Beyond Words: Sentiment Analysis of Croatian Language Attitudes

Barbara Kovačić
Ludwig-Maximilians-University
Center for Information and Language Processing
Oettingenstrasse 67, 80538 Munich, Germany
Barbara.Kovacic@campus.lmu.de

Anja Brnjaković, Gaurish Thakkar
University of Zagreb
Department of Information and Communication Sciences
Ivana Lučića 3, 10000 Zagreb, Croatia
{abrnjako, gthakkar}@m.ffzg.hr

Abstract. *Sociolinguistic research can benefit heavily from natural language processing (NLP) tools. Especially language attitude research can be facilitated by using sentiment analysis. By using existing tools for Bosnian-Croatian-Montenegrin-Serbian (BCMS), we try to develop a pipeline that can be used to collect sentiment towards specific attitudes by the example of the Croatian variety, henceforth referred to as the Croatian language. Therefore, we add sentiment labels at the sentence level to a dataset containing comments on language-related newspaper articles. Then we look at the automation of sentiment annotation for this resource, training different transformer models on that task by using the Machine Learning Toolkit MaChAmp. The experiments demonstrate varied performance across transformer models, with the BERTić model achieving the highest accuracy (0.71) in sentiment analysis of Croatian newspaper comments.*

Keywords. Sentiment analysis, Croatian language, Language attitudes

1 Introduction

The Croatian language is distinguished by its three primary dialects: Čakavian, Kajkavian, and Štokavian. Each dialect uses a different term for "what", specifically "ča", "kaj," and "što". The Štokavian dialect underpins the Standard Croatian language, which experienced a revival post-Yugoslavian war in the 1990s. In the aftermath, Croatia introduced a broad language policy (Mønnesland, 1997) aimed at maximising differentiation from the Serbian Standard language. On January 26, 2024, for the first time in history, the Croatian government legally regulated the use of the official and public use of the Croatian language. As announced on the official website of the Croatian government (Vlada Republike Hrvatske/Hina, 2024), the Štokavian dialect has been chosen as the model for the standard. The law prescribes the basic rules of the Croatian standard language in official and public use. The public use of the Croatian language is any oral or written language communication in a space intended for the public or a wider audience, including electronic ones. In media content, in addition to the use of the Croatian language, the Croatian dialect idiom can be used, especially when there are artistic reasons for this.

One unintended outcome of the Croatian language policy is the reduction in the number of Čakavian and Kajkavian dialect speakers. The changing attitudes of Croatian speakers towards the standard language, who increasingly believe it to be more correct than the dialects, may help to explain this phenomenon (Kalogjera, 2001). The concept of "language attitudes" refers to how individuals react to different language varieties and their users, influenced by psychological and sociocultural factors. These attitudes are a blend of cognitive aspects involving beliefs and stereotypes, affective aspects related to evaluations, and behavioural aspects that reflect observable actions and responses (Garrett, 2006).

Several approaches have been developed to investigate language attitudes, such as the societal treatment approach. Garrett (2006) defines this approach as an extensive research category that examines public domain sources, including government discourse, media, and literature. This method focusses on the sociocultural and political context of attitudes, aiming to offer a thorough understanding of the relationship between societal discourses, political events, and individual attitudes. In previous research, computational methods have been applied to study societal treatment, such as Durham's analysis of tweets about the Welsh

English Accent (Durham, 2016). While data collection was automated, the categorisation and analysis were done manually, making the process both time-consuming and resource-intensive.

Sentiment analysis presents a potential solution to this problem by reducing both the time consumption and resource intensity involved in manual categorisation and analysis by facilitating high-level sentiment classification of documents, sentences, or words (Pang et al., 2008). Therefore, we first annotate sentences containing language-related reader comments from different news websites (Bogetić and Batanović, 2020). By doing so, the dataset can be further analysed for language attitudes towards the Čakavian and Kajkavian dialect and their speakers. Finally, we test automatic annotation of sentiment with the current state-of-the-Art transformer models for Croatian, aiming to contribute the infrastructure for further automation of language attitude research.

2 Related Work

Durham is the first to combine sentiment analysis and language attitude research by the example of the Welsh English accent (Durham, 2016). His study aimed to analyse attitudes towards the Welsh accent expressed on Twitter over a nine-month period from September 2012 to May 2013. Data collection involved searching for tweets containing the terms “Welsh" and “accent", resulting in 87,165 extracted tweets. To manage the large number of tweets, only four days each month were selected for analysis. On those days, tweets were coded based on whether their attitudes were mostly positive, negative, or neutral. Categories included tweets showing love or appreciation, negativity, ambiguous attitudes, performance elements, and those discussing British dialects broadly. The analysis excluded retweets and replies to maintain focus. Monthly tweet counts varied between 9,000 to nearly 12,000, with peaks coinciding with events like the premiere of “The Valleys," a reality TV show featuring Welsh accents. A secondary coding phase refined categories based on thematic frequency, aligning findings with established frameworks for discussing attitudes towards accents.

In addition to that, sentiment annotation has been conducted on various BCMS linguistic resources in the past, resulting in a range of BCMS datasets available for sentiment analysis. The Croatian film review dataset, focusing on Croatian, offers approximately 10,000 sentences from film reviews, annotated at the sentence level and sourced from Thakkar et al. (2023). Another Croatian dataset, the dataset for stance and sentiment analysis from user comments, comprises 54 articles and 904 user comments from the newspaper “24 sata”, annotated at the comment level, published by Bošnjak and Karan (2019). The corpora of Thakkar et al. (2023) and Bošnjak and Karan (2019) are particularly relevant to our task because they include texts published by private users on the internet. Given that much of the input on language attitudes comes from social media and news site comment sections, these corpora can be valuable resources. We can use them either as raw data for analysing language attitudes or as a basis for extending our annotated corpus through data synthesis methods.

Furthermore, different technical approaches have been tested on improving existing methods. Babić et al. (2021) developed a framework for characterisation of COVID-19-related tweets in Croatian. Their research methodology involves several key steps focused on enhancing sentiment analysis using Croatian language data related to COVID-19. Initially, the CroCoV-cseBERT model was trained to embed COVID-19 tweets in Croatian as vectors, compared against a fastText baseline model. Four machine learning (ML) models - naive Bayes, random forest, support vector machine (SVM), and multilayer perceptron - were then trained for sentiment classification, using both embedding models. Following this, clustering techniques were applied to identify themes in the tweets, analysing sentiment trends and retweeting behaviours across different pandemic waves and clusters. The methodology developed by Babić et al. (2021) offers valuable insights into sentiment analysis within specific topics, like COVID-19, by employing specialised models and clustering techniques. This approach can be adapted to analyse language attitudes in Croatian texts by identifying dominant themes and sentiment trends in discussions about language policies, regional dialects, or linguistic debates. The ability to detect and categorise sentiment trends could reveal underlying language attitudes across different communities.

Another milestone of sentiment analysis for Croatian is the multitask learning approach, developed by Thakkar et al. (2021). This research employs cross-lingual knowledge transfer between Slovenian and Croatian datasets for sentiment analysis, leveraging their high mutual intelligibility among South Slavic languages. The study hypothesises Slovenian as an effective hub language for this purpose, given both datasets' shared text type (news articles). Using a shared encoder based on the CroSloEngual model, trained on Slovenian, Croatian, and English texts, the methodology includes a multitask learning setup across document-level, paragraph-level, and sentence-level tasks. The experimental setup encompasses five scenarios: zero-shot learning on Slovenian data for Croatian sentiment analysis at the document-level; multitask learning on Slovenian data across all three levels independently of Croatian data; traditional fine-tuning on Croatian data alone; simultaneous training on both Croatian and Slovenian data at the document-level; and training on both datasets but focusing only on document-level classification. The multitask learning approach by Thakkar et al. (2021) is highly relevant

for our research as it leverages cross-lingual knowledge transfer, particularly between Slovenian and Croatian, to improve sentiment analysis. This method could help us better understand language attitudes in multilingual contexts or regions where multiple South Slavic languages are spoken. Moreover, both use cases leverage the CroSloEngual BERT model, which has demonstrated strong performance, establishing it as the state-of-the-art model for Croatian sentiment analysis.

3 Dataset

This section details the dataset used for our sentiment analysis. We explain the steps taken to prepare and process the data, including the extraction of metadata and the selection of sentences for annotation. The section also describes our annotation process, where annotators classified sentiment and identified targets. Finally, we discuss the inter-annotator agreement, noting the challenges faced, and the measures implemented to ensure the dataset's reliability.

3.1 Original Dataset

We require a source containing natural language from public discussions about language itself, which is why we utilise an existing dataset provided by Bogetić and Batanović (2020) called MetaLangNEWS-COMMENTS-Hr[1]. The corpus is a collection of reader comments from 738 news articles, consisting of 21,533 texts and 823,459 tokens, published in the five-year period of January 1, 2015 - January 1, 2020. It was designed to facilitate research on metalanguage, as well as to provide insight into linguistic ideologies, language policy and planning and contemporary debates on language defining, naming, and standardisation. The corpus is available in .txt, .xml and CoNLL-U format. The mentioned formats were used for data preparation, extraction and processing.

3.2 Data Preparation, Extraction and Preprocessing.

The following is a brief description of our data collection and annotation process. We begin by extracting data from .xml files, where each file represents a news article. From these files, we extract metadata such as document ID, source name, article title, author, and text. Next, we process the corresponding CoNLL-U files, which contain the article comments split into sentences. We focus exclusively on sentences from the comments, as sentiment in news articles and titles tends to be neutral, while sentiment within reader comments can vary significantly depending on the topic. We then merge the metadata with the sentences by matching them to the document ID. This approach allows us to filter out news articles without comments and enhance data analysis possibilities, such as analysing sentiment by source or author. As a result, we generate a .csv file containing 57,257 sentences from 465 articles, containing the following metadata:

- global-id: Global document ID, in the form 'source-id'-'local-id'
- source-name: Full name of the source website
- article-title: Article title in its original script
- article-time: Date on which the article was published
- article-author: Name or initials of the article author
- article-text: Main text of the article in its original script
- sentence-id: ID of the sentence in the CONLL-u file
- sentence: target sentence

Due to time limitations, we pick the first 3,000 sentences to be annotated.

3.3 Annotation Environment

For our annotation environment, we use the Google Drive ecosystem. We create a folder on Google Drive for this purpose, where each annotator has their own Google spreadsheet. This spreadsheet includes the dataset from the aforementioned .csv file and features additional columns for "sentiment" and "target". In the "sentiment" column, we implement a drop-down menu allowing annotators to select the tags "positive", "neutral", "negative" or "mixed", the latter being used for sentences exhibiting both positive and negative sentiments, such as those with nested clauses. This setup aims to minimise spelling errors in the annotations.

3.4 Sentiment Annotation

For manual sentiment annotation, three annotators are involved in the process. The annotators are graduate students of information and communication sciences who are native speakers of the Croatian language and the Štokavian dialect. They have no prior experience in data annotation. Before annotation, annotation guidelines, describing each label with examples, were created. the guidelines were defined following the works from Mohammad (2016) and (Kralj Novak et al., 2015).

The annotators' task is to manually determine the polarity of each sentence and identify the aspect. This task is defined as a sentence-level sentiment analysis. As shown in the example below, annotators classify polarity into one of four categories: "positive", "neutral", "negative" and "mixed". For each sentence, they also identify the target object (aspect) - such as a person, place, subject, or other entity-toward which the opinion is directed.

[1] http://hdl.handle.net/11356/1370

- **Hr**: “ne pravite se ludi - ćirilica je u ovom slučaju negativan simbol koljača”
- **En**: “don’t pretend to be crazy - the Cyrillic alphabet in this case is a negative symbol of the butcher”
- **Label**: negative
- **Target**: Cyrillic alphabet

Each sentence is annotated three times by three different annotators. During the annotation phase, annotators must report any uncertainties and document them in the Notes column of their assigned spreadsheet. Difficult cases are discussed and incorporated into the annotation guidelines.

We also create a Google Sheets Dashboard that aggregates values from each annotation spreadsheet and colour-codes discrepancies in the labels: red for total disagreement, yellow for partial agreement, and green for full agreement. This overview facilitates discussions on cases where annotations differ and helps determine the final annotation. Once each annotator has completed 3,000 sentences in their spreadsheet, the entire dataset is re-annotated to produce the final annotated dataset. During the annotation process, it is noted that some sentences were poorly split, so these examples are removed from the dataset, resulting in a total of 2,938 annotated sentences. After the annotation and cleaning, The dataset includes 18 authors, 26 articles, 25 dates, and 5 sources. The final data characteristics for the first task is shown in Table 1.

Table 1: Distribution of Sentiment Labels in the Dataset. The table presents the distribution of sentiment labels within the final dataset, showcasing the number of instances for each sentiment category: Positive, Neutral, Negative, and Mixed. The total number of instances in the dataset is 2,938.

Label	Number of instances
Positive	189
Neutral	939
Negative	1662
Mixed	148
Total	2938

3.5 Inter-Annotator-Agreement

For calculating inter-annotator agreement, we use Fleiss’ Kappa, a statistical measure developed by Joseph Fleiss (1971). Fleiss’ Kappa measures agreement among multiple annotators by comparing their classifications against chance expectations. It involves calculating observed and expected agreement rates and normalising them. The Kappa values range from -1 (complete disagreement) to 1 (perfect agreement), with 0 indicating chance-level agreement. Values from 0.01 to 1.00 indicate increasing levels of agreement: slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (0.81-1.00). Using the Google Sheets Dashboard, Fleiss’ Kappa is calculated as 0.47, indicating a moderate level of agreement (0.41-0.60) among annotators. The moderate Kappa suggests that while there is a reasonable level of consensus among annotators, some variability remains. This variability can impact the reliability of the final results, indicating that further refinement or additional training of the annotators may be needed to improve consistency. It may also suggest that the dataset is ambiguous, which can impact the accuracy and performance of the language model trained on it.

4 Methodology

For our experiments, we use the machine learning toolkit MaChAmp (Van Der Goot et al., 2021). MaChAmp uses a pre-trained contextualised model as its initial encoder, fine-tuning the layers with an inverse square root learning rate decay and linear warm-up. Each task has a specific decoder trained for that task. For text classification, MaChAmp predicts labels using the embedding of the first token, inserting special tokens for tasks involving multiple sentences to delineate sentence boundaries. By default, the MaChAmp toolkit uses multilingual BERT (mBERT) (Devlin et al., 2019) as embeddings. mBERT supports over 100 languages with the largest Wikipedia, including Croatian and other varieties of BCMS. But the toolkit also allows training on other models.

Another multilingual transformer model including Croatian is CroSlo-Engual BERT (Ulčar and Robnik-Šikonja, 2020), hereinafter abbreviated as cseBERT. This model is trained on Croatian, Slovenian, and English corpora to support less-resourced languages. The training involved 5.9 billion tokens from monolingual corpora, including news articles and general web crawls. These corpora were preprocessed and deduplicated, ensuring high-quality data. Word piece vocabularies were created with the bert-vocab-builder tool, resulting in 49,601 tokens for cseBERT. The model employs the bert-base architecture, featuring a 12-layer bidirectional transformer encoder with 768 hidden units and 110 million parameters. Whole word masking was used for the masked language model task, and case information was preserved. This model is publicly available and provides a significant resource for enhancing NLP applications in Croatian, Slovenian, and English. Initially developed for part-of-speech tagging, named entity recognition (NER), and dependency parsing, cseBERT has also been applied to sentiment analysis projects like analysing COVID-19 communication on Twitter (Babić et al., 2021) and newspaper articles (Thakkar et al., 2021).

To be able to compare multilingual with monolingual approaches, we finally use MaChAmp in combination with BERTić (Ljubešić and Lauc, 2021). BERTić was trained over 8 billion tokens in BCMS. The model uses several existing datasets: the hrWaC

corpus (Croatian top-level domain), the srWaC corpus (Serbian top-level domain), the bsWaC corpus (Bosnian top-level domain), the cnrWaC corpus (Montenegrin top-level domain), and the Riznica corpus (Croatian literary works and newspapers). Additionally, new web crawls for Bosnian, Croatian, and Serbian domains were performed and labelled as CLASSLA web corpora, which were deduplicated to remove identical sentences. It also incorporates the cc100 corpora from CommonCrawl data, further deduplicated to remove overlaps with the WaC and CLASSLA corpora. The final dataset comprises 8.3 billion words. Text preprocessing preserved all Unicode characters. Common tasks executed on BERTić are part-of-speech tagging, NER, geolocation prediction, commonsense causal reasoning and hate speech detection.

Given that our dataset is rather small, we want to enhance it and also explore whether incorporating different styles of text can improve the performance of the language model. Therefore, we use the Croatian Film Review Dataset (Cro-FiReDa) (Thakkar et al., 2023)[2]. It is the first sentiment movie review dataset for the Croatian language. It includes reviews from a Croatian movie review website, covering genres such as adventure, series, and sci-fi. Each review consists of the review text, a summary, an overall assessment score, and the review date. The dataset contains 216 adventure-related reviews, 114 sci-fi reviews, and 76 series reviews. The sentiment annotation task was conducted at the sentence level, with each sentence labelled as "negative", "neutral", "positive", "mixed", or "other/sarcasm". Reviews were segmented into sentences and annotated by undergraduate students in linguistics and informatics. The annotation was facilitated by a deep-learning sentiment classification model trained on the SentiNews dataset, which includes Croatian and Slovenian news articles. The final dataset comprises 10,464 sentences, with 59 percent labelled as neutral. The dataset, annotation guidelines, trained models, and associated code are to be made publicly available. As MaChAmp requires the same format for all datasets used, we add a column called "sentence_i" to Cro-FiReDa. Additionally, we remove all sentences labelled as "other/sarcasm" because detecting sarcasm can be particularly challenging and may introduce significant noise into the data, complicating the sentiment analysis and potentially affecting the model's accuracy.

Moreover, we split our annotated dataset, hence after abbreviated as "LA" into a train (LA-train), a dev (LA-dev) and test (LA-test) dataset by using the ratio 70-10-20. By that, we get the division of the datasets, shown in Table 2.

[2] https://shorturl.at/Pvnr3

Table 2: Dataset train-test distribution. The table outlines the splitting of Cro-FiReDa and our annotated dataset LA, into training, development, and test sets.

	Cro-FiReDa	LA
Train	7480 sentences	2056 sentences
Dev	832 sentences	291 sentences
Test	2079 sentences	591 sentences

5 Experimental Setup

In total, we run 12 experiments, using four systems and three language models. Our reference system (REF) uses the complete Cro-FiReDa (Thakkar et al., 2023) dataset. By using it, we want to create reference values for the best case scenario of how well our dataset and language model can perform. Then, we have two baseline systems, called BASE-1 and BASE-2. In contrast to the reference system, we use the baseline system to see, how well the language model performs when trained on Cro-FiReDa and tested on our annotated dataset. While BASE-1 is trained and further developed with the Cro-Fi-ReDA datasets, the BASE-2 system uses the LA-dataset for further development. Our final system is the COM system, which is trained, developed and tested on our LA-dataset. Each system is then used once with mBERT, cseBERT and BERTić.

6 Results

Table 3 presents the results of our experiments. Here we can see that for REF, the best performance is achieved with the BERTić model, achieving the highest accuracy (0.80) and F1-score (0.77) among the three models. This suggests that BERTić is particularly effective at capturing the nuances required for the REF task, likely due to its specialised training on Croatian social media data. However, the relatively high F1-score indicates that while accuracy is high, there remains some challenge in correctly identifying all relevant instances, possibly due to nuances in sentiment and expression typical of newspaper comments compared to formal language.

The BASE-1 system performs better with cseBERT (0.57) and BERTić (0.56), showing significantly lower performance with mBERT (0.44). This variation highlights the sensitivity of the system's performance to the choice of pre-trained model, where domain-specific training appears to offer advantages over more general models like mBERT. The differences in source sentences between the Cro-FireDa dataset used by BASE-1 and the LA dataset may contribute to these disparities, as the linguistic and sentiment characteristics vary widely between these sources. Moreover, the substantial dataset imbalance in the LA dataset, with a predominance of negative and neutral labelled sentences compared to the movie review dataset, may skew performance metrics, particularly affecting the F1-score.

The BASE-2 system shows a similar trend to BASE-1, with an accuracy of 0.59 with cseBERT and 0.56 with BERTić. This consistency across different systems further underscores the influence of dataset origin on performance outcomes. The specific challenges posed by social media text in capturing sentiment accurately are evident here, affecting both accuracy and F1-score. The dataset imbalance also plays a role, as models trained on datasets with uneven class distributions may struggle to generalise well to new data, particularly in identifying less represented classes like mixed sentiment.

The COM system shows strong performance across all models, with the best accuracy using BERTić (0.71). However, its F1-score is highest with cseBERT (0.59) and notably lower with BERTić (0.41). This discrepancy between accuracy and F1-score suggests that while the model predicts many labels correctly, it may struggle with the more nuanced task of correctly identifying all instances of sentiment, particularly mixed sentiments, which can be expressed subtly in newspaper comments.

In summary, these results highlight several factors influencing performance, including the domain of training data (movie reviews vs. newspaper comments), the specificity of the pre-trained models used, and the inherent challenges of sentiment analysis in informal, user-generated content. The differences in annotated datasets and the level of agreement between annotators are likely another reason why model performance is not the same across all systems and metrics. Additionally, the dataset imbalance in the annotated dataset poses a challenge, impacting model training and evaluation, particularly in accurately predicting less represented sentiment classes.

7 Conclusion

In conclusion, our study utilises the MetaLangNEWS-COMMENTS-Hr corpus to explore sentiment analysis on newspaper comments, focussing on annotating and analysing 2938 sentences across 26 articles. Through annotation and cleaning processes facilitated by our annotation environment, we achieve a moderate inter-annotator agreement of 0.47 using Fleiss' Kappa. Employing MaChAmp and various transformer models, including mBERT, cseBERT, and BERTić, our experiments revealed nuanced performance variations influenced by dataset origin, model specificity, and the challenge of sentiment analysis in informal textual contexts. Despite these complexities, our findings underscore the viability of leveraging specialised models like BERTić for achieving robust sentiment analysis results in Croatian newspaper comments. The results show that automatic annotation works best when the language model is trained on a dataset using the same style of language.

Table 3: Results. The table presents the performance metrics (accuracy and F1-score) of different models (mBERT, cseBERT, BERTic) across various systems. REF represents the reference system, which has been exclusively trained and tested on the Cro-FiReDa dataset. BASE-1 represents our first baseline system, trained and developed on Cro-FiReDa before tested on our annotated LA dataset. In contrast, our second baseline system BASE-2 has used Cro-FiReDa only for training, while development and test have been conducted with the LA dataset. Finally, COM represents our experimental system which is only trained, developed and tested on the LA dataset. For each model, the accuracy (ACC) and F1-score (F1) are reported, highlighting the comparative effectiveness of each approach. The highest achieved values are shown in bold

	mBERT		cseBERT		BERTić	
	ACC	F1	ACC	F1	ACC	F1
REF	0.75	0.70	0.79	0.75	**0.80**	**0.77**
BASE-1	0.44	0.44	**0.57**	**0.54**	0.56	0.53
BASE-2	0.47	0.45	**0.59**	**0.54**	0.56	0.53
COM	0.68	0.54	0.68	**0.59**	**0.71**	0.41

8 Future Work

Further improvement to the approach can be achieved by increasing the size of the dataset. We can either further annotate the dataset or combine the existing dataset with the newspaper dataset from Bošnjak and Karan (2019). Not only would this provide more data using the same style of language, but also increase the variety of newspaper types, as the majority of our sentences are derived from articles from the Croatian newspaper "Večernji list", while Bošnjak and Karan's dataset consists of comments from the newspaper "24 sata". Another improvement of results can be achieved by balancing the dataset and adding more sentences, labelled as positive or mixed. Additionally, further improvement of training for the annotators as well as annotation guidelines can increase the inter-annotator agreement. By doing so, the annotations for the training dataset would increase in quality, leading to better training of the language model and increasing its accuracy. Also, we use the default parameters defined by MaChAmp. When we fine-tune these parameters, increased results are to be expected. Finally, to gain insights from the dataset, the next step is to identify the targets of the sentiment within the sentences. For that, we can either use the approach from Babić et al. (2021) of clustering topics from COVID-19-related tweets. An attractive alternative is also enhancing our experimental setup by training aspect detection and automatically defining a target within the annotation process. Here, we can use the targets, identified by the annotators, as a training base.

Limitations

The performance of pre-trained models in highly specialised or domain-specific tasks may be limited due to the broad coverage of topics in their training data. The pre-trained models learn from the data they are trained on, which can result in the introduction of any inherent biases in the training data. This bias can affect model outputs, particularly when the data do not represent all demographic, cultural, or social groups.

References

Babić, K., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., and Meštrović, A. (2021). Characterisation of covid-19-related tweets in the croatian language: framework based on the cro-cov-csebert model. *Applied Sciences*, 11(21):10442.

Bogetić, K. and Batanović, V. (2020). Annotated corpus of croatian language-related news comments MetaLangNEWS-COMMENTS-hr. Slovenian language resource repository CLARIN.SI.

Bošnjak, M. and Karan, M. (2019). Data set for stance and sentiment analysis from user comments on croatian news. In *Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing*, pages 50–55.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Durham, M. (2016). Changing attitudes towards the welsh english accent: A view from twitter. *Sociolinguistics in wales*, pages 181–205.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Garrett, P. (2006). Language attitudes. In *The Routledge Companion to Sociolinguistics*, pages 136–141. Routledge.

Kalogjera, D. (2001). On attitudes toward Croatian dialects and on their changing status. *International Journal of the Sociology of Language*, (147):91–100.

Kralj Novak, P., Smailović, J., Sluban, B., and Mozetič, I. (2015). Sentiment of emojis. *PloS one*, 10(12):e0144296.

Ljubešić, N. and Lauc, D. (2021). Bertić-the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 37–42.

Mohammad, S. (2016). A practical guide to sentiment annotation: Challenges and solutions. In *Proceedings of the 7th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pages 174–179.

Mønnesland, S. (1997). Emerging literary standards and nationalism. The disintegration of Serbo-Croatian. In *Actas do I Simposio Internacional sobre o Bilingüismo*, pages 1103–1113.

Pang, B., Lee, L., et al. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2):1–135.

Thakkar, G., Mikelić Preradović, N., and Tadić, M. (2021). Multi-task learning for cross-lingual sentiment analysis. In *2nd International Workshop on Cross-lingual Event-centric Open Analytics (CLEOPATRA 2021)*, pages 76–84.

Thakkar, G., Preradović, N. M., and Tadić, M. (2023). Croatian film review dataset (cro-fireda): A sentiment annotated dataset of film reviews. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 25–31.

Ulčar, M. and Robnik-Šikonja, M. (2020). Finest bert and crosloengual bert: less is more in multilingual models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings 23*, pages 104–111. Springer.

Van Der Goot, R., Üstün, A., Ramponi, A., Sharaf, I., and Plank, B. (2021). Massive choice, ample tasks (machamp): A toolkit for multi-task learning in nlp. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197.

Vlada Republike Hrvatske/Hina (2024). Sjednica vlade: Zakonom o hrvatskom jeziku osigurava se sustavna i stručna skrb o njemu. https://vlada.gov.hr/vijesti Last Accessed: 2024-06-27.