Target Two Birds With One STONE: Entity-Level Sentiment and Tone Analysis in Croatian News Headlines

Ana Barić[1†] **Laura Majer**[1†] **David Dukić**[1†] **Marijana Grbeša**[2] **Jan Šnajder**[1]
[1]University of Zagreb, Faculty of Electrical Engineering and Computing, TakeLab
{ana.baric, laura.majer, david.dukic, jan.snajder}@fer.hr
[2]University of Zagreb, Faculty of Political Science
marijana.grbesa@fpzg.hr

Abstract

Sentiment analysis is often used to examine how different actors are portrayed in the media, and analysis of news headlines is of particular interest due to their attention-grabbing role. We address the task of entity-level sentiment analysis from Croatian news headlines. We frame the task as targeted sentiment analysis (TSA), explicitly differentiating between sentiment toward a named entity and the overall tone of the headline. We describe STONE, a new dataset for this task with sentiment and tone labels. We implement several neural benchmark models, utilizing single- and multi-task training, and show that TSA can benefit from tone information. Finally, we gauge the difficulty of this task by leveraging dataset cartography.

1 Introduction

Sentiment analysis (SA) is a common method in media and communication studies used to examine how different topics, events, or actors are portrayed in the media. It has been used to address media bias, framing, agenda setting, priming, and negativity in the news; e.g., (Semetko and Valkenburg, 2000; Hopmann et al., 2010; Grbeša, 2012; Galpin and Trenz, 2019). While SA often involves entire news reports, news headlines emerge as a relevant unit of analysis as they play a central role in grabbing the attention of audiences in the digital environment (Kuiken et al., 2017; Scacco and Muddiman, 2020), optimize the relevance of the story for the headline scanning audience (Dor, 2003), and also act as a strong framing mechanism (De Vreese, 2005; Tankard Jr, 2001).

Although sentiment is often used interchangeably with tone, valence, affect, or polarity (cf. Boukes et al. (2020); Soroka et al. (2015)), here we consider sentiment and tone as distinct concepts. Sentiment is operationalized as a category that is always determined in relation to a particular entity, whereas tone is more general and captures the overall mood and polarity of the entire news story or another unit of analysis. This conceptualization draws on the distinction made by Lengauer et al. (2012) between the "actor-related dimension of negativity" and the "frame-related dimension of negativity". The former dimension corresponds to sentiment, while the latter corresponds to tone.

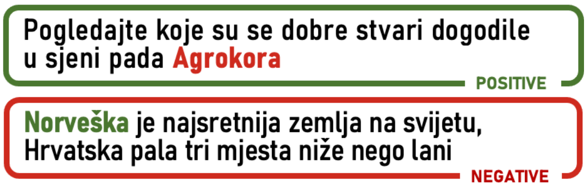

Pogledajte koje su se dobre stvari dogodile u sjeni pada Agrokora
POSITIVE
Norveška je najsretnija zemlja na svijetu, Hrvatska pala tri mjesta niže nego lani
NEGATIVE

Figure 1: Examples of headlines from the STONE dataset with contrasting tone and entity-level sentiment (red: negative, green: positive).

SA also has a long history in natural language processing (NLP) (Pang and Lee, 2008). The typical applications range from large-scale market research, product review analysis, and customer satisfaction estimation to voter profiling in political campaigns and media analysis. Typically, SA aims to determine the overall sentiment expressed in the text, thus corresponding to the "frame-related dimension of negativity". Often, this boils down to determining the sentiment *polarity* as either negative, neutral, or positive. In contrast, targeted sentiment analysis (TSA; Pei et al. (2019)) focuses on sentiment expressed toward specific targets. Specifically, entity-level sentiment analysis may be operationalized as TSA with the pre-extracted named entity (NE) mentions as targets, thus corresponding to the "actor-related dimension of negativity".

This paper addresses the TSA task for the Croatian language, more precisely, entity-level sentiment analysis from Croatian news headlines. To this end, we first propose a novel dataset for this task called STONE (Sentiment and TOne from NEws), with

[†] Equal contribution.

Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023), pages 78–85
May 6, 2023 ©2023 Association for Computational Linguistics

manually annotated entity-level sentiment and tone for news headlines. We investigate the relationship between targeted sentiment and tone, showing that there is a statistical dependence between the two. We then introduce and evaluate several neural benchmark models for this task. Building on our finding that targeted sentiment depends on the tone of a headline, we examine whether multitask modeling of TSA and tone can improve TSA prediction performance. Finally, to gauge the difficulty of the TSA on Croatian news headlines, we diagnose the dataset using the cartography technique of Swayamdipta et al. (2020), examining the relationship between annotator agreement and model correctness. The results show that, while TSA on Croatian news headlines is challenging for humans and state-of-the-art models, our benchmark models considerably outperform the baseline. We also show that using the tone signal in a multi-task setup can improve TSA performance further.

Our contribution is threefold: (1) a novel dataset for the task of entity-level sentiment and tone analysis in Croatian news headlines, which we make publicly available,[1] (2) neural benchmark models for this task in single- and multi-task setups, and (3) dataset diagnostics by means of dataset cartography. Our work brings valuable insights for TSA in the Slavic languages niche.

2 Related Work

The explicit emphasis on the target entity in TSA requires modeling the relationship between targets and their surrounding context. Previous work captured this target-context interaction by isolating the target entity with BIO tags and recurrent models (Hu et al., 2019; Li et al., 2019) or utilizing the attention mechanism (Zhang et al., 2016; Song et al., 2019). Another approach leveraged transformer-based models such as BERT for TSA tasks on user reviews (Gao et al., 2019; Rietzler et al., 2019) and Twitter data (Mutlu and Özgür, 2022) by focusing on target tokens for sentiment classification. In our work, we employ the target entity extraction method for the BERT-based model, but we simplify the extraction by using only target embeddings. Similar BERT-based approaches were adapted for the targeted sentiment analysis in the news domain on a sentence (Hamborg et al., 2021) and headline (Salgueiro et al., 2022) level for English and Spanish language, respectively.

In the realm of NLP for Slavic languages, our work is similar to that of (Pelicon et al., 2020), who annotated news articles for Slovene and Croatian language and performed sentiment classification of news articles on three levels of granularity (document, paragraph, and sentence level). However, they did not analyze sentiment toward specific named entities. Our work is most similar to (Baraniak and Sydow, 2021), who annotated and analyzed the dataset of news headlines for sentiment analysis toward target entities in English and Polish. We adopt a similar dataset design for annotating sentiment toward the target entity in news headlines for the Croatian language, but we also consider the general tone of the headline.

3 STONE Dataset

Our main contribution is STONE, a dataset containing headlines from Croatian news outlets labeled with sentiment towards NEs and the general tone of the headline. To compile the dataset, we first sampled the headlines from a database of news articles acquired by TakeLab Retriever,[2] a tool for analyzing Croatian online news media. To identify the NEs in the headlines, we ran the BERTić model fine-tuned for the task of NE recognition, which achieves an average F1-score of 89.21 on Croatian news hr500k dataset (Ljubešić and Lauc, 2021; Ljubešić et al., 2016). We retained only the headlines that contained at least one NE. If a headline contained several NEs, we randomly picked one as the target. Consequently, a headline may appear in our dataset several times with a different target.

We relied on a simple ternary annotation scheme, using the negative (NEG), neutral (NTR), and positive (POS) labels for both targeted sentiment and tone. While we considered more fine-grained schemes, such as the one proposed by Batanović et al. (2016), we decided to use the ternary one as it proved to be sufficient in preliminary annotation rounds. This aligns with our intuition that, as one of the primary purposes of news headlines is to draw attention and generate interest, sentiment and tone labels should capture the reader's immediate first impression rather than the result of a conscious and deliberate evaluation process.

The annotation was carried out by ten annotators using the Alanno tool for annotation management (Jukić et al., 2022). All annotators were native speakers of the Croatian language. The annota-

[1] https://takelab.fer.hr/retriever/stone.tar.gz

[2] https://retriever.takelab.fer.hr/

	Sentiment		
Tone	NEG	NTR	POS
NEG	3231	3024	524
NTR	344	3541	689
POS	251	1699	3827

Table 1: Contingency table of tone and sentiment judgments for the ten annotators.

tors worked independently, with six annotators per instance to account for the highly subjective nature of the task. The data annotation process was completed within 14 person-hours. The text of the articles was not made available to the annotators, only the headline. For each instance, the annotators labeled both the tone and targeted sentiment but were advised first to determine the tone of the headline and then the targeted sentiment, assuming this order – going from general to more specific – would make annotating easier.

The annotators were instructed to judge the tone and the sentiment based on their immediate impression of the headline. The guidelines further instructed them to consider the presence of epithets portraying the target entity or entire headline in a certain light and the context providing information about the nature of the event described in the headline. If none of these features were relevant, the annotators were told to rely on background knowledge of the subject in question. The annotators were also instructed to report erroneously identified NEs, and these headlines were discarded.

The final dataset contains 2855 headlines. For targeted sentiment, 1486 headlines were labeled as negative, 653 as neutral, and 716 as positive. Regarding the tone, 1262 headlines were labeled as negative, 666 as neutral, and 927 as positive. Inter-annotator agreement with the Fleiss-kappa metric is $\kappa = 0.416$ and $\kappa = 0.493$ for targeted sentiment and tone, respectively, which is considered a moderate agreement (Landis and Koch, 1977). A moderate level of agreement is expected, considering the highly subjective nature of these tasks.

Table 1 shows the contingency table of unaggregated sentiment and tone labels. Unsurprisingly, and as exemplified by Figure 1, targeted sentiment and tone are not always aligned, although in most cases they are. We used a chi-squared test to test the statistical dependence of targeted sentiment and tone. The two variables are significantly associated, with $\chi^2 = 8550.77, p < .01$.

4 Benchmark Models

The backbone of all our experiments was the BERTić model (Ljubešić and Lauc, 2021), based on Electra (Clark et al., 2020). Pre-trained BERTić achieves state-of-the-art performance on many NLP tasks in Slavic languages, including Croatian. We use BERTIć to produce contextualized representations of NEs in the headline for TSA.[3]

Gold Dataset. We compiled the gold dataset for evaluating benchmark models by aggregating the labels of the ten annotators for both sentiment and tone using a majority vote. To sidestep the problem of adjudicating labels in cases with no majority agreement, we removed all instances where there are ties in either sentiment or tone annotations. We leave alternatives, including adjudication steps, more fine-grained schemes, or label distribution prediction for future work. The so-obtained gold set contains 2307 instances, of which 508 are negative, 1151 are neutral, and 648 are positive sentiment instances. For tone, there are 428 negative, 1060 neutral, and 819 positive instances. We randomly split the gold set into training, validation, and testing in a 70:10:20 ratio.

Single-task Setup. We implemented one rudimentary baseline and four deep-learning benchmarks in the single-task setup for the TSA task. We use the univariate Bayes as a baseline, with class likelihood parameters estimated by computing the labels' distribution for lemmatized NEs appearing in the training set. Entities were lemmatized using CLASSLA (Ljubešić and Dobrovoljc, 2019). The intuition was that the sentiment of some NEs will be predictable regardless of the context they appear in. For out-of-vocabulary NEs, the prediction was made by sampling a label from the training set distribution conditioned on the NE type.

We implemented four benchmark models. In the first model (*Target*), the entire headline is fed to the model, followed by the extraction of only the target NE embeddings span. The second model (*Masked*) is fed with a headline where the target entity is replaced by a special [MASK] token. This tests the assumption that the targeted sentiment depends only on the context independently of the concrete entity. We experimented with including the NE type as a feature concatenated to the averaged embedding of an NE span before feeding it to the classification layer for both target (*Target+Type*)

[3] https://github.com/TakeLab/stone

and masked (*Masked+Type*) models. We fed the averaged contextualized embeddings of the NE span to the classification layer in each model.

Multi-task Setup. Annotation analysis in Section 3 revealed that there is a statistical dependency between tone and sentiment labels. We hypothesize that this dependency can be leveraged to obtain more accurate TSA predictions. To investigate this, we combined TSA and tone classification tasks into a multi-task training design (Zhang et al., 2022). The multi-task setup was implemented on top of BERTić with one linear classification head for tone classification and the other for targeted sentiment classification (*Target* single-task model). We implemented three multi-task setups. The first, *Alternate Batch* setup mimics the suggested procedure for dataset annotation, where, for each instance, the tone is annotated first, and the entity-level sentiment is annotated second. Within each mini-batch, we first present the model with tone instances, calculate the loss, and update the parameters based on the derivatives of the tone loss gradients. We then do the same for same-batch instances but this time for sentiment labels. We alternate between the two tasks during each epoch. This is in contrast to the second setup we considered, *Alternate Epoch*, where we first update model parameters depending on all tone training instances and then update the parameters based on all sentiment training instances using the appropriate classification head. Task-wise updates are conducted in a mini-batch fashion. The third setup is *Average Batch*, where we calculate the loss on a mini-batch level for both tasks and then update the model parameters based on the derivatives of the averaged batch loss.

Experimental Results. All results were obtained by averaging over five independent runs with different random seeds. BERTić-based benchmarks were trained for 10 epochs with a batch size of 16. We minimized cross-entropy loss and clipped gradients to 1.0. We used the AdamW optimization algorithm (Reddi et al., 2019) with a learning rate of 1e-5. The learning rate was adjusted with a linear learning rate scheduler that used zero warmup steps. We report macro-F1 scores and per-class F1 scores. TSA results are shown in Table 2 (corresponding results for tone are in A.2).

All neural models outperformed the Bayes baseline by a large margin. In a single-task setup, the *Target* model performed best, with *Target+Type* being the close second. Masked experiment results show that not knowing what the exact entity is or knowing only its NE type is detrimental to overall model performance, except for determining the neutral sentiment. *Average Batch* and *Alternate Batch* multi-task setups beat all single-task variants in terms of macro-averaged F1-score, with *Average Batch* reaching the highest score. This suggests that tone, incorporated through multi-task training, is beneficial for TSA model performance. Overall best negative and positive sentiment results were also obtained in multi-task setups. The performance on the neutral label was consistent across setups, presumably because the neutral instances make up the majority of the dataset.

Single-task	AVG	NEG	NTR	POS
Univariate Bayes	.214	.079	.079	.266
Target	.752	.738	.782	.737
Target+Type	.749	.737	**.787**	.723
Masked	.506	.393	.702	.422
Masked+Type	.589	.500	.720	.548
Multi-task				
Alternate Batch	.751	**.748**	.784	.720
Alternate Epoch	.755	.747	.779	.740
Average Batch	**.757**	.742	.779	**.749**

Table 2: TSA macro-averaged and per-class F1-scores for single-task (baseline and four models) and multi-task *Target* model. The best results by setup are in **bold**.

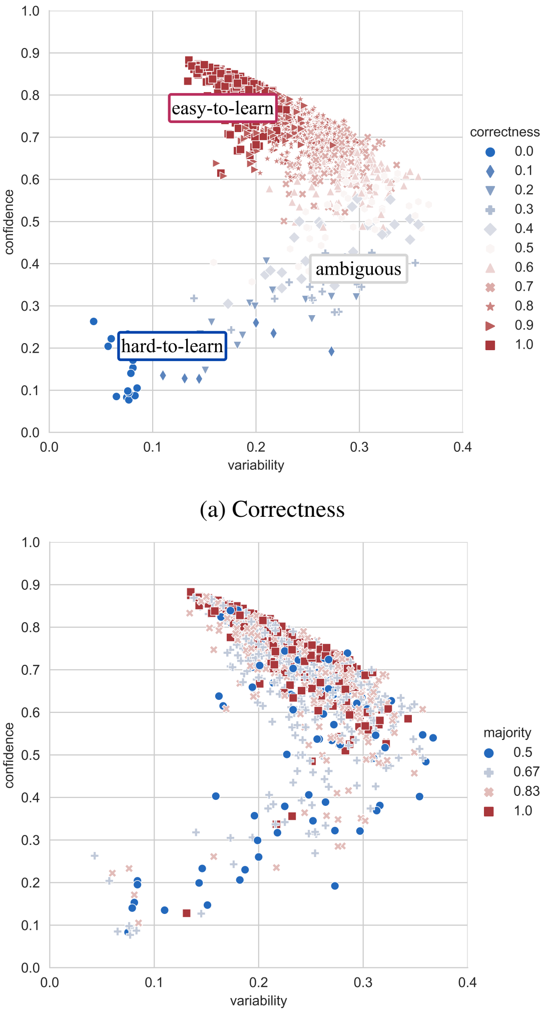
5 Dataset Diagnostics

The dataset cartography method (Swayamdipta et al., 2020) makes it possible to analyze the characteristics of a dataset in relation to model performance. The method uses three metrics – *confidence*, *variability*, and *correctness* – to measure the model's performance on the individual instances over training epochs. Instances may then be grouped into three regions reflecting their difficulty: *easy-to-learn* instances are of high confidence and low variability, *hard-to-learn* instances are of both low confidence and low variability, while other instances are considered *ambiguous*.

Figure 2a shows the cartography of the STONE dataset for the *Average Batch* model. The dataset exhibits patterns already observed for other NLP datasets. This visualization is especially useful for identifying hard-to-learn instances with critically low correctness. However, as noted by Swayamdipta et al. (2020), poor model perfor-

Category		Title	Gold label	Model prediction	Majority
LOCALITIES	1	Kaos u **Portugalu**: Policajci su pucali na nogometnoj utakmici (*Chaos in **Portugal**: Policemen fired shots at a football game*)	NEG	NTR	.67
	2	**Italija** bilježi stalni porast broja zaraženih, ali ministar zdravstva očekuje početak pada potkraj proljeća (***Italy** is recording a constant rise of infected, but the minister of health is expecting a decline by the start of spring*)	NEG	NTR	.67
QUOTES	3	**Bandić**: Informatika se u škole uvodi da bi se izbacio vjeronauk (***Bandić**: Informatics is being introduced in schools to cancel religious studies*)	NEG	NTR	.83
	4	**Michael Phelps**: Sad mi je tek jasno da sam bio jednaki šupak od čovjeka kao Jordan (***Michael Phelps**: It is only now clear to me that I was just as big of an asshole as Jordan*)	NEG	NTR	.50
INFERENCE	5	Posrtali su tamo gdje nisu smjeli! Ovo su utakmice koje su **Hajduk** koštale osvajanja naslova prvaka (*They failed where they shouldn't have! These are the matches which cost **Hajduk** the championship title*)	NEG	NTR	.67
	6	Kraljica u seksi kombinezonu: U ovom se vojvotkinja **Catherine** nikad ne bi pojavila (*Queen in a sexy overall: Dutchess **Catherine** would never be seen wearing this*)	NEG	NTR/POZ	.83

Table 3: Instances with low model performance, grouped into categories. The target entities are in **bold**. All instances have a correctness value ≤ .1, except example 3, which scored a correctness value of 1.



(a) Correctness

(b) Majority

Figure 2: STONE cartography with (a) correctness and (b) the majority metrics indicated with hue/shape.

mance may be due to ambiguity inherent to the instance rather than model limitations, and to distinguish between the two, it may be helpful to consider *human agreement* metric. We instead used the *majority* metric (the percentage of annotators that agreed on the gold label) to avoid the need to resolve ties for instances with no majority agreement stochastically. Figure 2b shows majority along with confidence and variability. Unlike in Figure 2a, one cannot identify prominent regions, suggesting there is no direct link between human consensus and the difficulty of an instance for the model.

Instead of looking at human consensus, to gain an insight into the phenomena the model seems to struggle with, we looked into instances with low correctness (≤ .1). Table 3 shows some examples. We preliminary identified three problematic categories of instances: (1) headlines with *localities* – the target entity refers to a location, and the sentiment is predominantly neutral, but in some cases the entity is a toponym that might be held responsible for the outcome. In this case, the negative evaluation may be transferred to the entity, which the model failed to infer; (2) headlines with *quotations* – the sentiment towards the speaker is usually neutral since no additional information is present, as shown in example 3. However, in example 4, the quote contains a negative observation about Phelps himself, which is atypical and failed to be recognized by the model; (3) evaluations based on *inference*, typical for headlines comprising multiple sentences. Instances such as examples 5 and 6 prove to be too difficult for the model while achieving sufficient consensus among the annotators.

6 Conclusion

We introduced a dataset for entity-level sentiment and tone analysis in Croatian news headlines. We tested neural benchmark models in a single- and multi-task setup, achieving the best results with representations of named entities and multi-task training. Dataset cartography identified several problematic cases for the model, which could be addressed in future work. Future work may also consider different framings of the TSA task.

Acknowledgments

This research was supported by the grant AIDWAS KK.01.2.1.02.0285. We acknowledge the insightful comments and suggestions provided by Irina Masnikosa, Nebojša Blanuša, and Mato Brautović.

References

Katarzyna Baraniak and Marcin Sydow. 2021. A dataset for sentiment analysis of entities in news headlines (SEN). *Procedia Computer Science*, 192:3627–3636.

Vuk Batanović, Boško Nikolić, and Milan Milosavljević. 2016. Reliable baselines for sentiment analysis in resource-limited languages: The Serbian movie review dataset. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2688–2696, Portorož, Slovenia. European Language Resources Association (ELRA).

Mark Boukes, Bob Van de Velde, Theo Araujo, and Rens Vliegenthart. 2020. What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, 14(2):83–104.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Claes H De Vreese. 2005. News framing: Theory and typology. *Information design journal+ document design*, 13(1):51–62.

Daniel Dor. 2003. On newspaper headlines as relevance optimizers. *Journal of pragmatics*, 35(5):695–721.

Charlotte Galpin and Hans-Jörg Trenz. 2019. Converging towards Euroscepticism? Negativity in news coverage during the 2014 European Parliament elections in Germany and the UK. *European Politics and Society*, 20(3):260–276.

Zhengjie Gao, Ao Feng, Xinyu Song, and Xi Wu. 2019. Target-dependent sentiment classification with BERT. *IEEE Access*, pages 154290–154299.

Marijana Grbeša. 2012. Framing of the president: Newspaper coverage of Milan Bandić and Ivo Josipović in the presidential elections in Croatia in 2010. *Politička misao: časopis za politologiju*, 49(5):89–113.

Felix Hamborg, Karsten Donnay, and Bela Gipp. 2021. Towards target-dependent sentiment classification in news articles. In *Diversity, Divergence, Dialogue: 16th International Conference, iConference 2021, Beijing, China, March 17–31, 2021, Proceedings, Part II 16*, pages 156–166. Springer.

David Nicolas Hopmann, Rens Vliegenthart, Claes De Vreese, and Erik Albæk. 2010. Effects of election news coverage: How visibility and tone influence party choice. *Political communication*, 27(4):389–405.

Minghao Hu, Yuxing Peng, Zhen Huang, Dongsheng Li, and Yiwei Lv. 2019. Open-domain targeted sentiment analysis via span-based extraction and classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 537–546, Florence, Italy. Association for Computational Linguistics.

Josip Jukić, Fran Jelenić, Miroslav Bićanić, and Jan Šnajder. 2022. Alanno: An active learning annotation system for mortals. *arXiv preprint arXiv:2211.06224*.

Jeffrey Kuiken, Anne Schuth, Martijn Spitters, and Maarten Marx. 2017. Effective headlines of newspaper articles in a digital environment. *Digital Journalism*, 5(10):1300–1314.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159.

Günther Lengauer, Frank Esser, and Rosa Berganza. 2012. Negativity in political news: A review of concepts, operationalizations and key findings. *Journalism*, 13(2):179–202.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6714–6721.

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

Nikola Ljubešić, Filip Klubička, Željko Agić, and Ivo-Pavao Jazbec. 2016. New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4264–4270, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

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

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*.

Jiaxin Pei, Aixin Sun, and Chenliang Li. 2019. Targeted sentiment analysis: A data-driven categorization. *arXiv preprint arXiv:1905.03423*.

Andraž Pelicon, Marko Pranjić, Dragana Miljković, Blaž Škrlj, and Senja Pollak. 2020. Zero-shot learning for cross-lingual news sentiment classification. *MDPI*.

Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. 2019. On the convergence of Adam and beyond. *arXiv preprint arXiv:1904.09237*.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860*.

Tomás Alves Salgueiro, Emilio Recart Zapata, Damián Furman, Juan Manuel Pérez, and Pablo Nicolás Fernández Larrosa. 2022. A Spanish dataset for targeted sentiment analysis of political headlines. *arXiv preprint arXiv:2208.13947*.

Joshua M Scacco and Ashley Muddiman. 2020. The curiosity effect: Information seeking in the contemporary news environment. *New Media & Society*, 22(3):429–448.

Holli A Semetko and Patti M Valkenburg. 2000. Framing European politics: A content analysis of press and television news. *Journal of communication*, 50(2):93–109.

Youwei Song, Jiahai Wang, Tao Jiang, Zhiyue Liu, and Yanghui Rao. 2019. Attentional encoder network for targeted sentiment classification. *arXiv preprint arXiv:1902.09314*.

Stuart Soroka, Lori Young, and Meital Balmas. 2015. Bad news or mad news? Sentiment scoring of negativity, fear, and anger in news content. *The ANNALS of the American Academy of Political and Social Science*, 659(1):108–121.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

James W Tankard Jr. 2001. The empirical approach to the study of media framing. In *Framing public life*, pages 111–121. Routledge.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. *Guide Proceedings*.

Zhihan Zhang, Wenhao Yu, Mengxia Yu, Zhichun Guo, and Meng Jiang. 2022. A survey of multi-task learning in natural language processing: Regarding task relatedness and training methods. *arXiv preprint arXiv:2204.03508*.

A Appendix

A.1 Dataset

Dataset Sampling. News headlines were taken from a news article database obtained through TakeLab Retriever, a tool for analyzing Croatian online news media. We used a stratified sampling technique on news outlet and date published attributes. A total of 3000 news headlines were sampled from 29 diverse news outlets, covering the time period between January 2000 and August 2022.

Annotation Guidelines. Annotation guidelines were given to annotators in the Croatian language. The instructions included the definition of named entities, targeted sentiment and tone for news headlines as well as annotation labels. We provided multiple annotation examples grouped by observed headline patterns.

The general guideline for annotating tone was to consider the impression of the headline, whereas for sentiment, it was the intentional impression towards the target entity. Further guidelines included: (1) when a headline contains a combination of positive and negative attributes toward the target entity, the final impression should be considered; (2) when the target entity's action expressed in the headline can be considered intrinsically negative or positive, this is transferred to the sentiment; (3) when the target entity is a toponym, it is crucial to identify whether it strictly represents a location (which is inherently neutral) or a metonymy (which can represent any sentiment); (4) when the headline contains a quotation, two cases are possible. If the chosen target entity is the author of the quote, the sentiment is usually neutral since no additional information is present. Otherwise, the attitude expressed by the author towards the entity is transferred to the sentiment of the target.

A.2 Tone Classification Results

Single-task	AVG	NEG	NTR	POS
Vanilla	**.773**	**.881**	**.598**	.840
Multi-task	AVG	NEG	NTR	POS
Alternate Batch	.761	**.881**	.575	.827
Alternate Epoch	.768	.875	.581	**.847**
Average Batch	.748	.876	.532	.835

Table 4: Tone classification macro-averaged and per-class F1-scores for single- and multi-task setups. The best results by setup are in **bold**.

Table 4 shows single- and multi-task tone classification results. The vanilla single-task tone model used BERTić with a classification layer on top. Multi-task setups are equivalent to the ones reported for TSA in Table 2. Results were averaged over five independent runs, using the same seeds, hyperparameters, and training procedures as for the TSA experiments.