applied sciences

MDPI

Article

Zero-Shot Learning for Cross-Lingual News Sentiment Classification

Andraž Pelicon [1,2,*], Marko Pranjić [2,3], Dragana Miljković [1], Blaž Škrlj [1,2] and Senja Pollak [1,*]

1 Jožef Stefan Institute, 1000 Ljubljana, Slovenia; dragana.miljkovic@ijs.si (D.M.); blaz.skrlj@ijs.si (B.Š.)
2 Jožef Stefan International Postgraduate School, 1000 Ljubljana, Slovenia; marko.pranjic@styria.ai
3 Trikoder d.o.o., 10010 Zagreb, Croatia
* Correspondence: Andraz.Pelicon@ijs.si (A.P.); senja.pollak@ijs.si (S.P.)

Received: 31 July 2020; Accepted: 25 August 2020; Published: 29 August 2020

Abstract: In this paper, we address the task of zero-shot cross-lingual news sentiment classification. Given the annotated dataset of positive, neutral, and negative news in Slovene, the aim is to develop a news classification system that assigns the sentiment category not only to Slovene news, but to news in another language without any training data required. Our system is based on the multilingual BERTmodel, while we test different approaches for handling long documents and propose a novel technique for sentiment enrichment of the BERT model as an intermediate training step. With the proposed approach, we achieve state-of-the-art performance on the sentiment analysis task on Slovenian news. We evaluate the zero-shot cross-lingual capabilities of our system on a novel news sentiment test set in Croatian. The results show that the cross-lingual approach also largely outperforms the majority classifier, as well as all settings without sentiment enrichment in pre-training.

Keywords: sentiment analysis; zero-shot learning; news analysis; cross-lingual classification; multilingual transformers

1. Introduction

Sentiment analysis is one of the most popular applications of natural language processing (NLP) and has found many areas of applications in customers' product reviews, survey textual responses, social media, etc. It analyzes users' opinions on various topics, such as politics, health, education, etc. In sentiment analysis, the goal is to analyze the author's sentiments, attitudes, emotions, and opinions [1]. Traditionally, such analysis was performed towards a specific entity that appears in the text [2]. A less researched, but nevertheless prominent field of research in sentiment analysis is to shift the focus from analyzing sentiment towards a specific target to analyzing the intrinsic mood of the text itself. Several works try to model feelings (positive, negative, or neutral) that readers feel while reading a certain piece of text, especially news [3,4]. In Van de Kauter et al. [5], the authors claimed that the news production directly affects the stock market as the prevalence of positive news boosts its growth and the prevalence of negative news impedes it. In the context of news media analytics, the sentiment of news articles has been used also as an important feature in identifying fake news [6] and biases in the media [7]. Rambaccussing and Kwiatkowski [8] explored the change in sentiment of news articles from major U.K. newspapers with respect to current economic conditions. Bowden et al. [9] took a step further and tried to improve the forecasting of three economic variables, inflation, output growth, and unemployment, via sentiment modeling. They concluded that, using sentiment analysis, out of the three variables observed, the forecasting can be effectively improved for unemployment.

In the last year, the use of pre-trained Transformer models has become standard practice in modeling text classification tasks. Among the first such models was the BERT (Bidirectional Encoder Representations from Transformers) model developed by [10], which achieved state-of-the-art performance on several benchmark NLP tasks, as well as in real-world applications, e.g., Google search engine [11] and chatbots [12]. The initial model was however pre-trained only on English corpora and could consequently be used only for modeling textual data in the English language. A new version of the BERT model, titled multilingual BERT or mBERT, soon followed. This model was pre-trained on unlabeled data in 104 languages with the largest Wikipedias using a joint vocabulary. Several studies noted the ability of the mBERT model to work well in multilingual and cross-lingual contexts even though it was trained without an explicit cross-lingual objective and with no aligned data [13,14].

In the context of sentiment analysis of news articles, we however identified two potential drawbacks of the mBERT model. The first is that the model accepts the inputs of a fixed length where the length is determined by the length of the context window, i.e., the maximum length of the input sequence during the pre-training phase. Since the training becomes computationally more expensive with the size of the context window, several standard implementations of the mBERT model have the context window set to a maximum length [15]. The standard solution for longer documents is therefore to cut the inputs to the length of the context window [16]. This method however potentially causes the loss of important information that could be present in the later parts of the document. Another potential drawback is that the input representations produced by the Transformer models may encode only a small amount of sentiment information. The pre-training objectives, namely the masked language modeling and next sentence prediction, are designed to focus on encoding general syntactic and certain semantic features of a language. The only explicit sentiment signal the models get is during the fine-tuning phase, when the models are generally trained on a much smaller amount of data.

The paper presents the advances achieved in the scope of the European project H2020 EMBEDDIA (www.embeddia.eu, duration 2019–2021), which focuses on the development of cross-lingual techniques to transfer natural language processing tools to less-resourced European languages with applications to the news media industry. In this paper, we present our approach to cross-lingual news sentiment analysis, where given an available sentiment-annotated dataset of news in Slovene [3], we propose a news sentiment classification model for other languages. In this paper, we focus on Croatian, where the news dataset is provided by 24sata, one of the leading portals in Croatia, and was labeled with the same sentiment annotation scheme as the Slovenian dataset in order to allow comparison in a zero-shot learning setting where no annotations in the target language are expected.

We identify three main contributions of this paper focusing mainly on the cross-lingual zero-shot learning setting. First, we gathered a sentiment-annotated corpus of Croatian news, where the annotation guidelines follow the annotation scheme of the Slovenian sentiment-annotated news dataset [3], therefore enabling cross-lingual zero-shot learning sentiment evaluation. Second, we tested several document representation techniques to overcome one of the shortcomings of the BERT models of not being capable of efficiently processing longer text documents. Last, but not least, we propose a novel intermediate training step to directly enrich the BERT model with sentiment information in order to produce input representations of better quality for sentiment classification tasks. These representations were then tested both in a monolingual setting, as well as in the zero-shot cross-lingual setting, where the model was tested on a different language without any additional target language training. Our experiments show that these representations improve the results in the monolingual setting and achieve a substantially better result than the majority baseline classifier in the cross-lingual setting.

The article is structured as follows. In Section 2, we first present the related work upon which our study builds. In Section 3, we present two datasets of news articles that are manually labeled in terms of sentiment: the existing Slovenian dataset [3] and the newly constructed Croatian test set. Section 4, where we present the methodology, is followed by Section 5, explaining the experimental

setup, with the training regime applied and the evaluation method. Section 6 presents the results of the experiments and discusses their impact, which is followed by qualitative inspection of the models in Section 7. Section 9 presents the conclusions of this work and ideas for future research.

2. Related Work

Traditionally, sentiment analysis was modeled through the use of classical machine learning methods, where especially learners such as support vector machines combined with the TF-IDF text representations proved to be widely successful [17,18]. Lately, however, deep neural networks have become more frequent for sentiment analysis and started outperforming the classical approaches. Mansar et al. [19] used convolutional neural networks (CNN), a variant of neural networks, which are heavily utilized for computer vision. With the help of the convolutional layer, they acquired word-level representations of individual news articles from the learning corpus and combined them with the sentiment score of the individual article, which was obtained with a simple, rule-based model. The attributes were used as input to the fully connected NN. Their model showed the best performance on the SemEval2017 challenge (Task 5, Subtask 2). Moore and Rayson [20] used two models for analyzing sentiment in financial news titles, a support vector machine and a bidirectional LSTM (Long-Short Term Memory) neural network. They reported the LSTM neural network to outperform the SVM modelsby 4–6%.

Several recent works also explored the problem of cross-lingual sentiment analysis. One of the earlier studies [21] employed machine translation to translate a large corpus of sentiment-annotated English training data for the development of a Chinese sentiment classifier. These translated data were then used in addition to the original Chinese data to train an SVM-based classifier. While machine translation can be a good solution for cross-lingual modeling, a quality machine translation system for a particular language pair may not exist or may be expensive to train. Furthermore, machine learning systems struggle with distant language pairs [22]. Zhou et al. [23] developed a cross-lingual English-Chinese attention-based neural architecture for sentiment classification. It utilizes a two-level hierarchical attention mechanism. The first layer of the model encodes each sentence separately by finding the most informative words. Then, the second layer produces the final document representation from lower-level sentence representations. The downside of their work is that the model uses aligned data in two languages, which are not readily available for every language pair. Ref. [24] proposed a representation learning method that utilizes emojis as an instrument to learn language-independent sentiment-aware text representations. The approach is however limited to text types where emojis regularly appear. The cross-lingual sentiment classification approaches presented above also do not address news analysis, but focus on shorter social media texts, where there is no need for adaptation to longer text sequences and they do not leverage cross-lingual Transformer models, such as mBERT, that have been recently introduced as the state-of-the-art for cross-lingual classification tasks. In this paper, we will bridge this gap by proposing a novel approach where we not only leverage standard transfer learning where pretrained language models are fine-tuned for specific classification tasks (in the same or another language), but introduce a novel intermediate training step for sentiment enrichment of BERT models.

The need for labeled data is seen as one of the main obstacles in developing robust cross-lingual systems for natural language processing, especially for low-resource languages. For this reason, research has been focused lately on models that can work in a zero-shot setting, i.e., without being explicitly trained on data from the target language or domain. This training paradigm has been utilized with great effect for several popular NLP problems, such as cross-lingual document retrieval [25], sequence labeling [26], cross-lingual dependency parsing [27], and reading comprehension [28]. More specific to classification tasks, Ye et al. [29] developed a reinforcement learning framework for cross-task text classification, which was tested also on the problem of sentiment classification in a monolingual setting. Jebbara and Cimiano [30] developed models for cross-lingual opinion target extraction, which were tested in a zero-shot setting, similar to ours. Their approaches rely on the

alignment of static monolingual embeddings into the shared vector space for input representation. Fei and Li [31] trained a multi-view cross-lingual sentiment classifier based on the encoder-decoder architecture used for unsupervised machine translation. Their systems showed state-of-the-art performance on several benchmark datasets. The difference from our work is that the datasets used are all product review datasets, which contain considerably shorter texts. Furthermore, as described in Section 1, product reviews contain the target of the modeled sentiment in the text, while news articles generally do not, which makes the two problems different on a more fundamental level.

Novel research has also been done on better input text representation techniques for classification tasks. Tan et al. [32] proposed a clustering method for words based on their latent semantics. The vectors composing the same clusters were then aggregated together into cluster vectors. The final set of cluster vectors was then used as the final text representations. This novel text representation technique showed improvement on five different datasets. Pappagari et al. [33] proposed a modification to the BERT model for long document classification in a monolingual setting. They utilized a segmentation approach to divide the input text sequences into several subsequences. For each subsequence, they obtained a feature vector from the Transformer, which they then aggregated into one vector by applying another LSTM- or Transformer-based model over it. This work has inspired part of our current research for obtaining better Transformer-based representation of long text sequences. Ref. [34] recently presented a Transformer architecture, which is able to produce input representations from long documents in an efficient manner. However, the model they produced based on this architecture was pre-trained only on English data.

3. Datasets

In this section, we present in detail the two datasets of sentiment-labeled news that were used in this experiment.

3.1. SentiNews Dataset in Slovene

We used the publicly available SentiNews dataset (available at https://www.clarin.si/repository/xmlui/handle/11356/1110) [3], which is a manually sentiment-annotated Slovenian news corpus. The dataset contains 10,427 news texts mainly from the economic, financial, and political domains from Slovenian news portals (www.24ur.com, www.dnevnik.si, www.finance.si, www.rtvslo.si, www.zurnal24.si), which were published between 1 September 2007 and 31 December 2013. The texts were annotated by two to six annotators using the five-level Likert scale on three levels of granularity, i.e., on the document, paragraph, and sentence level. The dataset contains information about average sentiment, standard deviation, and sentiment category, which correspond to the sentiment allocation according to the average sentiment score. The dataset statistics are:

- 10,427 documents;
- 89,999 paragraphs;
- 168,899 sentences.

For our news classification experiments, we used the document-level annotations, with 10,427 news articles and an imbalanced distribution of 3337 (32%) negative, 5425 (52%) neutral, and 1665 (16%) positive news, where the sentiment category corresponds to the sentiment allocation according to the average sentiment score. For intermediate training, we also leveraged paragraph-level annotations.

3.2. Croatian Sentiment Dataset

The Croatian dataset was annotated in the scope of project EMBEDDIA and for the purposes of testing cross-lingual classification; therefore, the annotation procedure fully matched the Slovenian dataset [3].

The data came from 24sata, one of the leading media companies in Croatia with the highest circulation newspaper. The 24sata news portal is one of the most visited websites in Croatia, and it

consists of a portal with daily news and several smaller portals covering news from specific topics such as automotive news, health, culinary content, and lifestyle advice. Portals included in the dataset are www.24sata.hr (daily news content, the majority of the dataset), as well as miss7.24sata.hr, autostart.24sata.hr, joomboos.24sata.hr, miss7mama.24sata.hr, miss7zdrava.24sata.hr, www.express.hr, and gastro.24sata.hr.

The dataset statistics are:

- 2025 documents;
- 12,032 paragraphs;
- 25,074 sentences.

As in [3], the annotators chose the sentiment score on the Likert [35] scale (corresponding to the question: Did this news evoke very positive/positive/neutral/negative/very negative feelings?), but for the final dataset, the average annotations were then three classes (positive, negative, and neutral). Annotations were done on three levels: document, paragraph, and sentence level. The distribution of positive, negative and neutral news texts of the document-level annotations used in this study is as follows: 303 (15.1%) positive, 439 (21.5%) negative, and 1283 (63.4%) neutral. They will be made available under a CC license upon acceptance of the paper. More details about inter-annotator agreement and annotation procedure are available in the Appendix A of this paper.

As one of the contributions of this paper is the evaluation of representation learning for long articles, we also provide the statistics of both datasets in terms of length. Table 1 compares the Slovenian and Croatian news datasets in terms of the length of annotated articles. It presents the average number of tokens per article, as well as the length of the longest and shortest articles in the respective datasets. We present the lengths in terms of the standard tokenization procedure where each word and punctuation mark counts as a separate token. However, the BERT model uses a different form of tokenization, namely the WordPiece tokenization [36]. Using this tokenization process, each word is broken into word pieces, which form the vocabulary of the tokenizer. The vocabulary is obtained using a data-driven approach: given a training corpus G and a number of word pieces D, the task is to select D word pieces such that the segmented corpus G contains as much unsegmented words as possible. The selected word pieces then form the vocabulary of the tokenizer. This approach is proven to handle the out-of-vocabulary words better than standard tokenization procedures. Since the inputs to the BERT model have to be tokenized according to this algorithm in order for the model to properly learn, we present the length statistics in terms of BERT's WordPiece tokenization model as well in the column "BERT tokens". We may observe that the average length of the articles in both datasets is relatively long in terms of the BERT tokens. Especially in the Slovenian dataset, which is used for training in this experiment, the average length of an article surpasses the maximum window size of the BERT model, which is set to 512 tokens in the implementation we are using for this work.

Table 1. Length of the articles in the Slovenian and Croatian datasets in terms of the number of tokens. The row "Tokens" presents the length in terms of the standard tokenization procedure, and the row "BERTtokens" presents the length of the articles in terms of BERT's WordPiece tokenization.

	Slovenian			Croatian		
	Min	**Max**	**Mean**	**Min**	**Max**	**Mean**
Tokens	10	2833	350	155	515	273
BERT tokens	19	4961	648	256	816	456

4. Methodology

We tested two approaches, one focusing on techniques for long document representation and the second one on improving the performance on the sentiment analysis task through intermediate pre-training.

In this work, we model sentiment in news articles, which are frequently longer than the BERT context windows, as discussed in Section 1. Therefore, in our first approach, we experiment with several methods for representing longer documents.

The second approach, presented in Section 4.4, proposes a novel technique for sentiment enrichment of mBERT. In standard BERT architectures, the pre-training phase of BERT consists of masked language modeling and next sentence prediction tasks, which are robust, but not necessarily relevant for sentiment classification, as discussed in Section 1. Therefore, we add an intermediate training step where, aside from masked language modeling, the sentiment classification is used as a learning objective. This model is then used for final fine-tuning. The role of intermediate training for BERT is still unexplored in NLP, with some initial experiments presented in [37].

4.1. Beginning of the Document

In the first experimental setting, we produced the document representations by using only the beginning part of the document. We first tokenized the document with the pre-trained multilingual BERT tokenizer. We then took the sequence of 512 tokens from the beginning of the document and fed them to the BERT language model. As proposed in Devlin et al. [10], we used the representation of the [CLS] token produced by the language model as the document representation. The [CLS] token is a special token prepended to every input of the BERT model, which, after fine-tuning, is used to represent the input sequence for classification tasks. We then sent this representation to the classification head composed of a single linear layer. This experiment mimics the usual usage of the BERT pre-trained models for text classification tasks and is included in this work for better benchmarking of other proposed text representation methods.

4.2. Beginning and End of the Document

For the second setting, we tried to produce the document representations by using the beginning and end of the document. The length of the input sequence was retained at 512 tokens. For sequences longer than 512 tokens after tokenization, we took 256 tokens from the beginning of the text and 256 tokens from the end of the text and concatenated them. We then fed the sequence to the BERT language model and used the [CLS] token vector from the last layer as the document representation. This document representation was then fed to the classification head composed of a linear layer.

4.3. Using Sequences from Every Part of the Document

In the third setting, we tried to compose our document representation by using information in the whole document.

For the language model fine-tuning phase, we tokenized each document and broke it into sequences of 512 tokens. We then used a sliding window that moved over all the subsequences in the order they appeared in the original sequence. Each subsequent window would overlap the first fifty tokens from the previous window. This way, we hoped our model would capture the relationships across sentence boundaries. We attached the document sentiment label to each of the subsequences from the same document. Such an oversampled dataset was then used to fine-tune the multilingual BERT language model with the attached linear layer for classification. This method is graphically presented in Figure 1.

After finetuning we again prepared each document in the dataset as described above and sent every subsequence of a particular document to the fine-tuned BERT model. We extracted the [CLS] vector representations from the last layer and combined them into a final document representation. This approach is inspired by the work of Pappagari et al. [33]. The main difference of our study is in the way the subsequence representations are merged into a document representation. In this work, we tested three different ways of combining the output vector representations into the final document representation.

- Using the most informative subsequence representation:

  In this approach, we tried to identify the most informative subsequence for the task at hand. As the BERT language model was fine-tuned on the sentiment classification task, we assumed some notion of the importance of different parts of the text was encoded directly into the vector representations. Using this line of thought, we defined the most informative subsequence as the subsequence with the highest euclidean vector norm. Formally, from the set of ordered subsequence representations: $S = \{x_1, x_2, \ldots x_n\}$ we chose: $x = argmax(||x||_2 : x \in S)$. We then used only this representation as the final vector representation and discarded the rest. The document representation is then sent into a two-layer fully connected neural network, which produces the final predictions.

- Averaging the representations of all subsequences:

  As the first approach is based on a strong assumption and it does not actually utilize the data from the whole document, here we combine all the vector representations of subsequences into one final document representation. We used a relatively naive approach of simply averaging all the vector representations to produce the final document embedding. The document representation is then sent into a two-layer fully connected neural network, which produces the final predictions.

- Using convolutional layers:

  In this approach, we extracted the most informative parts of the document with the use of 1D convolutional neural layers. We used a convolutional filter of size 2 with stride 2 that runs over the produced subsequence representations. This way, the convolutional filter processes the subsequences in pairs and extracts the most informative features from each pair of subsequences from each part of the document. Since we have documents of variable lengths that may be represented by a variable number of subsequences, all the representations were padded with zero vectors up to the maximum length of 6. We used 128 filters to produce 128 feature maps. We then mapped these maps to a final 128-dimensional document vector representation using a max pooling operation. The final embedding is then sent into a linear layer that produces the classification.

The advantage of the first two mapping operations is that, in comparison to the methods proposed in Pappagari et al. [33], they are more computationally efficient as we need to perform simple vector norm and averaging calculations to produce the final document representations. The third mapping operations uses a convolutional layer to map the different subsequences into one document representation. The convolutional networks have proven in the past to be competitive with other text-processing approaches in NLP [38]; therefore, our approach presents an alternative to the LSTM and Transformer-based sequence aggregation.

4.4. Sentiment Enrichment of the mBERT Model

In this approach, the aim is to to induce sentiment information directly into the vectorized document representations that are produced by the multilingual BERT model. To do so, we added an intermediate training step for the mBERT model before the fine-tuning phase. The intermediate training phase consists of jointly training the model on two tasks. The first task we used was the masked language modeling task as described in the original paper by Devlin et al. [10]. We left this task unchanged in hopes that the model would better capture the syntactic patterns of our training language and domain.

For the second task, we used the sentiment classification task, which mirrors the fine-tuning task, but is trained using a different set of labeled data. With this task, we tried to additionally constrain the model to learn sentiment-related information before the actual fine-tuning phase. The task was

formally modeled as a standard classification task where we tried to learn a predictor that would map the documents to a discrete number of classes:

$$\gamma : x \rightarrow C$$

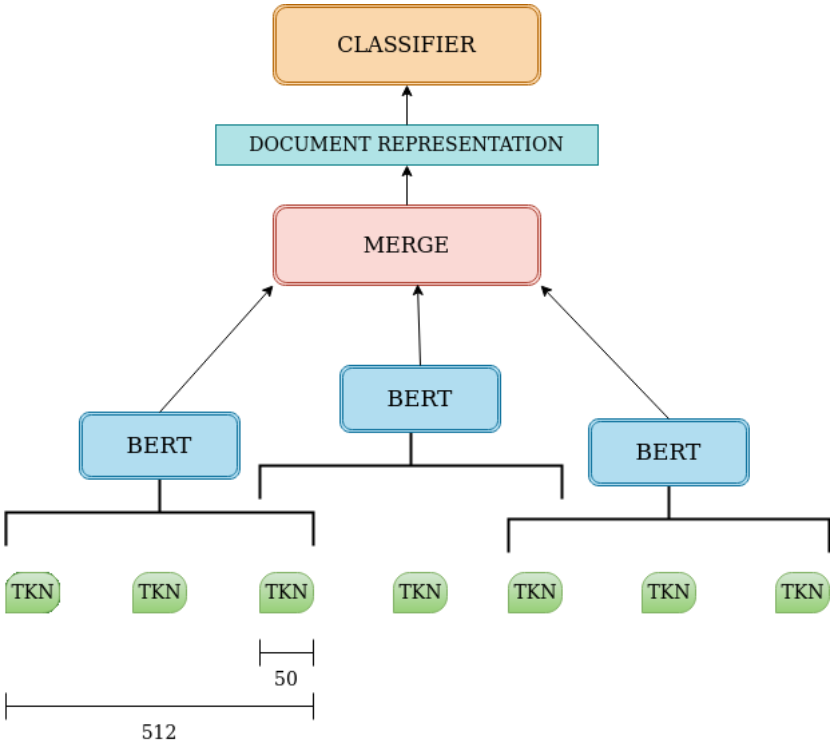


Figure 1. The document representation approach using a sliding window over the whole input sequence. Each subsequence is embedded using a fine-tuned BERT model, and all the subsequences are then merged into a final document representation, which is sent further as the input to the classifier. The length of the sliding window is 512 tokens. The first 50 tokens of each subsequent sliding window overlap with the last 50 tokens of the previous sliding window.

For each document x_i in the training set $S = \{x_1, x_2, \ldots, x_n\}$, we produced a document representation $d \in R^{1 \times t}$, where t is the dimension of the representation, by encoding the document with the mBERT model and taking the representation of the [CLS] token from the last layer. We sent this representation through a linear layer and a softmax function to map it to one of the predefined classes $C = \{y_1, y_2, \ldots, y_n\}$.

$$h = Linear(d, W) \tag{1}$$

$$\hat{y} = Softmax(h) \tag{2}$$

We calculated the loss of the sentiment classification task: $\mathcal{L}_s$ at the end using the negative log likelihood loss function

$$\mathcal{L}_s = -\log(\hat{y}_i)$$

where $\hat{y}_i$ is the probability of the correct class.
The final loss $\mathcal{L}$ is computed as:

$$\mathcal{L} = \mathcal{L}_{mlm} + \mathcal{L}_s$$

where $\mathcal{L}_{mlm}$ represents the loss from the masked language modeling task. The model is then jointly trained on both tasks by backpropagating the final loss through the whole network.

The original mBERT model is pre-trained on another task, namely next sentence prediction, which, according to the authors, helps the model learn sentence relationships. During training, the input for this task is treated as belonging to two separate sequences and the model has to decide if the two sequences follow one another in the original text or not. This information is useful for a variety of downstream tasks such as question answering. Since in this experiment we are dealing with a classification task, where the input is treated as being a part of the same sequence, we felt the additional training using the next sentence prediction task would not add much relevant information to the model so we omitted it in the intermediate training phase.

5. Experimental Setup

This sections describes the setup that we used to perform the experiments. It is divided into three subsections: the first subsection describes the regime we used for the fine-tuning phases; the second subsection describes the regime we used for the intermediate training phase; and the third subsection presents the evaluation of the trained models.

5.1. Fine-Tuning Phase

For the fine-tuning phase, we used the Slovenian news dataset [3] annotated on the document level (see Section 3), as the goal of our classification is to assign the sentiment label to a news article. We followed the suggestions in the original paper by Devlin et al. [10] for fine-tuning. We used the Adam optimizer with the learning rate of $2E-5$ and learning rate warmup over the first 10% of the training instances. For regularization purposes, we used the weight decay set to 0.01. We reduced the batch size from 32 to 16 due to the high memory consumption during training, which was the result of a long sequence length. For benchmarking purposes, we used the k-fold cross-validation training regime for the fine-tuning phase, where we split the dataset into k folds. In each cross-validation step, the k-1 folds are used as the training set, while the k-th fold is used as the testing set. The models in each cross-validation step were trained for 3 epochs. To avoid overfitting, we split the training folds into smaller training and development sets. After each epoch, we measured the performance on the development set and saved the new model parameters only if the performance of the model on the development set increased. For the document representation methods, described in Sections 4.1 and 4.2, the fine-tuning of the language model and the training of the classification head were performed end-to-end, while for the methods, described in Section 4.3, the classification heads were trained after the fine-tuning phase was completed. Otherwise, the training regime and the chosen hyperparameters were the same for all the experiments.

5.2. Intermediate Training Phase Regime

For the intermediate training phase, we utilized the proposed modified modeling objectives, described in Section 4.4. We used the Slovenian news dataset with annotations on the paragraph level. The annotations on this level of granularity were used because we wanted to perform the intermediate training phase on a different dataset than the one used for fine-tuning, but containing information relevant for the document-level sentiment classification task.

Since the annotated paragraphs were part of the same documents we used for the fine-tuning step, we took measures to prevent any form of data leakage. As described in Section 5.1, the fine-tuning phase was performed using 10-fold cross-validation. We performed the intermediate training in each cross-validation step, but excluded the paragraphs that were part of the documents in the k-th testing fold of the fine-tuning step from the dataset. We split the remaining data into a training and development set and trained the language model for a maximum of five epochs. At the end of each epoch, we calculated the perplexity score of the model on the development set and saved the new weights only if perplexity improved in the previous epoch. If perplexity did not improve for three

consecutive epochs, we stopped the training early. For this phase, we used the same hyperparameter settings as for the fine-tuning phase.

5.3. Evaluation

All the models were first trained and evaluated on the Slovenian dataset using 10-fold cross-validation as described in Sections 5.1 and 5.2. Next, the performance of the models from each fold was additionally tested on the Croatian test set to check the performance in the zero-shot learning setting (i.e., without any Croatian data used in training). The performances from each fold on the Croatian test set were then averaged and reported as a final result. The results for this set of experiments are presented in Table 2. The performance of the models was summarized using a standard classification metric, namely the macro-averaged F1 score, which is the appropriate measure given the highly imbalanced nature of the dataset (dominant neutral class). For completeness, we also separately report the precision and recall, both macro-averaged over all classes. Additionally, we also report the average F1 score performance of the model on the Slovenian and Croatian test sets. The performance of our models was compared to the baseline majority classifiers for both the Slovenian and Croatian datasets.

Table 2. Results of the document representation approaches. The first column shows the performance of models in the Slovenian 10-fold cross-validation setting; the second column is the average zero-shot performance on the Croatian test set; and the last column presents the average F1 score of the results on the Slovenian and Croatian datasets. Best results are marked in bold.

Model	Slovenian Cross-Validation			Croatian Test Set			Average
	Precision	Recall	F1	Precision	Recall	F1	F1
Majority classifier	17.34	33.33	22.76	0.20	0.33	25.00	/
Beginning of the document	65.45 ± 2.61	**62.83** ± 2.46	63.34 ± 2.29	57.74 ± 1.20	**53.91** ± 2.41	52.06 ± 2.64	57.70
Beginning and end of the document	64.72 ± 2.82	62.67 ± 2.69	63.33 ± 2.56	**59.00** ± 1.62	53.53 ± 3.64	**52.41** ± 2.58	**57.87**
Sequences from every part of the document							
Most informative subsequence	64.42 ± 2.44	62.09 ± 2.27	63.00 ± 2.34	57.87 ± 1.32	53.23 ± 2.82	52.30 ± 2.86	57.65
Averaging subsequence representations	**66.50** ± 3.13	62.00 ± 2.45	**63.39** ± 2.42	57.53 ± 1.14	52.95 ± 3.38	51.55 ± 3.93	57.47
1D CNN	63.96 ± 10.02	60.91 ± 5.22	61.58 ± 7.78	54.96 ± 5.48	53.31 ± 3.62	50.28 ± 4.65	55.93

6. Results

This section presents the results of the experiments conducted in the course of this study. We first present the results of the document representation approaches. The results are presented in Table 2. Next, for the best performing representation approach, we test our newly introduced technique for sentiment classification with intermediate training, and the results with and without the intermediate training objective are compared in Table 3. We also compare our results with the previous sate-of-the-art SVM and Naive Bayes models on the Slovenian dataset from [3], as well as with the neural network model based on LSTMs and TF-IDF from [39]. We note, however, that the testing regime in these experiments was not the same. In [3], the authors tested their models using five times 10-fold cross-validation, while in [39], the model was trained and tested on a random train-test split of the whole dataset with an 80:20 train-test split ratio. For this reason, the results are not directly comparable.

Table 3. Performance of the model using our intermediate sentiment classification training approach compared to the model without intermediate training. Additionally, we include the reported results from the related work using the same dataset. Best results are marked in bold.

Model	**Slovenian**			**Croatian**			**Average**
	Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**	**F1**
Majority classifier	17.34	33.33	22.76	0.20	0.33	25.00	/
Reported results from related studies							
SVM (from Bučar et al. [3]) 5 × 10 CV	/	/	63.42 ± 1.96	/	/	/	/
NBM(from Bučar et al. [3]) 5 × 10 CV	/	/	65.97 ± 1.70	/	/	/	/
LSTM+TF-IDF (from Pelicon [39])train-set split	/	/	62.5	/	/	/	/
Results from the current study							
Beginning of the document	65.45 ± 2.61	62.83 ± 2.46	63.34 ± 2.29	**57.74** ± 1.20	53.91 ± 2.41	52.06 ± 2.64	57.70
Beginning and end of the document with sentiment intermediate training	**67.19** ± 2.67	**66.00** ± 3.00	**66.33** ± 2.60	56.32 ± 1.88	**54.90** ± 2.36	**54.77** ± 1.39	**60.55**

As shown in Table 2, all the models using one of the tested document representation methods in this experiment performed better than the majority baseline classifier by a substantial margin. The best performing model on the Slovenian dataset (in terms of F1 score) utilizes document representations formed by simple averaging of the subsequence representations. The different document representation methods that were tested in this work do not seem to impact the model performance much as the performances of all our models differed only by a small margin when tested on the Slovenian data.

As far as absolute performance, we can see that the tested methods achieved F1 scores in the sixties for this particular Slovenian dataset with the best F1 score of 63.39 with averaging subsequence representations. When these models were tested on the Croatian test set in a zero-shot setting, the performance additionally dropped for approximately 11% with best the F1 scores achieving the low fifties. The best performing representation on the Croatian dataset uses the beginning and end of the document. Interestingly, the best performing model on the Slovenian dataset also saw the highest drop on the Croatian dataset of 11.84%. We additionally observed high variance of the CNN model compared to the other models.

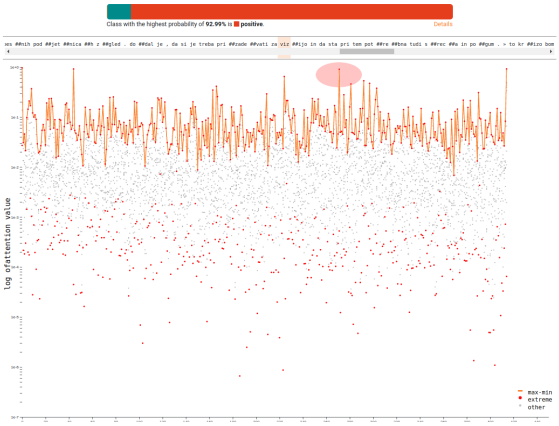
Since the three best performing document representation techniques were within a 0.06% difference on the Slovenian dataset, for experiments with intermediate training for sentiment enrichment, we opted for the document representation that used the beginning and ending of the input document as its average performance on the test sets of both Slovenian and Croatian languages was the highest. The results for the intermediate training experiment (Table 3) show that the model with the additional intermediate training step outperforms the model without the intermediate training step when using the same document representation technique. The results show three points better average performance on the Slovenian dataset and 2.68 points average improvement on the Croatian dataset in terms of the F1 score. Our model also manages to outperform the previous state-of-the-art models on the Slovenian dataset, achieving a 0.36 point increase in terms of F1 score, however this should be taken with precaution as the two evaluation settings differ.

7. Qualitative Exploration of the Models: Behavior of the Attention Space

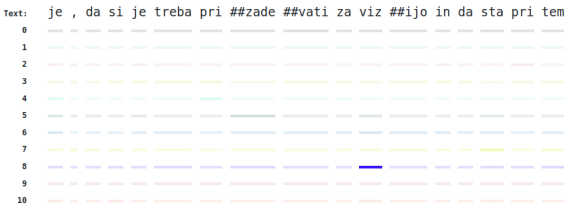
With the increasing use of neural language models, in recent years, the methodology aimed at the exploration of the human-understandable patterns, emerging from trained models, has gained notable attention. Models, such as BERT [40] and similar ones, can consist of hundreds of millions of parameters, which carry little useful information in terms of studying which parts of the model input were of relevance when making a prediction. To remedy this shortcoming, visualization methodologies are actively developed and researched for the task of better understanding the associations between the input token space and the constructed predictions.

The existing toolkits that offer the exploration of attention have been actively developed in recent years [41,42] and are widely used to better understand a given model's behavior. In this section, we exploit the recently introduced, freely available AttViz [43], an online toolkit for the exploration of the self-attention space of trained classifiers (http://attviz.ijs.si/). The tool is used to explore the behavior of the self-attention when considering positive, negative, and neutral classifications. The original tool was developed for instance-based exploration. In addition, we introduce a novel functionality of the tool aimed at the analysis of global attention values (per class analysis on the token collection level).

In the remainder of this section, we fist present a collection of selected examples, offering insight into the trained model's behavior. We begin by discussing selected positive instances, followed by neutral and negative ones. All the visualizations were done with the sentiment-enriched model that we trained in the course of this study. The main aim of this section is to explore the currently available means of inspecting trained neural language models. A positive example is shown in Figure 2.



(**a**) Sequence view.



(**b**) Token view.

Figure 2. Positive Example No. 41. The red ellipse (**a**) highlights one of the tokens (byte pairs) with the highest (normalized) self-attention—the token is part of the word "vizija" (translation: vision) (**b**). Note also the peaks at the beginning and the end; these peaks refer to the special tokens (e.g., [CLS]]).

The positive example was selected as it has a very high probability of being positive class and it showcases two main patterns that can be observed throughout the space of positively classified examples: first, only a handful of tokens are emphasized (if any), and second, there appears to be strong bias towards the first and the last token, indicating the potential effect of pre-training.

Next, we considered some of the examples classified as negative sentiment (see the example in Figures 3 and 4).

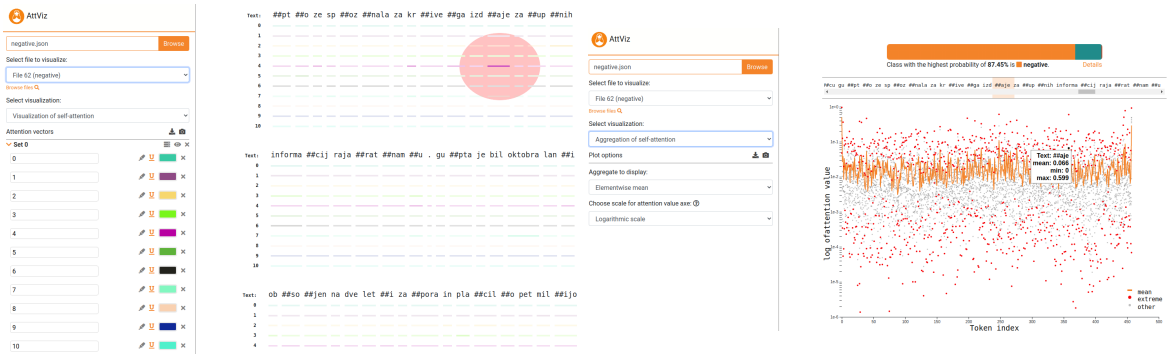


Figure 3. Negative Example No. 62. In this example, one of the highest attention values was around the token "izdaje" (translation: treason), which could be one of the carriers of the negative sentiment. Note that individual lines represent attention values for each of the ten attention heads. The document was classified with 87.45% probability.

The attention (highlighted red circle) peaks at the discussed token (translated as treason and negotiations respectively) can be observed, indicating that the neural language model picked up a signal at the token level during the association of the byte-paired inputs with outputs. Furthermore, we observed a similar pattern related to the starting [CLS] token, as well as the ending [SEP] token, i.e., token defining the end of the sentence. The pattern was consistent also throughout the neutral examples.

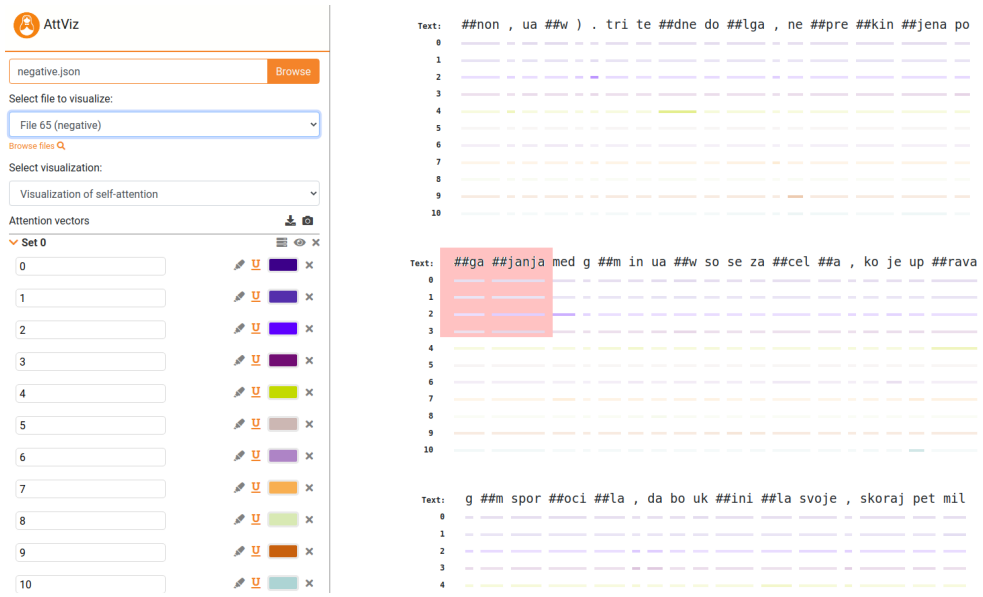


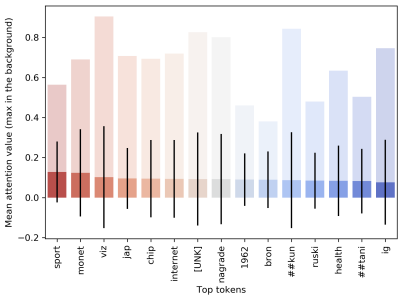
Figure 4. Negative Example No. 65. The highlighted region (red) corresponds to the term "pogajanja" (translation: negotiations), which appears to be associated with the classification of the observed text into the negative class.

The considered attention spaces offered insight into two main aspects of the trained model. First, the self-attention space, i.e., the space of the attention values alongside the attention matrix diagonals, offers relatively little insight into what the model learned. There are at least two main reasons for the observed behavior, as it appears to deviate from the reported explanations [43]. First, the considered documents are relatively long. Such documents give rise to a higher spread of the self-attention, smoothing out the individual peaks. Second, the wider spread of the attention could also be to the morphology-rich language considered (Slovene).

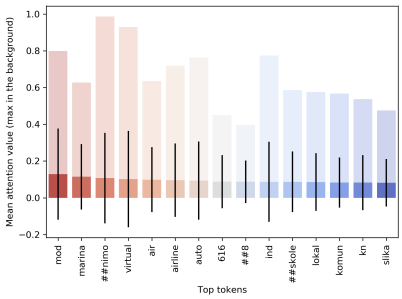
We next discuss the behavior of the global attention values both at the token, as well as the distribution level. The top 15 tokens according to the mean attention values are shown in Figure 5.

The presented results confirm the initial finding (e.g., Figure 2) that most of the attention space has high variability and, as such, does not directly offer interpretable insights; however, some meaningful results are also observed, e.g., the token with the greatest attention value for the positive class is sport. The final analysis we conducted was at the level of the global attention distributions. Here, we plotted the kernel density estimates of raw attention values across different types of instances. The results are shown in Figure 6.

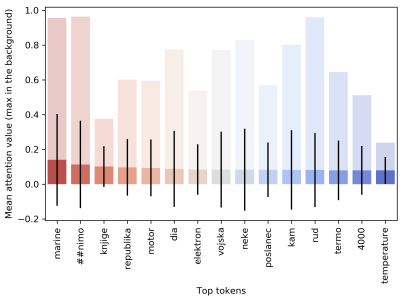
The distribution visualization indicates that the main differences emerge when considering the minimum value, a given token ever achieved; this result, albeit unexpected, potentially indicates that the attention is for classification of negative texts focused on a more particular subset of tokens, yielding a lower average subject to a skewed distribution. We finally offer quantile-quantile plots in Figure 7.



(**a**) Positive sentiment.

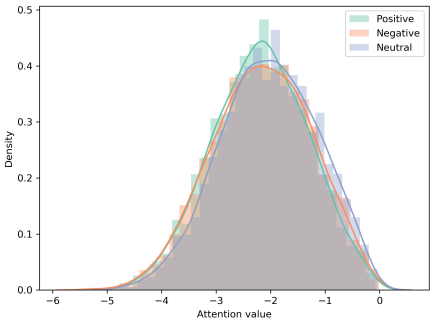


(**b**) Negative sentiment.

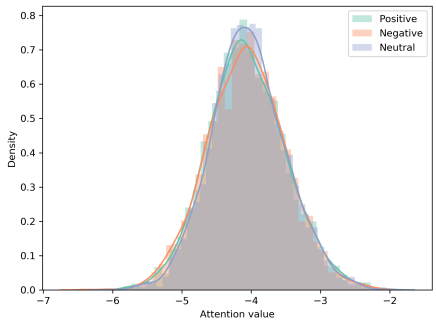


(**c**) Neutral sentiment.

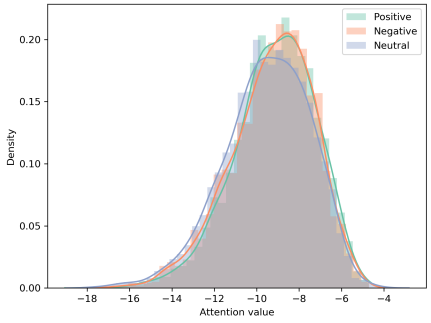
Figure 5. Visualization of token level attention. The figures represent the top 15 tokens according to the mean attention values. In the background, the maximum attention for a given token is also plotted. Note that the high standard deviation indicates little emphasis on the individual tokens.



(**a**) Maximum attention per token.

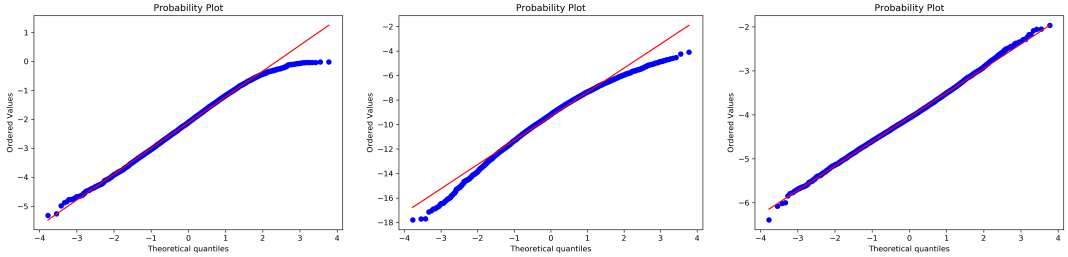


(**b**) Mean attention per token.



(**c**) Minimum attention per token.

Figure 6. Visualization of attention (log-transformed) distributions. It can be observed that the largest differences emerge when considering minimum attention. There, the negative texts' distribution is the most skewed. When considering maximum and mean distributions, however, no notable differences emerge.



(**a**) QQ plot: max attention. (**b**) QQ plot: min attention. (**c**) QQ plot: mean attention.

Figure 7. The quantile-quantile fits of the three considered attention distributions. It can be observed that the min and max attention distributions are skewed, indicating the presence of more extreme values.

The considered QQ-plots further confirm the observation that the skewed distribution of attention can be observed when considering min-max values; however, on average, the log transform could be interpreted to behave as a normal distribution; however, additional tests, such as Pearson's sample skewness (computed as $\frac{n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^3}{(n^{-1}\sum_{i=1}^{n}(x_i-\bar{x})^2)^{3/2}}$, where x_i is the i-th value out of n samples) could be conducted to further quantify the attention behavior.

8. Availability

The croatian news dataset with document-level sentiment annotations is available on the CLARIN repository under the Creative Commons license (CC-BY-NC-ND) (http://hdl.handle.net/11356/1342). The code for all the experiments is available on GitHub (https://github.com/PeliconA/crosslingual_news_sentiment.git).

9. Conclusions and Future Work

In this work, we addressed the task of sentiment analysis in news articles performed in a zero-shot cross-lingual setting. The goal was to successfully train models that could, when trained on data in one language, perform adequately also on data in another language. For this purpose, we used publicly available data of Slovenian news manually labeled for sentiment to train our models. Additionally, we gathered a new dataset of Croatian news and labeled it according to the guidelines for the annotation of the Slovenian dataset. This new dataset served as a test set for the zero-shot cross-lingual performance of our models.

We based our models on the multilingual Transformer-based model BERT, which has shown remarkable multilingual and cross-lingual performance. We however identified two potential drawbacks with the BERT model. The input window of the BERT model is fixed and relatively short. A widespread approach to this limitation is to shorten the input before sending it to the model for processing. While this approach is adequate for shorter texts, with longer documents, like news articles, it may cause severe information loss. The second drawback is that while BERT is pre-trained on a large collection of data, the only explicit sentiment signal it gets is during the fine-tuning phase on a usually small collection of labeled data.

To remedy the first potential drawback, we first tested several techniques for producing more informative long document representations. The techniques, which were described in detail, were partially inspired by earlier work, but to the best of our knowledge, they have not yet been tested in a cross-lingual setting. Our results show that all the techniques outperform the majority baseline classifier by a large margin, even when applied to the Croatian test set in a zero-shot setting where the model is not fine-tuned on Croatian data.

For the second identified limitation of the BERT model, we proposed a novel intermediate learning phase that encompasses the masked language modeling task and sentiment classification task. This phase is performed before the fine-tuning phase using a training set with separate annotations. The goal of this

phase is to induce the sentiment-related information directly into the BERT representations before the fine-tuning begins on the target task data. Results show that after fine-tuning, the sentiment-enriched model outperforms the models without the intermediate training phase both on the Slovenian dataset and on the Croatian test set in a zero-shot setting. Additionally, it slightly outperforms the current state-of-the-art on the Slovenian dataset, as reported in [3].

In the future, we plan to further test our proposed intermediate sentiment-enrichment phase with masked language modeling and sentiment classification tasks. Currently, the fine-tuning and the intermediate training phases share the dataset, but use labels on different levels of granularity: we used document-level labels for fine-tuning and paragraph-level labels for intermediate training. We would like to test how using training data from a very different training set would impact the performance of the proposed intermediate training step. We will also test the general transferability of this phase. Given a large enough corpus of sentiment-labeled instances that can be used for the intermediate training step, we would like to see if a Transformer-based model enriched with our proposed method can work well on sentiment tasks in different target languages and from different domains. Another interesting research area would be using topic modeling as a supplementary method for the news-related sentiment classification task. Such research would also test the underlying assumption that there is a positive correlation between the topic of a news article and the sentiment that a news article evokes in the readers. Even though the news articles in the datasets used for this work are not explicitly labeled for topics, they nevertheless deal with varying content and could support such research.

Author Contributions: A.P. and S.P. designed the study and developed its methodology. M.P. and D.M. provided the data for the study and guided the annotation process. Formal analysis of the study was done by A.P. Software for the experiments was written by A.P. Visualization of the trained models was done by B.Š. Validation of the results and supervision of the study was done by S.P. A.P., M.P., D.M., B.Š. and S.P. cotributed to the writing, reviewing and editing of the paper. All authors have read and agreed to the published version of the manuscript.

Funding: This research is supported by the European Union's Horizon 2020 research and innovation program under Grant Agreement No. 825153, project EMBEDDIA (Cross-Lingual Embeddings for Less-Represented Languages in European News Media). The work of A.P. was funded also by the European Union's Rights, Equality and Citizenship Programme (2014–2020) project IMSyPP (Innovative Monitoring Systems and Prevention Policies of Online Hate Speech, Grant No. 875263). We acknowledge also the funding by the Slovenian Research Agency (ARRS) core research program Knowledge Technologies (P2-0103). The results of this publication reflect only the authors' views, and the Commission is not responsible for any use that may be made of the information it contains.

Acknowledgments: We would like to thank 24sata, especially Hrvoje Dorešić and Boris Trupčević, for making the data available. We thank Jože Bučar for leading the annotation process.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A. Details on Croatian Dataset Construction

For the selection of articles, the time period was specified: approximately half of the articles were selected from the period from 1 September 2007 to 31 December 2013 in order to match the Slovenian dataset, while the other half were recent articles from last five years. From the initial set of articles, short and medium length articles were kept, leading to final selection. The articles were then cleaned and preprocessed, and the quality was checked (automatically and manually). The final dataset consisted of 2025 news articles. The sentiment annotation task was performed on three levels: document, paragraph, and sentence level.

For the selection of annotators, the condition of being native speakers was imposed, and we also considered the candidate's interest in the task.

The annotator were trained in two phases:

- In the first phase, we introduced the project EMBEDDIA and its goals. A referee introduced the web application for the annotation task. The annotators received basic guidelines, which were explained to them in detail by a referee. This was followed by the annotation of five articles,

which were annotated together on the three levels (sentence, paragraph, and document level). Using a five-level Likert scale: [35] (1—very negative, 2—negative, 3—neutral, 4—positive, and 5—very positive), the annotators annotated each article according to the following question: "Did this news evoke very positive/positive/neutral/negative/very negative feelings? (Please specify the sentiment from the perspective of an average Croatian web user)". Together with a referee, they discussed the individual instances, every single decision, and the annotation grade and resolved possible issues and doubts.

- In the second phase, all annotators annotated the same 25 articles individually. Afterwards, we analyzed the results of the annotation. The agreement (Cronbach's alpha measure) between the annotators on the document level was 0.816, which was a very good achievement with only 25 articles. We planned to achieve a 0.8 threshold. If the annotators had not achieved the planned threshold, they would repeat the second phase until they achieved it. The instances with lower agreement were discussed, and the issues were resolved.

Since a satisfying inter-annotator agreement was reached, the rest of the 2000 were annotated by different numbers of annotators. They followed the instructions they were given in the first and second phases.

To evaluate the process of annotation, we explored correlation coefficients using various measures of inter-annotator agreement at three levels of granularity, as shown in Table A1. The first three internal consistency estimates of reliability for the scores, shown in Table A1, normally range between zero and one. The values closer to one indicate more agreement, when compared to the values closer to zero. Cronbach's alpha values indicated a very good internal consistency at all levels of granularity. Normally, we refer to a value greater than 0.8 as a good internal consistency and above 0.9 as an excellent one [44]. The value of Krippendorff's alpha [45] at the document level of granularity implied a fair reliability test, whereas its values at the paragraph level and sentence level were lower. Fleiss' kappa values illustrated a moderate agreement among the annotators at all levels of granularity. In general, a value between 0.41 and 0.60 implies a moderate agreement, above 0.61 a substantial agreement, and above 0.81 an almost perfect agreement [46]. Kendall's values indicated a fair level of agreement between the annotators at all levels of granularity. Correspondingly, the Pearson and Spearman values range from –1 to 1, where 1 refers to the total positive correlation, 0 to no correlation, and –1 to the total negative correlation. The coefficients showed moderate positive agreement among the annotators, but their values decreased when applied to the paragraph and the sentence level. Usually, the values above 0.3 refer to a weak correlation, above 0.5 to a moderate correlation, and above 0.7 to a strong correlation [47].

Table A1. Results of dataset annotation: level of inter-rater agreement for document, paragraph, and sentence levels.

	Document Level			**Paragraph Level**			**Sentence Level**		
a_c	0.927			0.888			0.881		
a_k	0.671			0.565			0.548		
k	0.527			0.489			0.441		
	min	*max*	*avg*	*min*	*max*	*avg*	*min*	*max*	*avg*
r_p	0.544	0.824	0.682	0.488	0.719	0.572	0.425	0.706	0.558
r_s	0.557	0.762	0.669	0.474	0.702	0.548	0.42	0.696	0.54
W	0.508	0.73	0.625	0.449	0.656	0.513	0.389	0.649	0.504

Our results support the claim by [48] that it can be more difficult to accurately annotate sentences (or even phrases). In general, the sentiment scores by different annotators were more consistent at the document level than at the paragraph and sentence level.

The final sentiment of an instance is defined as the average of the sentiment scores given by the different annotators (as in the Slovenian news set). An instance was labeled as:

- negative, if the average of given scores was less than or equal to 2.4,
- neutral, if the average of given scores was between 2.4 and 3.6,
- positive, if the average of given scores was greater than or equal to 3.6.

References

1. Beigi, G.; Hu, X.; Maciejewski, R.; Liu, H. An overview of sentiment analysis in social media and its applications in disaster relief. In *Sentiment Analysis and Ontology Engineering*; Springer: Cham, Switzerland, 2016; pp. 313–340.
2. Mejova, Y. *Sentiment Analysis: An Overview*; University of Iowa, Computer Science Department: Iowa City, IA, USA, 2009.
3. Bučar, J.; Žnidaršič, M.; Povh, J. Annotated news corpora and a lexicon for sentiment analysis in Slovene. *Lang. Resour. Eval.* **2018**, *52*, 895–919. [CrossRef]
4. Liu, B. Sentiment Analysis and Opinion Mining. *Synth. Lect. Hum. Lang. Technol.* **2012**, *5*, 1–167. [CrossRef]
5. Van de Kauter, M.; Breesch, D.; Hoste, V. Fine-Grained Analysis of Explicit and Implicit Sentiment in Financial News Articles. *Expert Syst. Appl.* **2015**, *42*, 4999–5010. [CrossRef]
6. Bhutani, B.; Rastogi, N.; Sehgal, P.; Purwar, A. Fake news detection using sentiment analysis. In Proceedings of the IEEE 2019 Twelfth International Conference on Contemporary Computing (IC3), Noida, India, 8–10 August 2019; pp. 1–5.
7. El Ali, A.; Stratmann, T.C.; Park, S.; Schöning, J.; Heuten, W.; Boll, S.C. Measuring, understanding, and classifying news media sympathy on twitter after crisis events. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, Montreal, QC, Canada, 21–26 April 2018; pp. 1–13.
8. Rambaccussing, D.; Kwiatkowski, A. Forecasting with news sentiment: Evidence with UK newspapers. *Int. J. Forecast.* **2020**. [CrossRef]
9. Bowden, J.; Kwiatkowski, A.; Rambaccussing, D. Economy through a lens: Distortions of policy coverage in UK national newspapers. *J. Comp. Econ.* **2019**, *47*, 881–906. [CrossRef]
10. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv* **2018**, arXiv:1810.04805.
11. Schwartz, B. Google's Latest Search Algorithm to Better Understand Natural Language. Search Engine Land. 25 October 2019. Available online: https://searchengineland.com/welcome-bert-google-artificial-intelligence-for-understanding-search-queries-323976 (accessed one 28 August 2020).
12. Albarino, S. Does Google's BERT Matter in Machine Translation? Slator. 17 October 2019. Available online: https://slator.com/machine-translation/does-googles-bert-matter-in-machine-translation/ (accessed one 28 August 2020).
13. Pires, T.; Schlinger, E.; Garrette, D. How multilingual is Multilingual BERT? *arXiv* **2019**, arXiv:1906.01502.
14. Karthikeyan, K.; Wang, Z.; Mayhew, S.; Roth, D. Cross-lingual ability of multilingual bert: An empirical study. In Proceedings of the International Conference on Learning Representations, Scottsdale, AZ, USA, 2–4 May 2019.
15. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Brew, J. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv* **2019**, arXiv:1910.03771.
16. Xie, Q.; Dai, Z.; Hovy, E.; Luong, M.T.; Le, Q.V. Unsupervised data augmentation for consistency training. *arXiv* **2019**, arXiv:1904.12848.
17. Lin, K.Y.; Yang, C.; Chen, H.H. Emotion Classification of Online News Articles from the Reader's Perspective. In Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Sydney, Australia, 9–12 December 2009; Volume 1, pp. 220–226. [CrossRef]
18. Li, X.; Xie, H.; Chen, L.; Wang, J.; Deng, X. News Impact on Stock Price Return via Sentiment Analysis. *Knowl. Based Syst.* **2014**, *69*. [CrossRef]

19. Mansar, Y.; Gatti, L.; Ferradans, S.; Guerini, M.; Staiano, J. Fortia-FBK at SemEval-2017 Task 5: Bullish or Bearish? Inferring Sentiment towards Brands from Financial News Headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 817–822. [CrossRef]
20. Moore, A.; Rayson, P. Lancaster A at SemEval-2017 Task 5: Evaluation metrics matter: predicting sentiment from financial news headlines. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*; Association for Computational Linguistics: Vancouver, BC, Canada, 2017; pp. 581–585. [CrossRef]
21. Wan, X. Co-training for cross-lingual sentiment classification. In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, Singapore, 2–7 August 2009; pp. 235–243.
22. Guzmán, F.; Chen, P.J.; Ott, M.; Pino, J.; Lample, G.; Koehn, P.; Chaudhary, V.; Ranzato, M. The FLoRes evaluation datasets for low-resource machine translation: Nepali-english and sinhala-english. *arXiv* **2019**, arXiv:1902.01382.
23. Zhou, X.; Wan, X.; Xiao, J. Attention-based LSTM network for cross-lingual sentiment classification. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Austin, TX, USA, 1–5 November 2016; pp. 247–256.
24. Chen, Z.; Shen, S.; Hu, Z.; Lu, X.; Mei, Q.; Liu, X. Emoji-powered representation learning for cross-lingual sentiment classification. In Proceedings of the World Wide Web Conference, San Francisco, CA, USA, 13–17 May 2019; pp. 251–262.
25. Funaki, R.; Nakayama, H. Image-mediated learning for zero-shot cross-lingual document retrieval. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Lisbon, Portugal, 17–21 September 2015; pp. 585–590.
26. Rei, M.; Søgaard, A. Zero-shot sequence labeling: Transferring knowledge from sentences to tokens. *arXiv* **2018**, arXiv:1805.02214.
27. Wang, Y.; Che, W.; Guo, J.; Liu, Y.; Liu, T. Cross-lingual BERT transformation for zero-shot dependency parsing. *arXiv* **2019**, arXiv:1909.06775
28. Hsu, T.Y.; Liu, C.L.; Lee, H.Y. Zero-shot Reading Comprehension by Cross-lingual Transfer Learning with Multi-lingual Language Representation Model. *arXiv* **2019**, arXiv:1909.09587
29. Ye, Z.; Geng, Y.; Chen, J.; Chen, J.; Xu, X.; Zheng, S.; Wang, F.; Zhang, J.; Chen, H. Zero-shot Text Classification via Reinforced Self-training. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 3014–3024.
30. Jebbara, S.; Cimiano, P. Zero-Shot Cross-Lingual Opinion Target Extraction. *arXiv* **2019**, arXiv:1904.09122
31. Fei, H.; Li, P. Cross-Lingual Unsupervised Sentiment Classification with Multi-View Transfer Learning. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Seattle, WA, USA, 5–10 July 2020; pp. 5759–5771.
32. Tan, X.; Yan, R.; Tao, C.; Wu, M. Classification over Clustering: Augmenting Text Representation with Clusters Helps! In *Proceedings of the CCF International Conference on Natural Language Processing and Chinese Computing*; Springer: Cham, Switzerland, 2019; pp. 28–40.
33. Pappagari, R.; Zelasko, P.; Villalba, J.; Carmiel, Y.; Dehak, N. Hierarchical Transformers for Long Document Classification. In Proceedings of the 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), Singapore, 14–18 December 2019; pp. 838–844.
34. Beltagy, I.; Peters, M.E.; Cohan, A. Longformer: The long-document transformer. *arXiv* **2020**, arXiv:2004.05150.
35. Likert, R. A Technique for the Measurement of Attitudes. *Arch. Psychol.* **1932**, *140*, 5–55.
36. Wu, Y.; Schuster, M.; Chen, Z.; Le, Q.V.; Norouzi, M.; Macherey, W.; Krikun, M.; Cao, Y.; Gao, Q.; Macherey, K.; et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv* **2016**, arXiv:1609.08144
37. Pruksachatkun, Y.; Phang, J.; Liu, H.; Htut, P.M.; Zhang, X.; Pang, R.Y.; Vania, C.; Kann, K.; Bowman, S.R. Intermediate-Task Transfer Learning with Pretrained Models for Natural Language Understanding: When and Why Does It Work? *arXiv* **2020**, arXiv:2005.00628.

38. He, C.; Chen, S.; Huang, S.; Zhang, J.; Song, X. Using Convolutional Neural Network with BERT for Intent Determination. In Proceedings of the IEEE 2019 International Conference on Asian Language Processing (IALP), Shanghai, China, 15–17 November 2019; pp. 65–70.
39. Pelicon, A. Zaznavanje sentimenta v novicah z globokimi nevronskimi mrežami. In Proceedings of the Conference on Language Technologies and Digital Humanities 2020 (to appear), Ljubljana, Slovenia, 17–20 March 2020.
40. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 5998–6008.
41. Vig, J. Visualizing Attention in Transformer-Based Language Representation Models. *arXiv* **2019**, arXiv:1904.02679.
42. Vig, J.; Belinkov, Y. Analyzing the structure of attention in a transformer language model. *arXiv* **2019**, arXiv:1906.04284.
43. Škrlj, B.; Eržen, N.; Sheehan, S.; Luz, S.; Robnik-Šikonja, M.; Pollak, S. AttViz: Online exploration of self-attention for transparent neural language modeling. *arXiv* **2020**, arXiv:2005.05716.
44. George, D.; Mallery, P. *SPSS for Windows Step-by-Step: A Simple Guide and Reference, 14.0 Update*, 7th ed.; Allyn and Bacon, Inc.: Boston, MA, USA, 2006.
45. Krippendorff, K. *Content Analysis: An Introduction to Its Methodology*, 2nd ed.; Sage Publications: Thousand Oaks, CA, USA, 2004.
46. Landis, J.R.; Koch, G.G. The Measurement of Observer Agreement for Categorical Data. *Biometrics* **1977**, *33*, 159–174. [CrossRef] [PubMed]
47. Rumsey, D.J.; Unger, D. *U Can: Statistics for Dummies*; John Wiley: Hoboken, NJ, USA, 2015.
48. O'Hare, N.; Davy, M.; Bermingham, A.; Ferguson, P.; Sheridan, P.; Gurrin, C.; Smeaton, A. Topic-dependent sentiment analysis of financial blogs. In Proceedings of the TSA 2009—1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement, Hong Kong, China, 6 November 2009; TSA: Arlington County, VA, USA, 2009; pp. 9–16, ISBN 978-1-60558-805-6.