Croatian Film Review Dataset (Cro-FiReDa): A Sentiment Annotated Dataset of Film Reviews

Gaurish Thakkar and **Nives Mikelić Preradović** and **Marko Tadić**

Faculty of Humanities and Social Sciences, University of Zagreb, Zagreb 10000, Croatia

gthakkar@m.ffzg.hr, nmikelic@ffzg.hr,marko.tadic@ffzg.hr

Abstract

This paper introduces Cro-FiReDa, a sentiment-annotated dataset for Croatian in the domain of movie reviews. The dataset, which contains over 10,000 sentences, has been annotated at the sentence level. In addition to presenting the overall annotation process, we also present benchmark results based on the transformer-based fine-tuning approach.

1 Introduction

The goal of sentiment analysis is to classify the polarity of text (*e.g.*, positive, negative, neutral, or mixed). In this paper, we describe the process of annotating a sentiment analysis dataset in Croatian. As shown in the example below, the label indicates the sentiment polarity of the text.

Hr "I bio sam zadivljen i tijekom finalne borbene scene ."

En "And I was also amazed during the final battle scene."

- **Label** : positive

Croatian is a low-resource language in terms of sentiment analysis resources. There is currently no Croatian dataset for the domain of movie reviews. The dataset presented here is the first sentiment movie review dataset. The texts for the annotation campaign are taken from the Croatian movie review website and cover multiple genres, namely adventure, series (serija), and sci-fi. In addition to the other metadata described below, the website includes a summary of the entire text of the author's review. The dataset, annotation guidelines, trained models, and associated code will be made available to the public. In this work, we describe our entire workflow for creating the resource. We also present the experimental scores for the sentiment analysis task using pre-trained transformer models.

The rest of the paper is structured as follows: In Section 2, we review the related work on the dataset with regard to its annotation as well as modelling. In Section 3, we describe the annotation process in detail. In Section 4, we present the statistics of the annotated dataset before presenting the baseline scores in Section 5. We complete the paper with the conclusion, discussion, and future work in Section 6.

2 Related Work

In this section, we will highlight the related work on resources and models for sentiment analysis. Sentiment analysis is a well-researched field, and there are a number of resources for various languages, such as English (Maas et al., 2011; Pang and Lee, 2005; Keung et al., 2020), German (Cieliebak et al., 2017; Sänger et al., 2016; Clematide et al., 2012), French (Apidianaki et al., 2016), and Italian (Basile and Nissim, 2013). There are few resources available for Croatian sentiment analysis. The stance (and sentiment) annotated dataset (Bošnjak and Karan, 2019) contains comments submitted by users for online news articles. Pelicon et al. (2020) created a dataset for sentiment analysis of Croatian news articles and performed zero-shot classification using Slovene resources. Zhou et al. (2015) performed multiple levels of sentiment analysis on multilingual Wikipedia articles using machine translation. Öhman et al. (2020) compiled a parallel dataset for sentiment and emotion analysis based on movie subtitles. The dataset was created by manually annotating 25K Finnish and 30K English sentences, which were then projected onto 30 other languages, including Croatian. Agić et al. (2010) presented rule-based annotated Croatian news articles in the finance domain that captured the general sentiment of the text. Rotim and Šnajder (2017) compiled a dataset of gaming review text spans in Croatian that were tagged with positive and negative labels. There are also a few

Croatian sentiment lexicons, such as those developed by Ljubešić et al. (2020); Glavaš et al. (2012). The ParlaSent-BCS (Mochtak et al., 2022) dataset is another resource that has Croatian sentences in parliamentary debates tagged with sentiment polarity. Tikhonov et al. (2022) provide an overview of existing resources for East European languages, including Croatian.

3 Text and Annotations

In this section, we describe our annotation procedure in greater depth. First, we describe the backgrounds of the annotators. Second, the guidelines and methodology for annotation are explained. Third, the statistical aspects of the dataset are discussed.
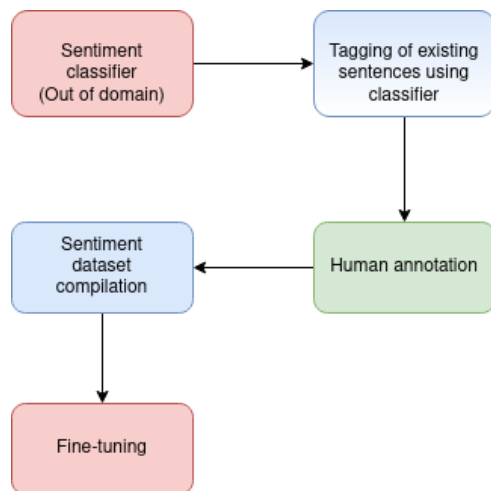
3.1 Annotation Procedure

Figure 1: The dataset creation process.

The task is defined as a sentence-level sentiment task in which each sentence in the training set is annotated with a single label. The dataset consists of professional reviews from the Croatian movie review website[1]. The adventure, TV series, and science fiction (sci-fi) genres were chosen as subcategories. Each review instance is accompanied by the following data fields:

1. **Review**: the text written by the professional reviewer.

2. **First impression**: short summary of the overall review.

3. **Overall assessment**: the score assigned by the reviewer. The reviewers rate the film on

[1]https://www.recenzijefilmova.com/

different scales, the scores range from (0-10) to (1-5) stars.

4. **Date**: date of the review.

In addition, the review text has formatted information about the title, IMDB rating, producers, actors, directors, genres, and date of release. The dataset contains a total of 216 adventure-related reviews, 114 sci-fi reviews, and 76 series reviews. We framed the sentiment annotation task as a sentence-level label correction task. The overall methodology is presented in Figure 1. Each review has undergone sentence segmentation, in which the entire review has been broken down into individual sentences. All reviews were sorted by sentence length and divided into groups so that each annotator received an equal amount of sentences, but at the same time, no annotator received partial review text. This was done to make sure that no student received a partial review. An empirical method was used to determine the N=23 groups. A minimum of three (and a maximum of five) annotators have annotated a single sentence. Each review was pre-annotated using the deep-learning sentiment classification model (Thakkar et al., 2021). The classification model was trained using the SentiNews dataset, which is composed of Croatian and Slovenian news articles in a multitask setup, and has reported an F1-score of 63.86. This step has sped up the annotation process, as annotators are no longer required to tag the sentence from scratch, but only correct the tag if it is incorrect. A total of 82 students participated in the study. All the annotators were undergraduate students of linguistics and informatics between the ages of 22 and 24. All the annotators were native Croatian speakers. The final label for a sentence is chosen by a majority vote.

3.2 Annotation Scheme

The guidelines for the annotation were largely adopted from Mohammad (2016). Learners were presented with five categories of sentiment: 1—negative, 2—neutral, 3—positive, 4—mixed, and 5—other/sarcasm. Evidently, the negative review is labelled as negative, while the positive review is labelled "positive". The release date and genre of the film are categorised as neutral facts. Sentences that have both positive and negative connotations are classified as "mixed". If figurative language exists, it is labelled as "other/sarcasm".

The annotation guidelines describe each instance of the label with multiple examples.

3.3 Web Interface

All of our annotation tasks used the online tool INCEpTION (Klie et al., 2018) because it enables simple semantic annotations. The platform simplifies the administration of annotation projects involving many annotators. Because we did not want participants to see each other's work, each group of students was assigned a separate project. Each user was subsequently able to view only the files assigned to him or her after logging into the system with his or her credentials. Before moving on to the next document, each user would perform the annotation process and lock the document. The locking mechanism signified the document's completion and allowed us to monitor its completion status and overall work status. Each student has averaged four hours on the assignment.

3.4 Inter-annotator agreement

Using Fleiss Kappa, we have measured the inter-annotator agreement of the dataset across multiple groups. The scores suggest moderate (0.41-0.60) to substantial (0.61-0.80) levels of annotator agreement. Table 1 lists the agreement for every label. During the phase of judging, the annotators were required to report any uncertainties. The majority of queries pertained to metadata present in the review text, such as the title. There were 843 disagreements in which there was no clear majority winner. These sentences were characterised by conditionals or mixed sentiments and were filtered out as they were not additionally annotated by anyone and will be taken up for future work.

Hr "Za one koji vole ovu vrstu filma , trebali biste biti u mogućnosti uživati , ali za lojalnog ljubitelja izvornog filma , ovaj se može vidjeti kao još jedan od najljepših ili najmanje omiljenih ."

En "For those who like this type of film, you should be able to enjoy it, but for a loyal fan of the original film, this one can be seen as another of the best or least favorite."

3.5 Corpus Statistics

Table 1 shows the statistics for the final sentiment annotated dataset. Out of 10,464 sentences, we have 59 percent neutral statements. This is clear because the majority of the text contains factual information about the movie/series. There are a total of 875 reviews that have text summaries associated with the main text. The mean number of space-separated tokens for review text and summary is 731 and 47, respectively.

Label	# of instances	agreement
neutral	6205	0.51
positive	2031	0.53
negative	1290	0.42
mixed	862	0.30
sarcasm	76	0.04
total	10464	

Table 1: Statistics of the sentiment dataset. Numbers represent sentences. Kappa statistics for each label

3.6 Dataset Analysis

Out of 10,388 samples, around 2,257 instances retained their original classification tag. The remaining 8,131, which constitute around 78 percent of the final dataset, were modified by the annotators. In these modifications, more than 50 percent of the changes (4,813 instances) were from negative label to neutral, followed by a positive to neutral annotation change (1,053 instances). The sentences that changed from non-neutral to neutral were mostly informative, similar to title sentences with polar words. We also sampled a few random reviews and checked the polarity of the individual sentences in the review, ignoring the neutral sentences. This number of positive and negative sentences does hint at the possibility of a relationship with the overall rating of the review given by the reviewer. For instance, if there were an equal number of positive and negative sentences, the movie would receive a 3/5 or 5/10 rating. On the other side, if the review contains more compliments, it will receive a rating higher than 3. Exactly 654 sentences in the groups received the same annotated class provided by the authors.

4 Experiments

4.1 Experimental Setup

We performed experiments for the task of sentiment analysis. To benchmark the dataset on sentiment classification, we use the fine-tuning approach proposed by Devlin et al. (2019). We used the CroSlo-Engual BERT (Ulčar and Robnik-Šikonja, 2020)

as our contextualised pre-trained language model and performed fine-tuning using a softmax classification head. CroSloEngual BERT was trained on corpora from Croatian, Slovenian, and English languages with a total of 5.9 billion tokens. For training, only the positive, negative, and neutral class instances were used. We divided the dataset into train tests in an 80:20 ratio and used 10% of the train set for development. We used a learning rate of 1e-05 and weight decay of 0.02 with early stopping on evaluation loss with patience of 4. A batch size of 16 was used during training. In addition, a hidden dropout and attention of 0.2 were used as regularization constants. Each of the experiments used a GPU with 24 GB of VRAM. Each epoch of sentiment training lasted longer than 20 minutes. In addition, we also present the results utilising the three strategies described in Pelicon et al. (2020). The reported approaches employ 10-fold cross-validation for training stage. A hidden layer (768,250), ReLU activation, and a softmax classification layer are used in the second and third methods (250, number of classes). The overlapped long texts used in the second technique are used to build an oversampled dataset. The third method averages all the vectors corresponding to the overlapping sentences, rather than oversampling them. The vectors are subsequently sent through a ReLU-equipped two-layered classification head.

5 Results

The scores for the sentiment task are reported using the F1, and accuracy (macro) metrics. In the case of fine-tuning setup, each experiment was performed five times with different random seeds, and the mean of all the scores is reported in Table 2. We also tested the XLM-RoBERTa (Conneau et al., 2019) and classla/bcms-bertic (Ljubešić and Lauc, 2021) language models, both of which were pre-trained in Croatian, but the results were no better than the CroSloEngual BERT. All the scores are comparable, as the final scores are reported on the same test set that was held out during the training phase.

5.1 Error Analysis

A manual error analysis points to two major categories of errors. First, there are instances in the annotated set that have polar labels for metadata about the movie. Second, the trained model also has problems dealing with conditionals. Two instances are provided below.

1. **Hr** "Ako tražite nešto zbog čega razmišljate , a usredotočujete se dosta na odnos , onda bi vas ova serija trebala zabaviti."

   En "If you are looking for something to make you think and focus a lot on the relationship, then this series should entertain you."

2. **Hr** "Kad bi samo satovi znanosti u školi bili zabavni."

   En "If only science lessons at school were fun."

5.2 Discussion

All the annotators were presented with a questionnaire to be answered upon the completion of the task. The questionnaire included basic questions like how much time was required on average, good and bad experiences, as well as suggestions for future enhancements. Apart from the enhancement of the user interface for the annotation tool, one common request was to include neutral-positive and neutral-negative. These were mainly sentences that were objective in nature, but invoked sentiment. For example,

Hr "Ocjena na IMDb.com mu je 6,4 / 10 , a na Tomatoesima malih 36%."

En "The rating on IMDb.com is 6.4 out of 10, and on Tomatoes it is 36%."

This was one of the sentences in which two annotators had tagged it neutral, while the other two had tagged it with a negative label.

6 Conclusion and Future Work

With this paper, we have presented the sentiment annotated movie review dataset for Croatian. We performed experiments using curated datasets for the sentiment analysis task for the Croatian language. Out of 21 unique categories of film reviews, to name a few, we have processed only three categories (adventure, series (serija), and sci-fi). In the future, we would like to use the gold-annotated dataset in a distant-supervised learning regime to perform sentiment classification on all the non-annotated reviews. Another area of research would be to formally evaluate how pre-suggestions of the model before manual annotation could influence annotators' decisions. For example, a systematic

Configuration	F1	Accuracy
FT †	79.78 (0.008)	84.71 (0.006)
CV ◇	71.19 (0.007)	80.43 (0.003)
Sampling average ◇	70.84 (0.003)	80.13 (0.002)
CV sampling ◇	70.69 (0.005)	80.18 (0.002)

Table 2: Results of the experiments. †: Devlin et al. (2019). ◇: Methods reported in Pelicon et al. (2020)

comparison of labelling sentences from scratch versus allowing people to correct/retain automated labels could be conducted. In addition, we would like to experiment with mixed and sarcasm-tagged sentences. The dataset also contains metadata, such as genres and document-level sentiment ratings, which can be explored in the future.

Acknowledgement

Limitations

Although the current dataset mainly covers the genres of sci-fi, adventure, and series, there are other genres (games and books) that are missing from the dataset. The models were trained on a 24 GB GPU. Hence, we expect this could limit reproducibility. The downside of the presented approach is the decision to use an existing classifier to pre-annotate the texts. The suggestions could bias the students.

Ethics Statement

The annotated dataset reported in this paper involved manual effort. This is an output of the annotation campaign conducted with students of linguistics and informatics in order to aid in learning about sentiment analysis as part of their coursework. The students were compensated with course credits at the end of the campaign.

References

Željko Agić, Nikola Ljubešić, and Marko Tadić. 2010. Towards sentiment analysis of financial texts in Croatian. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Marianna Apidianaki, Xavier Tannier, and Cécile Richart. 2016. Datasets for Aspect-Based Sentiment Analysis in French. In *International Conference on Language Resources and Evaluation*, Portoro, Slovenia.

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

Mihaela Bošnjak and Mladen Karan. 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, Florence, Italy. Association for Computational Linguistics.

Mark Cieliebak, Jan Milan Deriu, Dominic Egger, and Fatih Uzdilli. 2017. A Twitter corpus and benchmark resources for German sentiment analysis. In *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pages 45–51, Valencia, Spain. Association for Computational Linguistics.

Simon Clematide, Stefan Gindl, Manfred Klenner, Stefanos Petrakis, Robert Remus, Josef Ruppenhofer, Ulli Waltinger, and Michael Wiegand. 2012. MLSA — a multi-layered reference corpus for German sentiment analysis. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3551–3556, Istanbul, Turkey. European Language Resources Association (ELRA).

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Goran Glavaš, Jan Šnajder, and Bojana Dalbelo Bašić. 2012. Experiments on hybrid corpus-based sentiment lexicon acquisition. In *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, pages 1–9, Avignon, France. Association for Computational Linguistics.

Phillip Keung, Yichao Lu, György Szarvas, and Noah A. Smith. 2020. The multilingual amazon reviews corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9. Association for Computational Linguistics. Event Title: The 27th International Conference on Computational Linguistics (COLING 2018).

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

Nikola Ljubešić, Ilia Markov, Darja Fišer, and Walter Daelemans. 2020. 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*, pages 153–157, Barcelona, Spain (Online). Association for Computational Linguistics.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

Michal Mochtak, Peter Rupnik, and Nikola Ljubešić. 2022. The sentiment corpus of parliamentary debates ParlaSent-BCS v1.0. Slovenian language resource repository CLARIN.SI.

Saif Mohammad. 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, San Diego, California. Association for Computational Linguistics.

Emily Öhman, Marc Pàmies, Kaisla Kajava, and Jörg Tiedemann. 2020. XED: A multilingual dataset for sentiment analysis and emotion detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6542–6552, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the ACL*.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *Applied Sciences*, 10(17).

Leon Rotim and Jan Šnajder. 2017. Comparison of short-text sentiment analysis methods for Croatian. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 69–75, Valencia, Spain. Association for Computational Linguistics.

Mario Sänger, Ulf Leser, Steffen Kemmerer, Peter Adolphs, and Roman Klinger. 2016. SCARE — the sentiment corpus of app reviews with fine-grained annotations in German. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1114–1121, Portorož, Slovenia. European Language Resources Association (ELRA).

Gaurish Thakkar, Nives Mikelić Preradović, and Marko Tadić. 2021. Multi-task learning for cross-lingual sentiment analysis. In *Proceedings of the 2nd International Workshop on Cross-lingual Event-centric Open Analytics.*, pages 76–84, Ljubljana, Slovenija. CEUR Workshop Proceedings.

Alexey Tikhonov, Alex Malkhasov, Andrey Manoshin, George-Andrei Dima, Réka Cserháti, Md.Sadek Hossain Asif, and Matt Sárdi. 2022. EENLP: Cross-lingual Eastern European NLP index. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2050–2057, Marseille, France. European Language Resources Association.

M. Ulčar and M. Robnik-Šikonja. 2020. FinEst BERT and CroSloEngual BERT: less is more in multilingual models. In *Text, Speech, and Dialogue TSD 2020*, volume 12284 of *Lecture Notes in Computer Science*. Springer.

Yiwei Zhou, Alexandra Cristea, and Zachary Roberts. 2015. Is Wikipedia really neutral? a sentiment perspective study of war-related Wikipedia articles since 1945. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 160–168, Shanghai, China.

A Appendix

Task	Metric	Value
sentiment	learning rate	1e-5
	weight decay	0.02
	batch size	16
	epochs	10

Table 3: List of hyperparameters, model parameters and their values used during the experiments.

#	Category
1	adventure
2	new-films
3	biography
4	comedy
5	documentary
6	sci-fi
7	thriller
8	sport
9	war
10	western
11	mystery
12	crime
13	family
14	drama
15	music
16	history
17	action
18	romance
19	animation
20	fantasy
21	horror
22	series

Table 4: List of categories.