 Research questions

Our main question is the following:

> Will task-adaptive pre-training yield improvements in language model performance (F_1-score) in a targeted sentiment analysis task for Croatian headlines?

Our goal will therefore be to study the impact of TAPT on a number of language models. We will also explore the following sub-questions:

1. What challenges remain for our highest-performing model?
2. How does Croatian as a low-resource language affect our task?

1.3 Contributions

In our research, we make the following contributions to our field:

1. We expand on the work done by Barić et al. (2023) on targeted sentiment in Croatian news headlines.

2. We implement and expand on work in domain and task adaptation by Gururangan et al. (2020), by applying their method of TAPT to a new language and domain.

3. We contribute to the field of Croatian NLP, which focuses on a low-resource language. Thus, we contribute to the field of low-resource languages as well.

1.4 Outline

This work is broken into several chapters. In this current chapter, Chapter 1, we introduced our research interests of task adaptation, targeted sentiment analysis, and Croatian NLP. We situated our research in the context of the work at TakeLab FER.

The next chapter, Chapter 2, provides a literature review of previous work relating to these topics.

Afterwards, in Chapter 3, we describe the data that will be used both as pre-training and fine-tuning, both provided by TakeLab, for the models. We remark on some statistical and linguistic features of the data sets.

The chapter after that, Chapter 4, introduces the models which we will study, breaking them down into Croatian-specific and massively multilingual groups. This chapter also describes the training process and task, as well as evaluation metrics used.

Subsequently, Chapter 5 details the results of each stage of training as well as the final outcome of the task. We also perform an in-depth error analysis of the highest performing model.

In the following chapter, Chapter 6, we contextualise our work in our research interests. We then describe limitations, possible directions for future work, as well as biases and ethical concerns.

Finally, our concluding chapter, Chapter 7, provides an overview of our work and concluding remarks.

Chapter 2

Background and related work

2.1 Domain and task adaptation in transformer models

BERT (Devlin et al., 2018) is a powerful language model, but one that is still in need of further fine-tuning before deployment. There is, thus, a large amount of work done on domain and task adaptation for pre-built BERT models. 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 resistent 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. (2019), 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.

2.2 Sentiment analysis

Sentiment analysis, also referred to as opinion mining, is a broad class of NLP tasks which aim to extract sentiment from given text (Katrekar and AVP, 2005). Sentiment can take a variety of schemes such as polar (i.e. NEG, NTR, or POS), mixed (Kenyon-Dean et al., 2018) or even finer ones involving various numbers of emotions (Ekman et al., 1999; Demszky et al., 2020). Work in sentiment analysis has included identifying 'fake news' (Alonso et al., 2021), studying reviews of movies or products (Hu and Liu, 2004), or analysing opinions in tweets (Hasan et al., 2019). The expansion of Web 2.0 gave rise to more sources of data for sentiment analysis, in the form of blog posts, comments, and reviews, often with ratings included (Ravi and Ravi, 2015). Early attempts at sentiment analysis involve extensive feature engineering based on sentiment lexicons (Yi et al., 2003).

Recent years have given 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). Thet et al. (2010), for example, explore automated methods of identifying sentiment strength and orientation towards different aspects of films through reviews on discussion boards. For a survey of work on aspect-based sentiment analysis, see Zhang et al. (2022).

Another related fine-grained sentiment analysis task is targeted sentiment analysis (TSA), the task which we will be exploring. TSA differs from aspect-based sentiment analysis due to its aim on 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., 2022a).

2.2.1 News domain

Our interest in sentiment analysis is situated within the news domain, which encompasses text and media that concern news articles and journals. This is a domain that may be global or further sub-divided by region or topics, such as sports, finance, or current events. In recent years, much effort has been devoted to the detection of fraudulent or 'fake' news (Naredla and Adedoyin, 2022), particularly in the field of social and media forensics (Zhou et al., 2022b).

One particularly notable feature of the news domain are headlines. Headlines are unique due to the fact that they tend to summarise a body of text, identify the key topics, employ a telegraphic style of writing that lacks function words, and are designed explicitly to pull readers in. Tasks related to headlines can include tasks such as 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 ranging from 2003 to 2021.

Within the news domain, targeted sentiment analysis is also used for analysing headlines. 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.

2.2.2 Non-English sentiment analysis

Although sentiment analysis has been a crucial part of NLP tradition since the start of the field, 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, especially prior to the spread of the Internet. 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, particularly in more recent years. 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.

2.3 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 including Germany, the United Kingdom, the United States of America, and Chile. 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 history in a paper written in Croatian (Bozinovski, 2020). Despite its role in machine learning history, however, support for the language has dramatically plummeted in the intervening decades. Croatian may presently be considered a low-resource language, a language which lacks high-quality data or resources (Hedderich et al., 2020). Low-resource languages are contrasted with high-resource languages, such as English or German, for which there exists ample high quality human-annotated data sets, benchmarks, research oriented towards these languages, as well as institutional and corporate support (Bender, 2019). 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; Ulcar and Robnik-Sikonja, 2020).

2.3.1 Croatian NLP

A considerable amount of research in Croatian NLP is focused on building databases for the language. Such databases include emotion lexicons

(Ljubešić et al., 2020), affective word databases Ćoso et al. (2019), and a context-aware abusive language database (Shekhar et al., 2022). Both bilingual and monolingual corpora exist for the language. Some bilingual resources include an English-Croatian parallel corpus (Tadić, 2000) and a Croatian-Slovene bilingual treebank (Agić et al., 2014). Monolingual corpora include the Croatian Web Corpus, hrWaC (Ljubešić and Erjavec, 2011; Ljubešić and Klubička, 2014); the CLASSLA-hr corpus (Ljubešić, 2021); and the Riznica Croatian language corpus (Brozović Rončević et al., 2018). These data sets provide the bulk of training data for Croatian language models. For example, the hrWaC is employed by Ljubešić et al. (2013) and combined with a Slovene data set to produce an early Named Entity Recognition model.

Early research, however, was sparse, but the formation of the Regional Linguistic Data Initiative (ReLDI) (Samardžić et al., 2015) marked a large step forward. This initiative was formed as a means to bring together researchers, professionals and institutions in Switzerland, Serbia and Croatia to focus on lesser-researched regional languages, which also included Croatian. This resulted in a number of tools such as the part-of-speech tagger, `reldi-tagger`, which supports Croatian among other languages. An overview of more recent developments, including tools and data sets, brought in by the application of neural networks to Croatian, Slovene and Serbian NLP tasks is provided by Ljubešić and Dobrovoljc (2019).

The strong mutual intelligibility yet official political distinction between Croatian, Bosnian, Montenegrin and Serbian, has resulted in interest in automated differentiation these languages. This task is considerably difficult, even for human annotators, but there is some success with a feature-based Twitter geo-tagging method devised by Ljubešić et al. (2018). This task is performed again as one of the criteria for evaluating a BCSM-specific BERT model (Ljubešić and Lauc, 2021). Finally, this work is consolidated into an official benchmark for differentiating the languages, BENCHić (Rupnik et al., 2023).

In addition to massively multilingual models such as Multilingual BERT (mBERT) (Devlin et al., 2018) or XLM-RoBERTa (Conneau et al., 2019) that support the top 100 or so most-resourced languages, there also exist language-specific models. As of this writing, only two language-specific BERT models have been created. The first model is CroSloEngual BERT (pronounced 'Crosslingual BERT') or cseBERT, produced by Ulcar and Robnik-Sikonja (2020) as part of research on the impacts of transfer learning between related languages on BERT models in comparison to massively multilingual models. Ulcar and Robnik-Sikonja (2020) compare the performance of cseBERT as well as FinEstBERT, a Finnish and Estonian BERT model with an English base, to mBERT to demonstrate that their approach of transfer learning with fewer languages outperforms mBERT in tasks of named entity recognition, dependency parsing and part-of-speech tagging. However, later work show that cseBERT performs similarly or poorly in comparison to XLM-RoBERTa in some named entity recognition tasks (Prelevikj and Žitnik, 2021). 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 morpho-syntactic tagging, named entity recognition, social media geo-location and commonsense causal reasoning. Despite complications with its ELECTRA base, BERTić serves as a significant baseline for NLP research in the region. 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).

The fact that BERTić has been trained as a BCMS-focused model has also led to its use in the closely related but politically distinct, Serbian. We include them here as there is a high potential for application in Croatian as well. Such work include short text semantic similarity and 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). To our knowledge, no work has been done specifically for Bosnian or Montenegrin.

Sentiment analysis in recent years has become more of a focus in Croatian NLP spheres. Thakkar et al. (2023) provide a sentiment-annotated data set of Croatian film reviews, annotated on the sentence level. The work closest to ours is the recent paper by Barić et al. (2023), which introduces the SToNe data set, a data set 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.

2.4 Conclusion of background and related works

In this chapter, we provided an overview of works related to ours in terms of domain adaptation, news headlines, targeted sentiment analysis and Croatian NLP. We found that there is a lot of work related to all of these topics, although the intersection of them is particularly scarce. A few papers (Gururangan et al., 2020; Barić et al., 2023; Ulcar and Robnik-Sikonja, 2020; Ljubešić and Lauc, 2021) form the bulk of our inspiration. Although there is a wealth of work in all of these topics, our combination of approaches is, to our knowledge, a first.

Chapter 3

Data

In this chapter, we will describe the sets of data used for this research project. We will situate the data in its relevance to the domain and task at hand, as well as delve into information about the peculiarities of the data.

3.1 TakeLab Retriever headlines

The TakeLab Retriever (Ćurković et al., 2022) includes a web scraper that routinely trawls news articles from assorted Croatian web portals, scraping and performing text mining on each article. Each article contains headlines as well as meta-data such as publishing time. Currently text-mining includes named entity recognition, performed by a fine-tuned BERTić (Ljubešić and Lauc, 2021), as well as named-entity linking, which joins named entities found into a database of entities from Wikipedia. Retriever is intended to be used by political scientists and sociologists for the purpose of studying trends in news in Croatia. The platform also includes a front-end that allows users to input key phrases, select portals, and view a chart of frequency of a search term across a period of time.

The data set used for the purposes of our research was extracted from Retriever on 26 April, 2023. The data set consists of 8.34 million headlines. The time-span ranges from 1 January 2001 to day of retrieval. Headlines come from 42 different portals, which largely include Croatian new sites, although there are also three non-Croatian portals, of which two are Serbian and one is Bosnian. Some regional publications are also included, such as glas-slavonije.hr, a portal for news from the Slavonija region of Eastern Croatia, and slobodnadalmacija.hr, which consists of news from Dalmatia in the coastal region of Croatia to the south. Regional publications occasionally use regionalisms, such as the Ikavian spelling, especially in quotes. A considerably small proportion comes from blogs. Portals cover a diverse range of political leanings, including centrist news sources, tabloids, and fringe right-wing publications. See Table 3.3 for an overview of the distribution articles per portal.

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,

Statistic	Value
# of unique tokens	1.02 million
Average tokens per line	12.43
Longest headline by tokens	85
Longest headline by characters	420
Total number of tokens	91.38 million

Table 3.1: TakeLab Retriever headlines statistics

Corpus	Overlap
ENGRI	69.26%
hrWaC	62.22%
Riznica	59.77%
CLASSLA-hr	57.30%
CLASSLAwiki-hr	47.48%

Table 3.2: The amount of vocabulary overlap between the TakeLab Retriever headlines data set and a number of other Croatian data sets. The percentage indicates how many of the top 10,000 words are shared between data sets.

thus being particularly suitable for the purpose of task-adaptive pre-training (TAPT). See Table 3.1 for more general information about this data set.

In order to study the data set, we performed tokenisation using the ReLDI tokeniser, a rule-based tokeniser for Croatian provided by the CLASSLA Python package. The tokeniser identified 1.02 million unique tokens, including punctuation and case variations. Adapting the procedure for domain comparison from Gururangan et al. (2020), we compared the top 10,000 word frequencies with that of several other Croatian data sets: ENGRI, a news data set (Bogunović et al., 2021); Riznica, a literary corpus (Brozović Rončević et al., 2018); hrWaC, a general web corpus for the .hr TLD, dated pre-2015 (Ljubešić and Erjavec, 2011); CLASSA-hr, a more recent .hr TLD crawl from 2019-2020 (Ljubešić, 2021); and CLASSLAwiki-hr, a crawl of the Croatian wikipedia from 2021 (Ljubešić, 2021). Detailed numbers are provided in Table 3.2. For an analysis of the data set's vocabulary overlap for each model we test, see Table 4.3.

Our comparison showed that the headlines data set indeed aligned most closely to the news data set, while differing the most from the Wikipedia data set. The high degree of overlap was expected, considering that news headlines and news body text were part of a more general Croatian news domain. Differences were likely caused by the tendency for headlines to be short, telegraphic, and focused on named entities.

3.1.1 Pre-processing

We performed a de-duplication process, reducing the headline count by 11.91%, to 7.35 million headlines. De-duplication consisted of first removing exact matches and then a fuzzy match process in which all headlines are sorted by length and then recursively compared with subsequent headlines in terms of similarity. In many cases, duplicates existed because of the tendency for por-

Portal	Headlines
index.hr	1,079,324
24sata.hr	956,222
jutarnji.hr	857,050
vecernji.hr	850,612
tportal.hr	765,678
dnevnik.hr	610,363
net.hr	554,078
glas-slavonije.hr	479,840
slobodnadalmacija.hr	410,470
direktno.hr	305,799
rtl.hr	240,735
hr.n1info.com	210,648
narod.hr	185,780
hrt.hr	168,121
novilist.hr	121,089
dnevno.hr	118,619
telegram.hr	77,518
h-alter.org	68,499
face.ba	48,790
priznajem.hr	27,021
plusportal.hr	25,888
bug.hr	24,751
geopolitika.news	18,440
logicno.com	18,103
teleskop.hr	16,384
tris.com.hr	13,995
netokracija.com	13,640
intermagazin.rs	12,663
lupiga.com	12,343
hop.com.hr	10,431
tribun.hr	8,368
crol.hr	5,972
paraf.hr	5,669
svijetokonas.info	3,917
liberal.hr	3,795
forum.tm	3,727
istinomprotivlazi.info	2,888
2012-transformacijasvijesti.com	2,662
homunizam.wordpress.com	671
dokumentarac.hr	476
srbnovine.blogspot.com	309
novisvjetskiporedak.wordpress.com	143

Table 3.3: Distribution of headlines per portal in data set

tals to republish the text of other portals. We also pruned one-word headlines during this pre-processing stage, resulting in the deletion of 5,027 headlines.

To prevent data spillage, we performed fuzzy matching for headlines of similar lengths to remove headlines that appear in the SToNe validation and test set (described below in Chapter 3.2). Of these headlines, 17 could not be removed. We could not find them at all in the Retriever Headlines data set, perhaps due to an earlier pre-processing stage, so no further attempts were made to remove these 17 headlines. Headlines from the rest of the SToNe set, i.e. the training set, were maintained due the possibility that seeing the same data in two contexts could be beneficial (Gururangan et al., 2020).

For every 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.

3.2 SToNe

The second data set in use is the SToNe data set (Barić et al., 2023). The data set was provided by the team at TakeLab directly, with all annotation and processing already performed for a previous study. This data set is an annotated sub-data set of the TakeLab Retriever headlines data set containing named entities (NEs) as well as labels for the sentiment towards the NEs and the general tone of the headline. NEs are found by using a BERTić model fine-tuned on the NE recognition task, which achieves state-of-the-art results in the Croatian news domain (F_1 = 89.21) (Ljubešić and Lauc, 2021). Headlines that have several NEs may be included multiple times, with a different NE targeted. NEs were divided into four types: PER for people, ORG for organisations, LOC for locations, and MISC for everything else. The PER label notably also includes the names of ethnic groups, regardless of whether they refer to a single specific person, a non-specific person of the ethnic group, or the ethnic group as a whole. See Table 3.4 for a detailed description of the NE description and distribution across the data set.

3.2.1 Annotation

In this section, we will describe the annotation process carried out by TakeLab on the SToNe data set. Annotation was performed by ten annotators, all native speakers of the Croatian language and were Croatian by origin. All annotators were students of the University of Zagreb, although some were from the Faculty of Humanities and Social Sciences (*Filosofski fakultet*) and others from the the Faculty of Electrical Engineering and Computing (*Fakultet elektrotehnike i računarstva*). Annotators all lived in Zagreb or nearby areas, commuting regularly to the city for university, but did not necessarily all originate in Zagreb. The annotators were described of being of typical university age and as generally politically 'left-leaning'. Further details about the annotators were not provided. The tool used for annotation was ALANNO, a publicly available tool also developed by TakeLab (Jukić et al., 2023). Annotation was

Type	Count	Description
PER	1,187 (41.58%)	All physical people. This also specifically includes ethnic groups.
ORG	694 (24.31%)	An entity such as an organisation, association, government body or other similar collective
LOC	720 (25.22%)	All entities that refer to a geographic location such as countries, cities, towns, villages, rivers, lakes, and so on
MISC	256 (8.97%)	Everything that does not fit into the above categories, such as history events, holidays, titles of films, and so on

Table 3.4: Named entities with frequency and description. Note that this displays on the distribution of named entities that are the target of sentiment analysis. Headlines may additionally contain other named entities. Descriptions are largely taken from the internal annotations instructions document, translated from Croatian, with some clarifications added by the author.

carried out by each annotator independently within 14 person-hours, with each headline being assigned at random to precisely six annotators.

Annotators were presented with several tasks. First, a headline was presented with a bolded NE. In order to eliminate any false positives from the NE recognition stage, annotators were tasked to verify that the bolded text indeed was a NE and that the correct label was used. In the following example, we provide a sentence with three NEs. Croatia is labelled LOC as it refers to a country, COVID-19 is labelled MISC and refers to a specific outbreak of a virus related to a pandemic, and Krunoslav Capak is a physical person and thus belongs to PER.

> Podsjetimo, **Hrvatska** [LOC] je preliminarno dogovorila cjepiva protiv **Covida-19** [MISC] dovoljno za više od polovice stanovništva, s tri proizvođača, rekao je primarijus **Krunoslav Capak** [PER].
>
> As a reminder, **Croatia** [LOC] has preliminarily agreed on vaccines against **COVID-19** [MISC] enough for over half of the population, with three producers, said chief physician **Krunoslav Capak** [PER].

For both targeted sentiment and tone, a ternary annotation scheme was used as opposed to more granular approaches. The scheme consisted of *negative* (NEG), *neutral* (NTR), and *positive* (POS) labels. The authors had selected this approach with the aim to gather only the most immediate judgement of the annotators when evaluating a headline. Annotators were first instructed to look at headlines holistically, assigning a tone label to the general headline, and then focused specifically on the sentiment assigned to the named entity. When assessing sentiment, the annotators were instructed to pay attention primarily to sentiment from first impression, secondarily from wording of the headline, and lastly with general knowledge of the entity in question. The text of the article was not provided due to the intention of wanting the responses to be based solely on the text of the headline rather than any other added

Headline	Translation	Sentiment
SRAMOTA **U Solinu** se djeca nemaju gdje liječiti, roditelji očajni	SHAME In **Solin** children do not have anywhere to be treated, parents desperate	negative
U Solinu radi samo jedna pedijatrica, roditelji traže hitno rješenje	Only one pediatrician working in **Solin**, parents urgently seek solution	negative
U Solinu nastupio nedostatak liječničkog kadra	A lack of medical staff emerges in **Solin**	neutral

Table 3.5: A demonstration of how framing affects sentiment analysis in three different headlines. The target selected for each headline is **Solin**, a town in Dalmatia, and therefore falls under the type *location*.

context. See Table 3.5 for example of annotation in which the same event is described with differing sentiment assigned to the target.

The resultant annotated data set contains 2,855 headlines. 653 headlines were labelled NEG, 1,486 as NTR, and 716 as POS. There was an inter-annotator agreement rating of κ = 0.493 for targeted sentiment analysis according to the Fleiss-kappa metric, indicating moderate agreement (Barić et al., 2023).

3.2.2 Pre-processing

As done by Barić et al. (2023), 548 headlines were removed for having conflicting annotations, leaving 2,308 headlines in the final set used for our use. In order to retain the approach set out by the authors, we did not perform any additional pre-processing on the SToNe data set other than discarding tone information, which was not relevant to our research. The data set's original training, validation and evaluation split ratio of 70:10:20 was deployed as provided.

3.2.3 Exploratory analysis on SToNe training set

In interests of seeing what the final training set contains, we perform exploratory data set analysis on the training portion of the SToNe set. Out of 1,614 instances, 341 are NEG, 810 are NTR, and 463 are POS, indicating that targeted sentiment is predominantly neutral, split evenly between neutral and a polar sentiment, but skews positive (1.34 times) when polar.

When analysing the polar tags for each NE type, we find that PER is 1.5 times more likely to be POS than NEG, while MISC is 3.71 times more likely to be POS. ORG shows only a marginal skew towards POS (1.12 times) but differs from the general dataset bias. Conversely, LOC skews slightly towards NEG, with 1.24 times more NEG than POS. In short, these indicate that there is a significant bias towards positive sentiment for PER and MISC, while LOC tends to have a considerable negative bias. More detailed figures are presented in

	NEG	NTR	POS	Total
PER	157	264	236	657
ORG	95	181	106	382
LOC	72	312	58	442
MISC	17	53	63	133
Total	341	810	463	1,614

Table 3.6: An analysis of sentiment distribution across NE types in the SToNe training set.

Type	Unique	Count	Repetition
PER	561	657	14.61%
ORG	306	382	19.89%
LOC	308	442	30.32%
MISC	128	133	3.76%
Total	1,303	1,614	19.27%

Table 3.7: Unique forms of NE with a percentage of repetition for each type. A higher repetition value indicates that the same NE form appears multiple times, whereas a low value indicates a larger diversity of entities represented in the dataset.

Table 3.6.

Although the data set comes pre-labelled with NE type, we attempted to break down entities into sub-types to get a more detailed understanding of the domain represented. We sorted every type first alphabetically, then by frequency of appearance. See Table 3.7 for a count of unique forms of named entities of each type. Afterwards, we sampled the top 50. We noted types we found that also did not appear in the top 50 with examples.

Generally, we found that the location portion of the data set was predominantly represented by countries, but mostly China, Russia, Bosnia & Herzegovina, Italy and the USA. Neighbouring countries of Serbia and Montenegro were also present. The regional focus of the data, however, is very evident with the amount of Croatia-specific locations represented as well as the number of times Croatia itself is selected. Regions outside of Europe are hardly ever targeted except ones that have geopolitical (China, Russia, USA, Crimea) or cultural relevance (London, Buckingham Palace). Interestingly, although most countries are portrayed in neutral or negative light, Croatia is significantly more often portrayed positively than negatively. Detailed figures are provided in Table 3.8.

The Croatian context is also present when examining the distribution of the PER label. Local political figures predominated our sample, while global political figures were largely ones of particular geopolitical (Putin, Biden) or near-regional relevance (Vučić). Although, we found it very difficult to divide sports figures into regional or global, as many figures appeared to be affiliated with non-Croatian teams, the general trend seemed to be that the focus was on players of Croatian origin. Interestingly, there were rather few Croatian popular figures compared to global ones. Nationalities and ethnic groups included primarily Croatians, as expected, but also both Balkan and non-

Location type	Count	Examples
Other countries	22	China, Russia, Bosnia & Herzegovina, Italy, USA, Egypt, France, Germany, Poland, Slovenia
Cities, Croatia	12	Zagreb, Split, Rijeka, Osijek
Croatia	7	Croatia, Republic of Croatia
Cities, global	5	London, Belgrade, Madrid, Moscow
Continents	2	Europe
County	2	Istria, Međimurja
Geographic regions	1	Crimea
Stadiums	1	Poljud
Highway	1	A1
Parts of Zagreb	0	*Cvjetni trg* ('Flower Square')
Geography of Croatia	0	Plitvice Lakes, Zadar Peninsula
Other	0	Buckingham Palace, Camp David, Memorial Cemetary for War Victims

Table 3.8: Types of locations with number of occurrences of this type in a sample of the top 50 most common entities. Please note that entities are counted for each variation, i.e. five different declensions of Croatia will count as five.

Person type	Count	Examples
Political figure, local	21	Plenković, Milanović, Bandić, Tomašević, Marić, Beroš, Capak, Divjak, Jandroković, Josipa Rimac, Jovanović, Kosor, Kotromanović, Kovač, Kuščević, Medved, Orešković
Sports figure	7	Čilić, Ibrahimović, Kopić, Leo Messi, Mamić, Matić
Political figure, global	6	Putin, Vučić, Biden, Juncker, Macron
Popular figure, global	5	Adele, Bill Gates, Bradley Cooper
Ethnic group or nationality	5	Americans, Brazilians, Chinese, Macedonians, Germans
Popular figure, local	3	Lejla Filipović, Lidija Bačić, Nives Celzijus
Croat (ethnic group)	3	-
Regional identities	0	Dalmatians
Other	0	God

Table 3.9: Types of people with number of occurrences of this type in a sample of the top 50 most common entities. Please note that entities are counted for each variation, i.e. five different declensions of Plenković will count as five.

Organisation type	Count	Example
Sports	23	Hajduk, Dinamo, Inter, NBA, Barcelona, Formula 1
Political, local	10	HDZ, the (Croatian) government, SDP, Sabor, USKOK, Croatian Agency for Supervision of Financial Services
Political, global	6	EU, Al-Qaeda, United States Army, American Civil Liberties Union, Australian government
Corporations, local	5	Agrokora, HEP, Adris Group, Amfora
Corporations, global	4	Volkswagen, Airbus, Ara Shoes, Facebook
Musical acts, global	1	Foo Fighters

Table 3.10: Types of organisations with number of occurrences of this type in a sample of the top 50 most common entities. Please note that entities are counted for each variation, i.e. five different declensions of Hajduk will count as five.

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

Balkan ethnic groups.

Finally, organisations also show a high focus on Croatian sports clubs, government bodies, and political parties. It was again difficult to separate sports into local and global groups. See Table 3.10 for more details. We did not repeat this process for MISC due to sparsity of data. The most common MISC entity was 'Christmas' (holiday, mostly POS) but also included a number of both Croatian programmes and holidays.

A peculiarity of this data set is the fact that entities are not lemmatised. Although this generally has a minimal impact if any for analytical languages such as English due to the lack of inflection, this has a number of consequences for a highly-inflected language like Croatian. The first consequence is that it creates sparsity. For example, not only is Croatia spread out across different terms, such as 'Croatia', 'Republic of Croatia' and 'HR', but each term itself may be spread out across different declensions, resulting in the term appearing to be much less frequent. In Table 3.11, 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.

	Case	Number	Gender	NEG	NTR	POS	Total
Hrvat	NOM	SG	M	0	0	2	2
Hrvati	NOM	PL	M/Mix	1	1	2	4
Hrvatim*	DAT/LOC/INSTR	PL	UNK	0	0	1	1
Hrvatima	DAT/LOC/INSTR	PL	UNK	0	1	0	1
Hrvata	GEN	UNK	UNK	0	0	1	1
Hrvatica	NOM	SG	F	0	0	1	1
Total				1	2	7	10

Table 3.12: Distribution of the declension of *Hrvat* ('Croat') across different sentiment labels. An asterisk (*) denotes a suspected typographic error. We merge forms that are irrecoverable without context.

This leads us to the second consequence: namely that the grammatical function of the entity is partially recoverable by looking at the ending. In the case of *Hrvatska*, there is a tendency towards positive in nominative case, indicating the agent of an active verb or patient or theme of a pasisve 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, see Table 3.12.

3.3 Summary

In this chapter we presented the two data sets we have been provided for our research. The first data set is the TakeLab Retriever headlines data set, which we compared with other Croatian domains. We described our pre-processing process, which involved several rounds of de-duplication followed by concatenation. The second data set was the SToNe data set, a sub-set of the headlines set with targeted sentiment annotations. We described the annotation process of this data set as well as pre-processing steps taken. Finally, we performed exploratory analysis on the training set of this data set.

Chapter 4

Methodology

In this chapter, we will describe the methodology followed in order to carry out our experiment. We will first describe the models used, reasoning behind their selection, as well as any notable attributes of these models. Then, we will describe the two-stage training procedure, followed by a description of evaluation and metrics used for evaluation.

4.1 Models

This section describes the models used to perform the experiment. In total, we selected five separate models to examine how additional self-supervised pre-training with unlabelled task data (*task-adaptive pre-training* or TAPT, as described in Chapter 2.1) 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. Considering that there are very few Croatian-supporting models in existence, these five models together encompass a near totality of Croatian language modeling. The only notable model excluded from this roster is DistilBERT-Multilingual-Cased (Sanh et al., 2019). The decision to exclude this model is motivated by the fact that the model is designed around the goal of reducing size with significant compromises to performance and flexibility. These compromises result in performance worse than the typical baseline of Multilingual BERT (Ptiček, 2021). Additionally, DistilBERT is generally sidelined from the bulk of Croatian language modelling research, further rendering its inclusion unnecessary.

Models in our experiment can be divided into two main groups: *Croatian-dominant* and *massively multilingual*. Croatian-dominant models are models which are predominantly trained on Croatian or any of the closely related languages of Bosnian, Montenegrin or Serbian, but may also include models which are trained on other languages as well. Massively multilingual models, however, are models which are trained on 100 or more languages and may include Croatian in however limited quality in their training corpora. We provide an overview of the configuration of each model in Table 4.1 and a comparison of exposure to Croatian training data for each model in Table 4.2.

To examine how much the vocabulary of each model differed from the vocabulary in the TakeLab Retriever headlines data set, we ran each model's to-

Model	Base	Parameters	Layers	Vocab	Objective
BERTić	Electra	110 million	12	32,000	RTD
cseBERT	BERT	110 million	12	49,601	WWM
mBERT	BERT	110 million	12	119,547	MLM
XLM-RoBERTa-Base	BERT (RoBERTa)	125 million	12	250,002	MLM
XLM-RoBERTa-Large	BERT (RoBERTa)	334 million	24	250,002	MLM

Table 4.1: A comparison of all five models in use. RTD = Replaced token detection, WWM = Whole word masking, MLM = Masked language modelling.

Model	# of languages	Training breakdown
BERTić	1-4	Croatian (66.3%), Serbian (23.33%), Bosnian (9.42%), Montenegrin (0.95%)
cseBERT	3	English (47%), Croatian (31%), Slovene (23%)
mBERT	104	Includes Croatian, Bosnian, Serbian and Serbo-Croatian
XLM-RoBERTa	100	Includes Croatian (5.7G), Bosnian (18M) and Serbian (1.5G)

Table 4.2: The models compared based on the amount of training data in Croatian and related languages, if provided. XLM-RoBERTa models were merged due to being identical. Exact figures for Multilingual BERT and XLM-RoBERTa were not provided.

Model	Vocab used	Percentage of vocab
BERTić	30,187	94.33%
cseBERT	33,887	68.32%
mBERT	36,522	30.55%
XLM-RoBERTa	39,664	15.87%

Table 4.3: A comparison of each model in relation to the TakeLab Retriever headlines data set. *Vocab used* indicates the number of unique sub-tokens identified, while *percentage of vocab* indicates how much the vocabulary overlaps with the model's own subword token vocabulary.

Model	Subtokens (millions)	# of UNK tokens	Percentage UNK
BERTić	142.72	459	<.001%
cseBERT	146.64	798,487	0.54%
mBERT	187.55	774,053	0.41%
XLM-RoBERTa	160.86	117	<.0001%

Table 4.4: Length of the data set in subtokens, as per each model's tokeniser, followed by the number of out-of-vocabulary (UNK) tokens.

keniser across the data set. We compared the final vocabulary size with that of the model's actual vocabulary size. We also counted the number of subtokens in total and compared the number of out-of-vocabulary (UNK) subtokens. See Table 4.3 for a breakdown of these statistics.

4.1.1 Croatian-dominant

This sub-section describes the two different Croatian-dominant models selected for evaluation. We analyse their training corpora and compare them to the TakeLab Retriever headlines corpus we use for pre-training. We also describe any peculiar attributes they may have and provide justifications for their use.

BERTić

The first model, BERTić is a model trained exclusively on corpora derived from Bosnian, Croatian, Montenegrin and Serbian sources (Ljubešić and Lauc, 2021). Taking into consideration the large degree of overlap that these corpora have in terms of vocabulary and structure, BERTić is as close as possible to being a monolingual model for Croatian. We illustrate the similarity of some of the different TLD corpora in Table 4.5, which shows that Croatian and its closely related neighbouring languages share more vocabulary than the hrWaC corpus does with other Croatian domains (see Table 3.2.

Linguistic overlap notwithstanding, a large proportion of BERTić's training corpus comes from Croatian sources; BERTić's training corpus altogether consists of 8.39 billion tokens, of which approximately 5.56 billion are Croatian. The headline data set highly overlaps BERTić's vocabulary (94.33%) with only 459 out-of-vocabulary tokens (<.001% of tokens). We take this to

	HR	BS	ME	SR
HR	**100.0**	80.08	59.29	67.73
BS	80.08	**100.0**	68.48	75.23
ME	59.29	68.48	**100.0**	66.90
SR	67.73	75.23	66.90	**100.0**

Table 4.5: Comparison of word frequencies (top 10,000) for different web corpora for Bosnian, Croatian, Montenegrin and Serbian, sourced from corresponding top-level-domains, {bs,hr,me,sr}WaC (Ljubešić and Klubička, 2014). HR = Croatian, BS = Bosnian, ME = Montenegrin, SR = Serbian.

be indicative of the degree to which BERTić is specified for Croatian language modelling.

A notable quirk of BERTić, however, is that it is trained with the ELECTRA objective, or 'Efficiently Learning an Encoder that Classifies Token Replacements Accurately' (Clark et al., 2020). Rather than performing masked language modelling, as done with traditional BERT models (Devlin et al., 2018), ELECTRA models are trained on a *replaced token detection* (RTD) task. RTD involves concurrently training a smaller generator and larger discriminator. The generator is used to generate a number of plausible replacements for a random token, which are then used to replace the token in unlabelled text. The discriminator is then tasked to evaluate the likelihood for each word in a string and identify the replacement. Upon completion of training, the generator is discarded and the discriminator is used for downstream tasks. Clark et al. (2020) claim such an approach to be effective with regards to sample size in comparison to masked language modelling due to the fact that loss is calculated across an entire span of text rather than just for the masked token. BERT-level results are achieved with significantly less training and possibly surpassed if fully trained. The sample-effectiveness of this approach is of particular benefit to low-resource languages such as Croatian, making full use of the available unlabelled corpora. Crucially, this approach 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.

We include BERTić in our roster of models not only because it boasts state-of-the-art performance in all tasks related to Croatian, but also because it serves as the main focus of previous work done by Barić et al. (2023) on entity-level sentiment analysis in Croatian headlines. Its inclusion is thus an invaluable part of this research.

CroSloEngual BERT

CroSloEngual BERT (pronounced 'Crosslingual BERT') or cseBERT is a trilingual model trained on English, Croatian and Slovene (Ulcar and Robnik-Sikonja, 2020), designed to explore the phenomenon of cross-lingual knowledge transfer between a limited number of related languages as opposed to monolingual models or massively multilingual models. Notably, in compar-

ison to BERTić, cseBERT excludes pre-training from Bosnian, Montenegrin or Serbian corpora, instead relying on transfer learning from less-related languages. The intuition behind the author's approach is that feature-rich English data can translate to better performance in Croatian, with Slovene data also providing transfer learning.

cseBERT is trained on a corpus of 5.9 billion tokens, predominantly composed of English, with a smaller portion in Croatian. Slovene makes up the smallest portion. This means that cseBERT's Croatian pre-training corpus is not only smaller than that of BERTić, but that cseBERT has been pre-trained on the least amount of data altogether out of all models examined here. Ulčar and Robnik-Šikonja (2020) report that the training data set specifically uses a combination of the publicly available Croatian Riznica literary corpus (Brozović Rončević et al., 2018) and hrWaC web corpus (Ljubešić and Klubička, 2014), both used by BERTić, as well as an unreleased private corpora. The Slovene portion derives from the Slovene Gigafida 2.0 corpus (Krek et al., 2020). The authors did not report on the English corpus used for pre-training. The headline data set moderately overlaps cseBERT's vocabulary (68.32%) despite its training set only containing 31% Croatian, possibly indicating considerable transfer from English and Slovene in particular. Perhaps due to its considerably small training corpus, cseBERT also encounters the most out-of-vocabulary tokens out of all the models in the headlines data set: 798,487 tokens, or 0.54% of the tokens.

Unlike BERTić but also unlike other BERT models, cseBERT is trained on the *whole word masking* (WWM) task, also known as the Cloze task or procedure (Taylor, 1953). WWM differs from the standard BERT training procedure by masking entire words, requiring the target model to recover the whole word rather than just WordPiece sub-word tokens (Schuster and Nakajima, 2012). In some cases, the model may have to recover several sub-word tokens in a row, depending on how the masked word is tokenised by the tokeniser. This task is significantly more challenging as it demands that the model rely on surrounding context rather than surrounding sub-tokens. Such an approach has been shown to result in a more robust model in some cases (Cui et al., 2021).

We justify the inclusion of cseBERT for two main reasons. The first reason is that, prior to the introduction of BERTić, cseBERT was the state-of-the-art model for Croatian and remains competitive if not top in some tasks (Ljubešić and Lauc, 2021). It thus continues to be relevant for Croatian-related comparisons. The second reason is that its training objective is a lot more similar to that of the other BERT models compared to BERTić. Essentially, cseBERT is trained through a generative rather than discriminative task. This reduces the ambiguity of whether observed differences can be attributed to the pre-training task alone or to the language specificity of the model. By having a language-specific (albeit not monolingual) model that is also trained in a manner similar to the massively multilingual BERT models, the inclusion of cseBERT strikes us as sensible, especially in lack of resources, in terms of data, time, and computing power, to train new monolingual models using masked language modelling from scratch.

4.1.2 Massively multilingual

This subsection describes the three additional massively multilingual models included in the comparison. We give an overview of the three models under consideration, delve a bit into the composition of their training corpus, and examine their pre-training procedure.

These models were 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 introduced by Devlin et al. (2018). In this task, 15% of the sub-word tokens are randomly masked and must be recovered by the model. The token is either replaced with a special [MASK] token (80% of the time), a random token (10%) or the original unchanged token (10%). Loss is calculated for the prediction of the replaced token. This task is on occasion also erroneously referred to as the Cloze task (Taylor, 1953), although that task involves masking whole words rather than sub-words.

Massively multilingual models have been the subject of a considerable amount of research in BERT-focused NLP sub-fields, especially with low-resources languages. Although usually only one such model is included as a baseline, we decided to include and compare three for the sake of completion.

4.1.3 Multilingual BERT

Multilingual BERT (mBERT) is the original massively-multilingual BERT model introduced in the original landmark paper by Devlin et al. (2018). For our research, 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. This corpus includes the Croatian, Bosnian, Serbian and Serbo-Croatian editions.

Although we were not able to determine the precise proportions of the pre-training corpora that contains Croatian, Bosnian, Serbian and Serbo-Croatian text, we nevertheless determined that none of these language editions were within the top 10 largest Wikipedia corpora (Wu and Dredze, 2020). Furthermore, we found the inter-lingual corpora vocabulary overlap to be quite high (Table 4.6, which may indicate that portions of the corpora were redundant. On the other hand, the overlap between the Croatian Wiki corpus and the headlines corpus appeared to be the lowest out of all Croatian corpora (Table 3.2). When compared with the vocabulary of the headlines data set, mBERT also showed a considerable amount of unfamiliarity (0.41% out-of-vocabulary) despite our expectations that its diverse multilingual vocabulary would likely be able to capture non-Croatian noise text. These facts all suggested that mBERT may have the least exposure to the target domain.

We justify the inclusion of mBERT on the following grounds. First, mBERT has become a baseline for comparison for non-English tasks, especially for low-resource languages that lack several monolingual models or lack data necessary to produce one. In fact, mBERT is used by both Ljubešić and Lauc (2021) and Ulcar and Robnik-Sikonja (2020) in their evaluations of Croatian transformer models. Second, mBERT has been trained on, by far, the least

	HR	BS	SR	SH	SL
HR	**100.0**	73.03	62.35	73.65	28.29
BS	73.03	**100.0**	67.68	72.89	27.51
SR	62.35	67.68	**100.0**	72.14	26.00
SH	73.65	72.89	72.14	**100.0**	28.85
SL	28.29	27.51	26.00	28.85	**100.0**

Table 4.6: Vocabulary overlap (top 10,000 words) for CLASSLAwiki corpora consisting of various Wikipedia language editions. Slovene is included to emphasise its comparative dissimilarity to the other languages. HR = Croatian, BS = Bosnian, SR = Serbian, SH = Serbo-Croatian, SL = Slovene.

amount of in-domain Croatian out of all models. Its performance can thus reveal how TAPT affects models that are further away from the task or domain.

4.1.4 XLM-RoBERTa

Two variants of XLM-RoBERTa were also included: a smaller Base model and a larger Large model. The two models differ only in size, such as the number of parameters and layers (4.1). Other details, such as training corpora size and vocabulary size, are identical. Due to their similarities, we will be describing them together in this heading.

XLM-RoBERTa consists of two crucial parts. First, RoBERTa (Robustly Optimised BERT Pretraining Approach) was introduced by Liu et al. (2019) in response to suspected shortcomings in the per-training approach to BERT. Such suspicions hinted that BERT was not properly optimised and may have been under-trained. The authors performed a new set of hyper-parameter tuning and continued training with a larger amount of data. Their findings showed significant improvements across many benchmarks including GLUE (Wang et al., 2018). Second, the XLM approach to multilingual training was described by Conneau et al. (2019) as a way to work with supervised and unsupervised data across several languages. The authors applied this approach to a RoBERTa base to produce XLM-RoBERTa, which produced state-of-the-art results in multilingual tasks.

XLM-RoBERTa differs from mBERT in a few notable ways. First, XLM-RoBERTa makes use of Byte-Pair Encoder (BPE) tokenisation, introduced by (Sennrich et al., 2015), 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. However, more striking is the sheer amount of data to which XLM-RoBERTa is exposed in pretraining. 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 (Wenzek et al., 2019).

Although we were unable to perform a training set overlap comparison due to the lack of time and resources to perform a word frequency count on the cc100-hr data set, we found that XLM-RoBERTa had the lowest number of out-of-vocabulary tokens while also using the lowest amount of its total vocabulary when tokenising the headlines data set. This may suggest that,

while using only a little portion of its very large vocabulary, the amount used is still enough to represent almost every subtoken from the data set, including non-Croatian noise data.

We felt the inclusion of both XLM-RoBERTa versions to be suitable for a number of reasons. First, XLM-RoBERTa is often used in comparisons with monolingual models. Its performance can reveal how differences in training architecture affect gains through language adaptation. Second, RoBERTa is also the base model of the work by Gururangan et al. (2020) on domain- and task-adaptive pre-training. Its inclusion thus serves to replicate how their findings may adapt to other languages and domains.

Finally, the inclusion of the second XLM-RoBERTa model, XLM-RoBERTa-Large, can show how much model size alone can impact performance and rates of improvement. The decision was also inspired by Ljubešić (personal communication, 2023), who had informed us that his research on BCMS now tended to include large multilingual models.

4.2 Training

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, see Chapter 2.1 for an in-depth explanation), which adapted the models to the general unlabelled data of the task. For each model, we also produced a version for comparison which omitted this stage. The second stage consisted of fine-tuning. In this stage, each model was trained on the labelled SToNe data set with loss calculated on the validation portion of the set. We go more in-depth about these two stages below.

4.2.1 Task-adaptive pre-training

The first stage of training was TAPT in which the models were trained with their original pre-training objective using the Takelab Retriever headline data set. All models retrained the original parameters as specified by the respective papers. The data set was concatenated and then divided into equal-sized chunks for every model except BERTić. In the case of BERTić, we encountered many difficulties related to the ELECTRA implementation overall. In the end, on the advice of some associates of Ljubešić and Lauc (2021), we used the `SimpleTransformers` library to automate training, which did not allow concatenation.

In consideration of time and resources at our disposal, as well as the sheer amount of differences between all models, we did not perform any hyperparameter tuning. We only trained the models for three epochs rather than the 100 epochs used by Gururangan et al. (2020), even if results indicated that the models appeared to be under-trained.

4.2.2 Fine-tuning

In the second stage of training, we performed fine-tuning using the annotated SToNE data set with the goal of targeted sentiment analysis. We closely followed the process of the *Target* baseline from Barić et al. (2023) and fed all models the headline and target NE embeddings span only. Loss was calculated based on the validation portion of the data set. We left other approaches from the paper, such as *Target+NE Type* and multi-batch combinations with tone, to future research.

Each model was tested after 10 epochs of fine-tuning. We experimented with as few as 3 and as many as 50 epochs but found the 10-epoch fine-tuning process used by Barić et al. (2023) to strike an optimal balance.

4.3 Task

The final models were tasked on their ability to predict the sentiment labels selected by the annotators. Models were given a headline and NE target and expected to predict one of three labels sentiment labels (NEG, NTR or POS) for the target of each headline. Targets could have either implicit or explicit sentiment and may also have multiple named entities with different or conflicting sentiments. The models were expected to predict the label for *only* the NE target.

4.4 Evaluation metrics

In this section, we will discuss metrics employed to perform evaluation on the models after pre-training and after fine-tuning, each using a different metric. The first stage, consisting of TAFT, employs *perplexity*, whereas the final evaluation uses an average of F_1*-scores* across five seeds. However, for the purposes of analysis, *precision* and *recall* across all labels will also be considered.

4.4.1 Perplexity

To evaluate the results of TAPT, we used perplexity. Perplexity is a well-established metric for language model evaluation which measures the confidence with which a language model is able to predict the outcome of a task. Higher perplexity values indicate that the model is more 'surprised' and thus lacks confidence or is unfamiliar with the given material, while lower values indicate more confidence. A decrease in perplexity after training therefore indicates that a model has 'learned' from the process. In the case of Transformer models, perplexity is calculated by taking the exponential of the cross-entropy loss from the validation set.

4.4.2 Precision, recall, and F_1-score

Our final evaluation and analysis will make use of *precision*, *recall*, and F_1*-score*, metrics which derive from information retrieval and continue to see

standard use in NLP. The overall evaluation will employ F_1-score, including a comparison of gains or losses from TAPT training, while the error analysis will make use of precision and recall metrics to examine areas of improvement in the highest-performing model. We provide an overview of these metrics in the context of our research below.

1. **Precision** is defined as the ratio between the true positives (*TP*) and the sum of true positives (*TP*) and false positives (*FP*). Essentially, it measures how successfully it has applied a prediction class, i.e. POS, when cases are indeed that class. A model with high precision for POS will accurately predict POS labels while minimally predicting this particular label in cases where the gold label is NTR or NEG.

2. **Recall** is defined as the ratio between the true positives (*TP*) and the sum of true positives (*TP*) and false negatives (*FN*). This measures the amount which a prediction class, i.e. POS, has been applied correctly compared to the total count of this class. For example, high recall will mean a model has predicted POS in labels where POS is indeed appropriate, while not missing other instances of POS.

3. ***F*$_1$-score** is defined as the harmonic mean between precision and recall. It balances both precision and recall into one metric.

When considering overall performance, we will use macro averages rather than weighted averages. We believe that macro averages are appropriate in our evaluation due to the equal importance of our classifier in recognising every type of sentiment.

4.5 Summary

In this chapter, we have described all the models in use, performed a comparison between their Croatian training data and the TakeLab Retriever headlines data set, and indicated reasons behind their inclusion. We split the models into two main groups: Croatian-dominant and massively multilingual. Then we described the two-stage training procedure and the targeted sentiment analysis task. Finally, we discuss our main evaluation metrics, perplexity and F_1-score.

Chapter 5

Results and analysis

This chapter provides the final results of our comparison. We first reveal the perplexity decreases for each model and provide a quick summary of the results. Then, we introduce a comparison between the different models used. We compare the models on F_1-score as well as on percent change with task-adaptive pre-training (TAPT). Finally, we perform an error analysis of the highest-performing seed of the highest-performing model on the test set.

5.1 Pre-training results

All models showed a drop in perplexity after the task-adaptive pre-training stage. This indicated that all models indeed 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. The high value indicates 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 5.1 for all pre-training results.

5.2 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

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 5.1: Perplexity across pre-training

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-RoBERTa-Base	0.669	0.633	0.726	0.648
+ TAPT	0.728	0.702	0.763	0.719
XLM-RoBERTa-Large	0.728	0.723	0.749	0.713
+ TAPT	**0.771**	**0.770**	**0.793**	**0.750**

Table 5.2: Comparison of F_1-scores for all models with and without task-adaptive pre-training (TAPT). Highest performing results are indicated in **bold**.

Model	AVG	NEG	NTR	POS
BERTić	-1.208%	1.664%	-0.519%	-4.839%
cseBERT	-0.975%	-1.695%	1.759%	-2.691%
mBERT	**10.000%**	**15.273%**	4.360%	**11.943%**
XLM-RoBERTa-Base	8.819%	10.900%	5.096%	10.957%
XLM-RoBERTa-Large	5.907%	6.501%	**5.874%**	5.189%

Table 5.3: 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. The largest increase for each label is indicated in **bold**.

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 5.2.

The sheer difficulty of the task is evidenced here through the performance of the models. Although models exhibited decent performance, none of the models manage to exceed an F_1-score of 0.8 consistently across all seeds. This was the case not just for the average across all labels, but also for every label individually.

An entirely different picture is painted when examining the results through the gains (Table 5.3). All Croatian-dominant models experience decreases in performance with TAPT, with BERTić decreasing 1.208% in F_1-score after the added pre-training. cseBERT experiences a similar but slightly smaller decrement, 0.975%. On the other hand, all massively multilingual models experience performance boosts with TAPT.

5.3 Error analysis

In this section, 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

	Precision	Recall	F_1-score	Support
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 5.4: Final results for each label in terms of precision, recall and F_1-score for XLM-RoBERTa-Large.

room for improvement for XLM-LT. We provide an overview of final scores, Table 5.4, and a confusion matrix of the results, Table 5.5.

Considering that errors are not all equal, 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.

The logic behind this subdivision is that opposite errors are significantly rarer than neutralising or polarising errors. Although it is important for a classifier to be able to accurately identify any label correct, opposite errors are significantly more severe and a higher frequency may indicate a problem that needs to be resolved.

Immediately, it can be seen 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 evidenced by the lower precision for both NEG and POS compared to NTR. XLM-LT is shown to over-predict NEG labels in particular, as seen in Table 5.5. Meanwhile, POS gets equal values for precision and recall (and thus for F_1-score as well). We presume that this is simply a mathematical coincidence, as the breakdown of the false positives and false negatives for POS differ, with a slight tendency to label POS as NEG rather than the other way around. That said, XLM-LT very rarely produces opposite errors, i.e. it rarely predicts POS or NEG when NEG or POS is expected, respectively. Such errors only make up 18.37% of errors, or 3.9% of the predictions overall. In both cases, it is almost much more likely to predict a NTR tag as NEG (9.95% of NTR gold labels) or as POS (10.41%).

Overall, precision is a bit weaker than recall, with most of the weakness coming from neutralisation errors, predicting neutral instead of a polar sentiments.

5.3.1 Results by named entity types

We now examine how XLM-LT's predictions hold up when broken down to different named entity (NE) types. See Table 5.6 for a classification chart

	NEG	NTR	POS	Support
NEG	**92**	16	7	115
NTR	22	**176**	23	221
POS	11	19	**96**	126

Table 5.5: Confusion matrix for the labelling performance of XLM-RoBERTa-Large + TAPT. The rows indicate gold labels whereas the columns indicate predictions.

		Precision	Recall	F_1-score	Support
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 5.6: Results for XLM-RoBERTa-Large + TAPT distributed across named entity types. Areas in which performance is below 0.750 have been noted in **bold**.

filtered by NE type, which we will be referencing throughout our analysis. We provide a breakdown of error types by NE type as well, Table 5.7, which we justify by the decision of Barić et al. (2023) to include NE type during their trials. We also propose some suspected causes of errors.

PER

We observe that PER is a particularly weak NE type for our classifier, having the lowest F_1-score of all (0.755) as well as lowest precision (0.753). The weaknesses in particular appear to be the result of a tendency to under-predict NTR labels, producing polarisation errors by assigning NEG or POS. NEG is considerably over-predicted, resulting in the lowest precision out of all type+sentiment combinations (0.682), although its rather high recall (0.833) indicates that it manages to label PER+NEG successfully quite often. Our previous observation of PER+POS being over-represented in the training set is a likely culprit of this decreased performance in PER overall.

While only 18.37% of errors overall are opposite errors (POS is predicted for NEG or vice versa), 12 of such errors (66.67% of polar errors, or 12.24% of

Type	Target	Prediction	PER	ORG	LOC	MISC	Total
Opposite	NEG	POS	5	1	0	1	7
	POS	NEG	7	3	1	0	11
Neutralising	NEG	NTR	4	6	5	1	16
	POS	NTR	7	5	6	1	19
Polarising	NTR	NEG	14	4	3	1	22
	NTR	POS	15	4	1	3	23
Total			52	23	16	7	95

Table 5.7: A breakdown of error type by NE type.

	Abbreviations	Total MISC	Percentage
Training	90	694	12.97%
Test	34	114	29.82%
Errors	9	20	45.00%

Table 5.8: A tally of abbreviation composition of MISC in the training, test, and error set.

errors overall) are associated with the PER label. Considering that PER makes up 45.89% of the NE types in the test set, this indicates that 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. In the following example, which is labelled NEG but predicted POS by our model, it appears that the author is judging the wardrobe choice of Ava and then extracting a quote from an interview with her for irony:

> **Ava** došla polugolih grudi: Ja sam natprosječno inteligentna
>
> **Ava** came with partially nude breasts: I am of above-average intelligence.

Our classifier is likely failing to detect the sarcasm present, which is being produced through contradiction. Sarcasm is a crucial part of building a sentiment analysis classifier (Onan and Toçoğlu, 2021) and would need to be examined in the future, especially for PER, where irony is frequently deployed.

ORG

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, evidenced by the lower recall for ORG+NTR (0.780).

Abbreviations appear to cause particular difficulty for ORG. We observe that 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, see Table 5.8. We believe that, while abbreviations overall cause grammatical anomalies by introducing a hyphen to indicate declension, some of these issues may be caused by issues with tokenisation and

NE	Approach	Headline
HDZ	Stemmed	Zaoštreni odnosi u stranci: Plenković će se riješiti pobunjenika u **HDZ**-u ne podrže li u Saboru odluke vodstva
		Strained relations in the party: Plenković will get rid of rebels in the **HDZ** if they do not support the decisions of the leadership in Parliament
HDZ-	Stemmed with hyphen	Škorini uvjeti za vlast s **HDZ-**om: kultura, poljoprivreda i MUP
		Škora's conditions for power with the **HDZ**: culture, agriculture and MUP
HDZ-a	Complete (with hyphen and case ending)	REZULTAT ĆE VAS IZNENADITI Usporedili smo politike **HDZ-a** i SDP-a
		THE RESULT WILL SURPRISE YOU We compared the policies of **HDZ** and SDP

Table 5.9: Demonstration of inconsistent approaches to selecting the span of the same entity, HDZ, an ORG named entity. All three headlines included are from the error subset of the test set. The expected behaviour is the last form, with the full case ending in tact.

span labelling in both the test and training set which surface much earlier upstream, during pre-processing and annotation of the SToNe data set before it was presented to us. See Table 5.9 for some examples taken from the ORG error set.

LOC

The classifier's performance in LOC, like ORG, is almost opposite that of PER. LOC+NTR is without a doubt 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 but POS in particular. It is clear that the classifier makes a lot of neutralisation errors, tending to predict a NTR sentiment with locations. This performance reflects the biases of LOC in the training set very well, further suggesting that lack of representation in LOC+NEG and LOC+POS may be responsible.

In Table 5.10, 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, a look into case for LOC specifically is possible because there is enough case variation in our data set and cases are easily discernible 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

		Precision	Recall	F_1-score	Support
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 5.10: Results for XLM-RoBERTa-Large + TAPT, LOC named entity types, divided further by case. Areas in which performance is below 0.750 have been indicated in bold. We omit INSTR because it only occurs once and no errors were made.

such as semantic roles to parse the sentiment of input.

Immediately, 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**

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. Considering the fact that precision and recall (and F_1-score) are the same for each of these labels, we are not able to make any particular conclusions about biases that the classifier has for nominative case. 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.

MISC

Finally, performance in MISC can be improved (F_1= 0.765) but it is likely hindered by the relative rarity of this NE type. MISC+NEG has rather low recall (0.667), likely caused by its rarity in the training set, appearing only 17 times (1.05% of the training set). Although the classifier has excellent recall for MISC+POS, its low precision combined with low recall for MISC+NEG and MISC+NTR indicate that it is over-predicting POS for MISC NEs.

Considering that the NE type is already uncommon and errors are few, there is little to be said of statistical significance about the errors. However, we will bring attention to one error in particular in the following headline:

> Novo liječenje **Covida-19** dramatično smanjuje broj bolesnika na intenzivnoj
>
> New treatment for **COVID-19** dramatically reduces the number of patients in intensive care

The classifier makes an opposite error here, predicting POS while the label is NEG. COVID-19 appears three times in the training set, each with NEG sentiment. We suspect that this may indicate that the classifier is failing to understand different aspects of the entity 'treatment for COVID-19'. Essentially, it is failing to perform a kind of aspect-based sentiment analysis. The logic is that COVID-19 should be labelled NEG because it is a virus that has caused a very serious global pandemic in 2020, whereas a treatment for it should be POS. However, XLM-LT may be failing to understand that the goal is to find the sentiment towards the virus itself rather than the treatment.

A particular challenge with this type of error, as well as the error described above in Chapter 5.3.1, is that the sentiment in question is much more implicit. This once again indicates that the model may have issues with implicit sentiments.

5.3.2 Conclusion of error analysis

This concludes our error analysis. We have identified a number of potential strengths and weaknesses as well as underlying causes of issues in XLM-RT. Our overall proposal is either to expand the training data overall or curate it to contain more examples of areas of weakness. Some issues, such as span errors in abbreviation labelling, are specific to our task and may require directly editing the training data. Other issues, such as irony and inferred sentiment, are much deeper challenges common to the task of sentiment analysis at large which will require further research or new approaches.

5.4 Summary

In this chapter, we examined the final results of our two-stage training process. We compared the perplexity decreases for each model before and after training. We also compared how much each model improved with the addition of task-adaptive pre-training (TAPT), noting that there appeared to be a correlation between how overlap there was between the original pre-training and

the TAPT data. Performance decreased with Croatian-focused models, while multi-lingual models dramatically improved.

Subsequently, we performed an in-depth error analysis of one run from XLM-RoBERTa-Large with TAPT (XLM-LT), our highest-performing model. We found that XLM-LT performed rather well with NTR, but not so well with NEG and POS. Overall, recall was better than precision. We also analysed the performance of specific NE types and qualities of each NE type. We proposed the following potential issues:

1. There may be issues with parsing irony.
2. Abbreviations cause issues and some of the annotation data seem to contain span errors, causing case-endings to be inconsistently marked.
3. Locations are overwhelmingly predicted to be NTR, but a look at different grammatical cases indicates that there may be some unintended shortcuts being learned. Particularly, the classifier seems to associate certain semantic roles with certain sentiments.
4. There may also be issues with separating different aspects of an entity. This means that XLM-LT might have some underlying problems with targeted sentiment analysis that also falls under the category of aspect-based sentiment analysis.
5. There are issues with implication, as more implicit sentiments seem to cause issues for the classifier.

We finally conclude that these issues may be addressed by acquiring more data and, in the case of abbreviations, making modifications where appropriate.

Chapter 6

Discussion

In this chapter, we expand on our findings, contextualising them in a bigger picture of task adaptation with transformer models as well as low resource languages, especially Croatian. We reflect on our findings in relation to our research interests. Then, we remark on limitations of our research and discuss potential future directions for research in the future, including different approaches to analysis, data handling, and training of different models. We conclude with a discussion on bias and ethical concerns.

6.1 General remarks

6.1.1 Domain and task adaptation

One of our main sources of inspiration was Gururangan et al. (2020), who introduce the concept of task-adaptive pre-training (TAPT). This kind of pre-training consists of continuing pre-training using unlabelled task-related data, which the authors found yielded improvements in RoBERTa across four domains and eight classification tasks. Gururangan et al. (2020) tested this 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. The 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. Further analysis can be performed by including DistilBERT, an even smaller model. If our suspicions hold up, then it should see larger gains than mBERT while performing still worse.

Lastly, we propose the existence of languages as 'super-domains', which apply in particular to low-resource 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.

6.1.2 Low resource language

Although we have touched on this point several times in passing, an underlying theme in working with Croatian is its status as a low resource language. This aspect of the language becomes very apparent throughout the process of our research, as we continuously encountered obstacles relating to lack of data, especially in ways we have not anticipated. This may either be directly related to the lack of data, such as encountering little means to extend the training for a particular model because the model has already seen all data for the language, or indirectly, such as models using alternative but less accessible architectures, such as ELECTRA or whole word masking, to make use of all available resources.

One of our takeaways from this work is that low resource languages require a special kind of attention that high resource languages do not. Work with low resource languages entails not only building and annotating data sets, but also finding ways to mitigate the lack of data. This means continuing to explore creative means of maximising sample efficiency but also weighing the costs and benefits of techniques such as transfer learning from another language. On top of that, low resource 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 necessarily become a 'must-do' but rather be included as part of a diverse toolbox of approaches domain and task adaptation if seen fit.

6.2 Limitations

6.2.1 Pre-training objectives

Although most BERT models use a masked language modelling (MLM) training objective which masks a certain proportion of sub-word tokens, not ev-

ery 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. These approaches to pre-training naturally have consequences on our approach.

First, 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 pre-training corpus contains the least amount of task-related data.

Time and resource constraints restricted our approach. 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.

6.2.2 Replicability

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. Even then, the data set would have to be the precise one retrieved from Retriever in order for all results to be the same. Unfortunately, this is beyond our control, although we suspect that there will not be significant changes.

A second limitation is one which relates more generally to 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 computers 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.

6.3 Future work

In this section, we outline a few different possibilities for future work. We roughly order our suggestions in the following manner:

1. **Data** We lay out ways directions to go with data, whether expanding or making use of our current data.
2. **Training** We discuss possible alternative approaches to training, domain and language adaptation in particular.
3. **Models** We look into ways specific models can improve as well as shortcomings to our approach that may have resulted in their decreased performance.
4. **Evaluation** We propose further means of evaluation, which may give a better impression of the performance of the models.

6.3.1 Exploring stylistic variation

Despite being short pieces of text, headlines tend to exhibit a large amount of stylistic variation which relate to the category of news they belong to. These variations can be influenced factors such as type of publication (newspaper or tabloid); topic (celebrity gossip, sports, politics, opinion); partisanship (political leanings towards the left, right or centre); and other factors. These variations introduce their own nuances in lexicon, leading to potentially different ways for targeted sentiment to be expressed. This is also the case for Croatian headlines. We see a potential for future research in domain-oriented headline analysis and see whether the model performs notably worse in some domains than others. We predict that including domain information, either classified by hand or automated through another transformer model, may have a positive effect on the performance of a model.

6.3.2 Mixed headlines

Discarding mixed data is a common practice in sentiment analysis, but one which results in the loss of what is already a low quantity of data (Kenyon-Dean et al., 2018). In our case, as noted in the data section, we replicated the process of Barić et al. (2023) by removing 548 headlines for having conflicting annotations. However, we acknowledge that the removal of 548 headlines essentially treats 19.19% of the 2,855 headlines from the SToNe data set as noise. Classifiers trained on such a 'de-noised' data set end up not learning about conflicting sentiments.

Barić et al. (2023) leave handling of these remaining 548 annotations up to future work. Although we did not end up using these mixed headlines, we will now discuss a few ways these headlines may be included. The most obvious method of use would be to simply include headlines that only conflict on tone but not sentiment. This will include only a trivial number of new examples, 45. Aside from this, the authors propose a few possible schemes. One scheme, which they refer to as 'adjudication' could include asking annotators

to make a final decision on the label for conflicting annotations. Another possibility would be 'fine-grained schemes', which would entail representing the annotations in more fine-grained manners, such as in a decimal scale rather than categorical labels.

An approach proposed by Kenyon-Dean et al. (2018) is to include a COMPLICATED label. Although such a label is difficult to use and often rare, it may be a direction worth exploring especially considering the often ambiguous nature of Croatian headlines.

6.3.3 Additional annotation and alternative approaches

It has been well-observed that the sheer amount of data that large language models see allow them to outperform human annotators in tasks such as masked language modelling, other tasks may prove to be especially difficult to annotate, leading to poor performance. Regardless of how the models themselves are tuned, all tasks ultimately boil down to the quality of the annotation and the difficulty of the task. Above, we discuss how to handle mixed sentiments. However, expanding the SToNe data set to be much larger will certainly benefit the models and allow there to be samples of some of the more challenging patterns.

In addition to expanding the data set, parts of the data set may be altered to include more information. Although Barić et al. (2023) experiment with removing the target embeddings and/or including NE type, there remain some other potential directions for representing the targets. Previously in Chapter 5.3, we identified potential issues with abbreviations (Table 5.9) as well as different grammatical cases (Table 5.10). We concluded that abbreviations may need to be modified to have consistent representations across the entire data set. Here we propose a few other ways for targets to be represented in the STonE training set.

Lemma

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, either in place of or in conjunction with the named entity as it appears in the SToNe data set. Such an addition could either be done by hand or a lemmatiser such as `reldi-tagger` (Ljubešić and Dobrovoljc, 2019), F_1= 98.17.

Representing the named entity in lemma form puts a focus on the named entity itself rather than without context. We believe that such an appropriate may positively impact implicit sentiments which rely on local or world knowledge.

Case or semantic label

We observed in Chapter 5.10 a potential statistical relationship between semantic role and sentiment. Although we focus mostly on locations, it is pos-

sible that other NE types also exhibit such statistical relationships. Our example in Table 3.12 with *Hrvat* alludes to such a relationship. Future work could incorporate this information in one of two ways. The first way is by passing case to the classifier, which is predominantly a grammatical function. The second way is by classing semantic label to the classifier, thus indicating the semantic role of the target. We predict that semantic label maybe be more informative than case, although the inclusion of case has not yet been researched extensively in sentiment analysis for highly inflective languages such as Croatian.

6.3.4 Domain adaptive pre-training

As discussed previously, Gururangan et al. (2020) show not only benefits from self-supervised pre-training with unlabelled data from a task (TAPT) followed by the usual fine-tuning procedure, but also by performing a preliminary stage of in-domain pre-training before all both of these training stages. This process, also referred to as *domain-adaptive pre-training* (DAPT), is one which we have entirely skipped in our study. There is, nevertheless, further potential for improvement for by incorporating more in-domain data.

In our case, this could be more general unlabelled data from the Croatian news domain. We have seen previously (see Table 3.2) that the headlines data set most closely matches the vocabulary of Croatian news data set. Aside from the publicly available ENGRI data set (Bogunović et al., 2021), TakeLab also may have all article text at their disposal. In either case, a large amount of unlabelled in-domain text exists which could serve as a stage of domain-adaptation. Following the work by Gururangan et al. (2020), there may be potential gains worth exploring using a larger in-domain data set.

Finally, building on our previous discussion of language super-domains, there also exists the potential to perform both LAPT and DAPT. Combining both LAPT and DAPT, in that order, 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. Although we suspect this may be redundant 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. This approach, thus, primarily applies to XLM-RoBERTa, Multilingual BERT and potentially cseBERT. 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.

6.3.5 Improving BERTić

Although BERTić remained a consistently strong performer, it was quite possible that the model did not reach its full potential. In this section, we will explore ways that BERTić's performance itself may be improved.

First, it is highly likely that BERTić has been under-trained and would benefit from significantly more epochs of pre-training. This is indicated by the

fact that its perplexity was exceptionally large before the TAPT stage (190,601.81) and still considerably large after (3383.42, compare to the max of 3.64 in XLM-RoBERTa-Base among multilingual models). Due to time and resource constraints, we were unable to test BERTić with different numbers of epochs. Future work should explore training BERTić on more epochs to see if this yields increases rather than decreases in performance. This is largely in the name of fairness, although we do not predict that BERTić will perform dramatically better.

It is worth noting, however, that there are a number of other considerations to make when working with BERTić. Even though the pre-training process of ELECTRA is sample-efficient and therefore suitable with limited corpora, it should be noted that its implementation can pose unpredictable challenges in comparison to more well-supported, traditional transformer models. The complexity of the training process makes it difficult to compare the model with other models, especially when a pre-training stage is involved. Pre-training ELECTRA either requires a time-consuming re-implementation process to account for the simultaneous training of the generator and discriminator heads, or the use of more restrictive libraries such as SimpleTransformers, which automate such a process but require specific types of data. We took the latter approach, which may have resulted in lowered performance. Further research could thus involve a proper implementation of the pre-training process with concatenation to match that of the rest of the models in our test.

Finally, it is possible that the WordPiece tokeniser itself can be a bottleneck for performance. Another approach to tokenisation can be done using ByT5 encoding (Xue et al., 2021), an approach recommended by one of the authors of BERTić (Lauc, personal communication, 2023). Such an approach would use token-less byte encoding, which may yield better performance and less susceptibility to noise.

Aside from such changes, due to scarcity of resources, short of massive production of new data, there is little to be done in terms of extending the pre-training corpus itself. BERTić has already been trained on an already large portion of the Croatian data available; we reiterate that it is likely that the TakeLab Retriever headlines data set or any other similarly scraped online data set from the .hr domain is redundant.

6.3.6 Improving cseBERT

Many of the above suggestions also apply to cseBERT. Like BERTić, cseBERT may also benefit from more epochs of training, given that its perplexity was quite high before training (279.35) and still above the multilingual models after (12.07), while still remaining well below that of BERTić. This may be explained by the fact that its whole word masking task is considerably more difficult than to the masked language modelling task. Additional epochs may allow cseBERT to reach levels of perplexity similar to that of XLM-RoBERTa and mBERT.

On the other hand, we do not expect XLM-RoBERTa-Base or mBERT to benefit from further pre-training, given their already quite low perplexity scores after training.

6.3.7 XLM-RoBERTa-XL and larger

We found that the pre-training and fine-tuning procedure already yielded such considerable gains that XLM-RoBERTa-Base leapt from poor performance to outperforming cseBERT entirely. Meanwhile, XLM-RoBERTa-Large without pre-training already outperformed every model but BERTić, and took the lead in the model line-up after pre-training. These dramatic gains in performance hint at the possibility of even larger models being better adept at handling the task.

Although we found that performance increases diminished as model size increased (see Table 5.3), it is quite likely that an even larger model will outperform BERTić without the need for TAPT, while notably increasing in performance with TAPT, if a bit less than smaller models. 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. The issue of diminishing returns, however, still holds. XLM-RoBERTa-Large already took the most amount of resources to train compared to other models in its cohort. A larger model may not perform better enough to justify the resources required to train it, especially if it may require more epochs to reach optimal performance. The limit on performance is ultimately affected by the quality of data available. Whether XLM-RoBERTa-XL is worth considering is entirely left for future research to decide.

6.3.8 Re-evaluating performance

In Chapter 5.3, we provided an analysis of errors produced by our highest performing model, XLM-RoBERTa-Large + TAPT. Although our review identifies patterns in errors, such as named entity types that cause issues or possible challenges in sentiment analysis, the results are not comprehensive enough to make deeper, statistically backed conclusions about the behaviour of the classifier. There exists the possibility that these issues are merely reflections of statistical anomalities in the data set we studied. Tuning the model too closely to our observations our may thus result in over-fitting, potentially causing a boost in performance on test data, but causing the model's performance to generalise poorly to new data.

One manner to study generalisability and robustness of the model is behavioural testing. Ribeiro et al. (2020) introduce CHECKLIST, a software suite which includes a matrix of linguistically-informed tests designed to generate large amounts of new data. The authors propose three types of tests:

1. *Minimal functionality tests*, in which the failure rate of a model is measured across a data set with a particular feature included, such as negation;

2. *Invariance tests*, in which pairs of sentences are compared with a feature added, with the expectation that no change will occur with the added stimuli; and

3. *Directional expectation tests*, which also compare pairs of sentences, but with the expectation that a particular kind of change will occur.

Although previous work has been done by Lee and Bloem (2023), who perform behavioural testing on Croatian-supporting model, BERTić, and Multilingual BERT on the closely related language Serbian, such a procedure has not been performed in the context of sentiment analysis in Croatian. Given the high degree of inflection, freedom of word order, and richness of morphology in Croatian, all of which are factors that typically pose challenges to Croatian language modelling, there is a lot of potential for variance bugs that must be checked.

Grammatical variance aside, it is also unknown to what extent the models are basing their predictions on knowledge associated with particular terms as opposed to the structure of the headline. Sentences in the data set alone are too diverse to ensure invariance. For example, if a sentence where 'Croatia' is the target were to be altered such as only 'Croatia' were replaced with 'Japan' or 'United Arab Emirates', would the sentiment remain the same? Such questions necessitate further research and vital to ensure that the model is properly functioning. Considering that some targets are expected to elicit changes but others are not, both invariance and directional expectation tests should be taken into consideration.

Finally, behavioural testing can identify where precisely issues with implicature may lie. In our error analysis, we examine a case where the phrase *liječenje COVIDa-19* ('treatment for COVID-19') receives a positive sentiment for COVID-19, even though the disease is implied to be negative. In this case, the construction in question, *liječenje* + noun (GEN) predominantly refers to 'treatment for an illness' and much more rarely 'treatment for a person (with an illness)'. This failure implies an issue with implicature that deserves further testing, and such a test could include a minimal functionality test in which COVID-19 is replaced with other diseases, health conditions, or names of pandemic, epidemic or endemic events.

6.3.9 Comparative error analysis with and without TAPT

Our error analysis could also be expanded through a comparative analysis of errors produced by one model, with and without TAPT. This analysis would be justified by the fact that improvements with TAPT are not equal across all models, but are in fact unevenly distributed. In fact, no two of the models follow the same pattern of improvement or regression. An analysis could reveal for example, how BERTić managed to only improve in NEG while cseBERT only improved in NTR, or improvement for mBERT is dramatically lower in NTR than either of the XLM-RoBERTa models while having strong gains overall. Considering that an in-depth error analysis with and without TAPT is time consuming, we leave re-evaluation as future work.

6.4 Biases and ethics

Finally, in this section we will discuss some potential biases or ethical concerns related to our work.

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. As we have observed in our data chapter, Chapter 3.1, a small minority of articles in the TakeLab Retriever 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 POS 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.

Lastly, the purpose of this research is, again, ultimately to create a model which will be integrated into the TakeLab Retriever text processing pipeline. 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.

6.5 Summary

In this chapter, we discussed our findings in the context of task adaptation and low resource languages. We also commented on how limitations had affected our research while proposing paths for future developments based on our work. A quick summary of the key points of this chapter follows:

1. General remarks
   (a) TAPT seemed to primarily work if the unlabelled task-related data was new.
   (b) Bigger models benefited less, but still benefited as long as the original, general pre-training data differed enough.
   (c) Languages can be regarded as a super-domain with the possibility of an additional pre-training stage before DAPT or LAPT being beneficial for multi-lingual models.
   (d) Low-resource languages have a lot of unique issues that need to be considered when working with them. TAPT can be included as part of 'the toolbox' for low-resource languages, but used only if applicable.
2. Limitations
   (a) Many of our models had different pre-training objectives, which may have affected our findings. However, we can at least claim that TAPT helps with massively multilingual models.
   (b) Working with neural network-based models always causes issues with replicability. However, we believed that we have done as much as we could to mitigate these issues.
3. Future work
   (a) We proposed a number of ways to expand the current training data set, including identifying news sub-domains, more fine-grained labels, and linguistic information.
   (b) We discussed ways to improve the different models, including DAPT and our proposed language adaptive pre-training (LAPT). We also focus particularly on BERTić, which we believe may have underperformed with TAPT due to being under-trained, although it may simply have seen too much BCMS data already.
   (c) We proposed a number of alternative means for evaluating our models, including behavioural testing and comparative error analyses between TAPT and non-TAPT models.
4. Biases and ethics
   (a) We discussed language identity concerns as well as regional biases.
   (b) We made a statement, stressing that, considering its fallibility, our model should be treated only as a tool for statistical study of general trends.

Chapter 7

Conclusion

This work, performed as a part of internship for TakeLab FER, focused on developing sequence classification models for the task of targeted sentiment analysis, a continuation of work by Barić et al. (2023). We specifically trained these models 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 research resulted in a new high-performing model, XLM-RoBERTa-Large (F_1= 0.771), although there is still room for improvement.

Recall in Chapter 1, our original research question:

> Will task-adaptive pre-training yield improvements in language model performance (F_1-score) in a targeted sentiment analysis task for Croatian headlines?

Our findings indicated that TAPT indeed yielded improvements on some language models, but not all. Instead, we found that each massively multilingual model improved in different ways, while Croatian-dominant models apparently decreased in performance.

Finally, we also introduced the following sub-questions:

1. What challenges remain for our highest-performing model?
2. How does Croatian as a low-resource language affect our task?

For the first question, we found that the highest-performing model (XLM-RoBERTa-Large with TAPT) still suffered from linguistic issues such as irony-detection, understanding aspects, and implicature. There were also some bugs possibly 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 a low-resource 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 an-

other 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 low-resource 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.

Bibliography

Ž. Agić, J. Tiedemann, D. Merkler, S. Krek, K. Dobrovoljc, and S. Može. Cross-lingual dependency parsing of related languages with rich morphosyntactic tagsets. In *Proceedings of the EMNLP'2014 Workshop on Language Technology for Closely Related Languages and Language Variants*, pages 13–24, 2014.

M. A. Alonso, D. Vilares, C. Gómez-Rodríguez, and J. Vilares. Sentiment analysis for fake news detection. *Electronics*, 10(11):1348, 2021.

K. Babić, M. Petrović, S. Beliga, S. Martinčić-Ipšić, M. Matešić, and A. Meštrović. Characterisation of covid-19-related tweets in the croatian language: Framework based on the cro-cov-csebert model. *Applied Sciences*, 11 (21), 2021. ISSN 2076-3417. doi: 10.3390/app112110442. URL https://www.mdpi.com/2076-3417/11/21/10442.

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

A. Barić, L. Majer, D. Dukić, M. Grbeša, and J. Šnajder. Target two birds with one stone: Entity-level sentiment and tone analysis in croatian news headlines. 05 2023.

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

V. Batanović and M. Miličević Petrović. Cross-level semantic similarity for Serbian newswire texts. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1691–1699, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.180.

E. Bender. The# benderrule: On naming the languages we study and why it matters. *The Gradient*, 14, 2019.

I. Bogunović, M. Kučić, N. Ljubešić, and T. Erjavec. Corpus of croatian news portals ENGRI (2014-2018), 2021. ISSN 2820-4042. URL http://hdl.handle.net/11356/1416. Slovenian language resource repository CLARIN.SI.

L. A. M. Bostan, E. Kim, and R. Klinger. 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, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.194.

P. Bourgonje, J. Moreno Schneider, and G. Rehm. 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, Sept. 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4215. URL https://aclanthology.org/W17-4215.

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

D. Brozović Rončević, D. Ćavar, M. Ćavar, T. Stojanov, K. Štrkalj Despot, N. Ljubešić, and T. Erjavec. Croatian language corpus riznica 0.1, 2018. ISSN 2820-4042. URL http://hdl.handle.net/11356/1180. Slovenian language resource repository CLARIN.SI.

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

R. Bugarski. 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, 2019.

Y. Chen and S. Skiena. 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, 2014.

K. Clark, M. Luong, Q. V. Le, and C. D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. *CoRR*, abs/2003.10555, 2020. URL https://arxiv.org/abs/2003.10555.

A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116, 2019. URL http://arxiv.org/abs/1911.02116.

B. Ćoso, M. Guasch, P. Ferré, and J. A. Hinojosa. Affective and concreteness norms for 3,022 croatian words. *Quarterly Journal of Experimental Psychology*, 72(9):2302–2312, 2019.

Y. Cui, W. Che, T. Liu, B. Qin, and Z. Yang. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514, 2021.

K. Dashtipour, S. Poria, A. Hussain, E. Cambria, A. Y. Hawalah, A. Gelbukh, and Q. Zhou. Multilingual sentiment analysis: state of the art and independent comparison of techniques. *Cognitive computation*, 8:757–771, 2016.

D. Demszky, D. Movshovitz-Attias, J. Ko, A. S. Cowen, G. Nemade, and S. Ravi. Goemotions: A dataset of fine-grained emotions. *CoRR*, abs/2005.00547, 2020. URL https://arxiv.org/abs/2005.00547.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

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

U. C. Eiken, A. T. Liseth, H. F. Witschel, M. Richter, and C. Biemann. Ord i dag: Mining norwegian daily newswire. In T. Salakoski, F. Ginter, S. Pyysalo, and T. Pahikkala, editors, *Advances in Natural Language Processing*, pages 512–523, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-37336-0.

P. Ekman et al. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16, 1999.

J. Golubović and C. Gooskens. Mutual intelligibility between west and south slavic languages/ . *Russian linguistics*, pages 351–373, 2015.

S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, and N. A. Smith. Don't stop pretraining: Adapt language models to domains and tasks, 2020.

M. R. Hasan, M. Maliha, and M. Arifuzzaman. Sentiment analysis with nlp on twitter data. In *2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)*, pages 1–4. IEEE, 2019.

M. A. Hedderich, L. Lange, H. Adel, J. Strötgen, and D. Klakow. A survey on recent approaches for natural language processing in low-resource scenarios. *CoRR*, abs/2010.12309, 2020. URL https://arxiv.org/abs/2010.12309.

M. Hu and B. Liu. Mining and summarizing customer reviews. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

M. Jabreel, F. Hassan, and A. Moreno. 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, 2018.

L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao. Target-dependent twitter sentiment classification. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 151–160, 2011.

J. Jukić, F. Jelenić, M. Bićanić, and J. Snajder. ALANNO: An active learning annotation system for mortals. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 228–235, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.eacl-demo.26.

A. Katrekar and B. D. A. AVP. An introduction to sentiment analysis. *GlobalLogic Inc*, 2005.

K. Kenyon-Dean, E. Ahmed, S. Fujimoto, J. Georges-Filteau, C. Glasz, B. Kaur, A. Lalande, S. Bhanderi, R. Belfer, N. Kanagasabai, et al. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, 2018.

S. Khan and M. Shahid. Hindi/bengali sentiment analysis using transfer learning and joint dual input learning with self attention, 2022.

E. V. Kotelnikov. Current landscape of the russian sentiment corpora. *CoRR*, abs/2106.14434, 2021. URL https://arxiv.org/abs/2106.14434.

S. Krek, Š. Arhar Holdt, T. Erjavec, J. Čibej, A. Repar, P. Gantar, N. Ljubešić, I. Kosem, and K. Dobrovoljc. Gigafida 2.0: The reference corpus of written standard Slovene. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3340–3345, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.409.

R. Kulkarni. A Million News Headlines, 2018. URL https://doi.org/10.7910/DVN/SYBGZL.

S. Lee and J. Bloem. 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, 2023.

H. Li, X. Cheng, K. Adson, T. Kirshboim, and F. Xu. 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, May 2012. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2012/pdf/640_Paper.pdf.

Z. Li, Y. Wei, Y. Zhang, X. Zhang, and X. Li. 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, 2019.

C. Lin, S. Bethard, D. Dligach, F. Sadeque, G. Savova, and T. A. Miller. Does bert need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association*, 27(4):584–591, 2020.

H. Liu, D. He, and S. Chan. Fraudulent news headline detection with attention mechanism. *Computational Intelligence and Neuroscience*, 2021:1–7, 2021.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL http://arxiv.org/abs/1907.11692.

A. Ljajić, N. Prodanović, D. Medvecki, B. Bašaragin, and J. Mitrović. Uncovering the reasons behind covid-19 vaccine hesitancy in serbia: Sentiment-based topic modeling. *Journal of Medical Internet Research*, 24(11):e42261, 2022.

N. Ljubešić. Text collection for training the BERTić transformer model BERTić-data, 2021. ISSN 2820-4042. URL http://hdl.handle.net/11356/1426. Slovenian language resource repository CLARIN.SI.

N. Ljubešić and K. Dobrovoljc. 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, Aug. 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-3704. URL https://aclanthology.org/W19-3704.

N. Ljubešić and T. Erjavec. hrwac and slwac: Compiling web corpora for croatian and slovene. In I. Habernal and V. Matousek, editors, *Text, Speech and Dialogue - 14th International Conference, TSD 2011, Pilsen, Czech Republic, September 1-5, 2011. Proceedings*, Lecture Notes in Computer Science, pages 395–402. Springer, 2011.

N. Ljubešić and F. Klubička. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden, 2014. Association for Computational Linguistics.

N. Ljubešić and D. Lauc. 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, Apr. 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.bsnlp-1.5.

N. Ljubešić, I. Markov, D. Fišer, and W. Daelemans. The lilah emotion lexicon of croatian, dutch and slovene. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media. Barcelona, Spain (Online), ACL, pp. 153–157, December, 2020*, pages 1–5, 2020.

N. Ljubešić, M. Stupar, T. Jurić, and Agić. Izgradnja modelov za prepoznavanje imenskih entitet za hrvaščino in slovenščino. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave*, 1(2):35–57, dec. 2013. doi: 10.4312/slo2.0.2013.2.35-57. URL https://journals.uni-lj.si/slovenscina2/article/view/6925.

N. Ljubešić, M. Miličević Petrović, and T. Samardžić. Borders and boundaries in bosnian, croatian, montenegrin and serbian: Twitter data to the rescue. *Journal of Linguistic Geography*, 6(2):100–124, 2018. doi: 10.1017/jlg.2018.9.

X. Ma, P. Xu, Z. Wang, R. Nallapati, and B. Xiang. 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, Nov. 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-6109. URL https://aclanthology.org/D19-6109.

I. Makogon and I. Samokhin. 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, 2021.

G. L. Martin, M. E. Mswahili, and Y.-S. Jeong. Sentiment classification in swahili language using multilingual bert, 2021.

M. Mochtak, P. Rupnik, and N. Ljubešič. The parlasent-bcs dataset of sentiment-annotated parliamentary debates from bosnia-herzegovina, croatia, and serbia. *arXiv preprint arXiv:2206.00929*, 2022.

M. M. Mutlu and A. Özgür. A dataset and bert-based models for targeted sentiment analysis on turkish texts, 2022.

N. R. Naredla and F. F. Adedoyin. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1):100064, 2022. ISSN 2667-0968. doi: https://doi.org/10.1016/j.jjimei.2022.100064. URL https://www.sciencedirect.com/science/article/pii/S2667096822000088.

A. Onan and M. A. Toçoğlu. A term weighted neural language model and stacked bidirectional lstm based framework for sarcasm identification. *IEEE Access*, 9:7701–7722, 2021. doi: 10.1109/ACCESS.2021.3049734.

M. A. Palomino and F. Aider. Evaluating the effectiveness of text preprocessing in sentiment analysis. *Applied Sciences*, 12(17), 2022. ISSN 2076-3417. doi: 10.3390/app12178765. URL https://www.mdpi.com/2076-3417/12/17/8765.

I. Papadimitriou, K. Lopez, and D. Jurafsky. Multilingual bert has an accent: Evaluating english influences on fluency in multilingual models, 2023.

I. Pavlopoulos. Aspect based sentiment analysis. *Athens University of Economics and Business*, 2014.

M. Prelevikj and S. Žitnik. Multilingual named entity recognition and matching using bert and dedupe for slavic languages. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 80–85, 2021.

M. Ptiček. How good bert based models are in sentiment analysis of croatian tweets: comparison of four multilingual berts. pages 175–182, 2021.

K. Ravi and V. Ravi. A survey on opinion mining and sentiment analysis: tasks, approaches and applications. *Knowledge-based systems*, 89:14–46, 2015.

M. T. Ribeiro, T. Wu, C. Guestrin, and S. Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442.

A. Rietzler, S. Stabinger, P. Opitz, and S. Engl. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*, 2019.

A. Rogers, O. Kovaleva, and A. Rumshisky. A primer in bertology: What we know about how BERT works. *CoRR*, abs/2002.12327, 2020. URL https://arxiv.org/abs/2002.12327.

P. Rupnik, T. Kuzman, and N. Ljubešić. 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, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.vardial-1.11.

H. Saif, M. Fernandez, Y. He, and H. Alani. Evaluation datasets for twitter sentiment analysis: a survey and a new dataset, the sts-gold. 2013.

T. A. Salgueiro, E. R. Zapata, D. Furman, J. M. Pérez, and P. N. F. Larrosa. A spanish dataset for targeted sentiment analysis of political headlines, 2022.

T. Samardžić, N. Ljubešić, and M. Miličević. Regional linguistic data initiative (ReLDI). In *The 5th Workshop on Balto-Slavic Natural Language Processing*, pages 40–42, Hissar, Bulgaria, Sept. 2015. INCOMA Ltd. Shoumen, BULGARIA. URL https://aclanthology.org/W15-5306.

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

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

R. Sennrich, B. Haddow, and A. Birch. Neural machine translation of rare words with subword units. *CoRR*, abs/1508.07909, 2015. URL http://arxiv.org/abs/1508.07909.

R. Shekhar, M. Karan, and M. Purver. Coral: a context-aware croatian abusive language dataset. *arXiv preprint arXiv:2211.06053*, 2022.

C. Strapparava and R. Mihalcea. SemEval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL https://aclanthology.org/S07-1013.

M. Tadić. Building the croatian-english parallel corpus. In *Proceedings of the Second International Conference on Language Resources and Evaluation*, pages 523–530. Citeseer, 2000.

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

G. Thakkar, N. Mikelic Preradovic, and M. Tadić. 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, May 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.bsnlp-1.4.

T. T. Thet, J.-C. Na, and C. S. Khoo. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6): 823–848, 2010.

M. Ulcar and M. Robnik-Sikonja. Finest BERT and crosloengual BERT: less is more in multilingual models. *CoRR*, abs/2006.07890, 2020. URL https://arxiv.org/abs/2006.07890.

M. Ulčar and M. Robnik-Šikonja. High quality ELMo embeddings for seven less-resourced languages. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4731–4738, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.582.

J. Vásquez, H. Gómez-Adorno, and G. Bel-Enguix. Bert-based approach for sentiment analysis of spanish reviews from tripadvisor. In *IberLEF@ SEPLN*, pages 165–170, 2021.

X. Wan. 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, Oct. 2008. Association for Computational Linguistics. URL https://aclanthology.org/D08-1058.

A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. *CoRR*, abs/1804.07461, 2018. URL http://arxiv.org/abs/1804.07461.

G. Wenzek, M.-A. Lachaux, A. Conneau, V. Chaudhary, F. Guzmán, A. Joulin, and E. Grave. Ccnet: Extracting high quality monolingual datasets from web crawl data. *arXiv preprint arXiv:1911.00359*, 2019.

S. Wu and M. Dredze. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.repl4nlp-1.16. URL https://aclanthology.org/2020.repl4nlp-1.16.

C. Xiang, J. Zhang, F. Li, H. Fei, and D. Ji. A semantic and syntactic enhanced neural model for financial sentiment analysis. *Information Processing & Management*, 59(4):102943, 2022.

L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models, 2021.

J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack. Sentiment analyzer: Extracting sentiments about a given topic using natural language processing techniques. In *Third IEEE international conference on data mining*, pages 427–434. IEEE, 2003.

A. Žagar and M. Robnik-Šikonja. Slovene superglue benchmark: Translation and evaluation. *arXiv preprint arXiv:2202.04994*, 2022.

M. Zhang, Y. Zhang, and D.-T. Vo. Gated neural networks for targeted sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.

W. Zhang, X. Li, Y. Deng, L. Bing, and W. Lam. A survey on aspect-based sentiment analysis: tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, 2022.

P. Zhou, Z. Wang, D. Chong, Z. Guo, Y. Hua, Z. Su, Z. Teng, J. Wu, and J. Yang. Mets-cov: A dataset of medical entity and targeted sentiment on covid-19 related tweets. *arXiv preprint arXiv:2209.13773*, 2022a.

Y. Zhou, Q. Ying, Z. Qian, S. Li, and X. Zhang. Multimodal fake news detection via clip-guided learning, 2022b.

S. Ćurković, D. Dukić, M. Petričević, and J. Šnajder, Nov 2022. URL https://retriever.takelab.fer.hr/.