LLMs for Targeted Sentiment in News Headlines: Exploring the Descriptive–Prescriptive Dilemma

Jana Juroš **Laura Majer** **Jan Šnajder**
University of Zagreb Faculty of Electrical Engineering and Computing
TakeLab
{jana.juros, laura.majer, jan.snajder}@fer.hr

Abstract

News headlines often evoke sentiment by intentionally portraying entities in particular ways, making targeted sentiment analysis (TSA) of headlines a worthwhile but difficult task. Due to its subjectivity, creating TSA datasets can involve various annotation paradigms, from *descriptive* to *prescriptive*, either encouraging or limiting subjectivity. LLMs are a good fit for TSA due to their broad linguistic and world knowledge and in-context learning abilities, yet their performance depends on prompt design. In this paper, we compare the accuracy of state-of-the-art LLMs and fine-tuned encoder models for TSA of news headlines using descriptive and prescriptive datasets across several languages. Exploring the descriptive–prescriptive continuum, we analyze how performance is affected by prompt prescriptiveness, ranging from plain zero-shot to elaborate few-shot prompts. Finally, we evaluate the ability of LLMs to quantify uncertainty via calibration error and comparison to human label variation. We find that LLMs outperform fine-tuned encoders on descriptive datasets, while calibration and F1-score generally improve with increased prescriptiveness, yet the optimal level varies.

1 Introduction

News framing impacts information perception, shapes public opinion, and guides discussions on key topics (Semetko and Valkenburg, 2000). News headlines – succinct and attention-grabbing introductions to full news stories – often evoke sentiment by portraying entities in specific ways. Targeted sentiment analysis (TSA) is the task of determining the polarity of sentiment expressed towards the target entity (Pei et al., 2019). While sentiment analysis is inherently challenging due to subjectivity, TSA introduces additional complexity by requiring the differentiation between targeted and overall sentiment.

For subjective tasks like TSA, the choice of data annotation paradigm is crucial. Rottger et al. (2022) identified two contrasting paradigms: *descriptive* and *prescriptive*. The descriptive paradigm encourages subjectivity and diverse interpretations, typically with brief guidelines. In contrast, the prescriptive paradigm discourages subjectivity by providing detailed interpretation guidelines.

Fine-tuned encoders such as BERT (Devlin et al., 2019) show strong TSA performance across various languages (Wu and Ong, 2021; Zhang et al., 2020; Mutlu and Özgür, 2022). However, using these models in different languages or domains requires new fine-tuning, and adapting them to low-resource languages necessitates pre-trained models and labeled data. In contrast, large language models (LLMs) offer a versatile approach to TSA across various domains by leveraging their broad linguistic and world knowledge, as well as in-context learning (Brown et al., 2020), without the need for annotated datasets or fine-tuning. However, LLMs performance is often inconsistent and contingent on prompt design (Mizrahi et al., 2024), making it challenging to identify optimal settings. Furthermore, it is unclear how specific TSA criteria, defined during annotation, can be transferred using zero- and few-shot prompting.

In this paper, we compare the zero- and few-shot performance of open and closed-source LLMs to fine-tuned encoder models on datasets annotated following the descriptive or prescriptive paradigm. We then explore the influence of prompt design on the performance of LLMs for the prescriptive TSA dataset of Croatian news headlines. Similar to crafting effective annotation guidelines, finding the appropriate *level of prescriptiveness* is essential in prompt design. The recent use of LLMs as data annotators (Wang et al., 2021; Pangakis et al., 2023; Alizadeh et al., 2023) further invites a direct comparison of annotation paradigms and prompt design: less prescriptive prompts give more interpretive freedom, while highly detailed prompts

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 329–343
August 15, 2024 ©2024 Association for Computational Linguistics

constrain it. Building on this parallel, we evaluate the predictive accuracy of LLMs using prompts constructed from annotation guidelines with different levels of prescriptiveness, ranging from plain zero-shot to elaborate few-shot prompts matching annotation guidelines.

Another interesting connection between annotation and prompting is label variation. Regardless of whether subjectivity is encouraged, some human label variation is inevitable in subjective tasks and may be leveraged to improve model performance (Mostafazadeh Davani et al., 2022). Similarly, LLM inconsistency, typically viewed as a limitation, can diversify responses to emulate human label variation. Recent LLM uncertainty quantification methods (Rivera et al., 2024; Xiong et al., 2023; Tian et al., 2023) can be used for the same purpose. Building on this idea, we assess LLMs' capability to quantify predictive uncertainty in TSA of headlines using calibration error and compare label distribution with human label variation.

Our experiments mainly focus on a Croatian dataset labeled with TSA on news headlines accompanied by detailed, prescriptive annotation guidelines. Additionally, we evaluate zero-shot LLMs and BERT on English, Polish, and Spanish TSA datasets with less prescriptive guidelines. Our contributions include (1) comparing LLMs and BERT for TSA on news headlines in four languages, (2) evaluating the effect of prompt prescriptiveness on LLMs' predictive accuracy, and (3) assessing calibration error and label distribution across models based on prompt prescriptiveness. This study offers valuable insights into LLMs' zero- and few-shot potential for TSA of news headlines.

2 Related Work

Sentiment analysis of news headlines is an important task that has garnered significant attention in prior work (Agarwal et al., 2016; Joshi et al., 2016; Aslam et al., 2020; Nemes and Kiss, 2021; Rozado et al., 2022). In addition to overall sentiment, TSA is crucial for understanding how entities are portrayed in news articles. Cortis et al. (2017) apply TSA on financial headlines, where sentiment is less implicit and topically constrained. Dufraisse et al. (2023) and Steinberger et al. (2011) present multilingual datasets for TSA in news articles. Hamborg and Donnay (2021) present a dataset for TSA on English news articles reporting on political topics, while (Balahur et al., 2013) focus on quotes from news articles. Overcoming the need for a labeled dataset, LLMs present a possible solution for TSA due to their in-context learning (ICL) abilities and broad background. Huang et al. (2020) conducted an analysis to identify and mitigate the entity bias of LLMs trained for sentiment analysis on Wikipedia and news articles. Chumakov et al. (2023) leverage both few-shot learning and fine-tuning with GPT models on mixed-domain Russian and English datasets to model sentiment effectively without domain-specific data.

3 Datasets and Models

Our experiments utilize, to our knowledge, the only two available datasets for TSA in general news headlines, alongside one domain-specific dataset. These datasets cover four languages and employ different annotation styles.

STONE. The STONE dataset (Barić et al., 2023) offers overall sentiment and targeted sentiment along with extracted target entities for Croatian news headlines, using ternary labels (positive, neutral, negative). Each of the 2855 headlines has 6 labels assigned by 6 annotators, with inter-annotator agreement (IAA) of κ = 0.416 (moderate agreement). Annotators were instructed using prescriptive, detailed guidelines (obtained from the authors upon our request). If a headline contained multiple entities, the target entity was chosen randomly and disclosed to the annotators.

SEN. The SEN (Baraniak and Sydow, 2021) dataset includes 3819 English and Polish news headlines, each featuring targeted sentiment labels and corresponding target entities. It comprises a Polish part (SEN_pl), an English part (SEN_en_r) annotated by volunteer researchers, and an English part (SEN_en_amt) annotated using Amazon Mechanical Turk. The reported Fleiss' kappa IAA are κ = .459, κ = .309, and κ = .303, respectively. Unlike STONE, SEN lacks raw labels, providing only an aggregated gold label per headline (positive, neutral, and negative), and was annotated using vaguer annotator guidelines, adhering more to the descriptive paradigm.

Spanish. The Spanish dataset (ES) of Salgueiro et al. (2022) comprises 1976 headlines concerning the 2019 Argentinian Presidential Elections. Three annotators assigned ternary labels to each headline with masked targets, with IAA of $\alpha = .62$. The authors do not disclose annotation guidelines, which

			SEN		
	STONE	**ES**	en_amt	en_r	pl
GPT 3.5	61.3	64.2	66.1	61.5	60.0
GPT 4	65.9	**67.0**	**68.8**	63.2	**69.5**
Neural Chat	59.8	63.0	66.3	**63.8**	58.1
Llama 3	53.5	60.5	59.2	52.7	51.2
Phi-3	43.5	61.7	58.3	52.7	47.3
Gemma	48.4	60.5	60.0	52.7	51.2
BERT*	**74.9**	66.7	63.6	56.2	61.9

Table 1: F1 scores across languages and datasets

suggests the straight-forward descriptive paradigm.

Models. We experiment with four open-source models: Neural Chat (NC) (7B), Llama 3 (8B), Phi-3 (3.8B), and Gemma (9B), pitted against two proprietary OpenAI models – GPT-4 Turbo (560B) and GPT-3.5 Turbo (175B) (OpenAI et al., 2023) (cf. Appendix A.3 for more details).

4 Experiments and Results

4.1 Predictive Accuracy

We first evaluate the LLMs' accuracy of TSA on headlines and compare them to top-performing BERT* models. We use the BERT models specifically pre-trained for each language – RoBERTa-base (Liu et al., 2019) for English, BERTić (Ljubešić and Lauc, 2021) for Croatian, BETO (Cañete et al., 2023) for Spanish and Polish-RoBERTa-base-v2 (Dadas et al., 2020) for Polish – and fine-tune each for TSA on the corresponding training set (cf. Appendix A.1 for dataset split sizes and A.2 for hyperparameters details). For LLMs, we use zero-shot prompting on the test set, using basic prompts outlining the task and the target classes (cf. Appendix B.1).

Table 1 presents the F1 scores on the test set portions of each dataset. On the descriptive datasets (SEN and ES), LLMs outperform BERT-based models. GPT-4 achieves the highest F1 score on the Polish SEN and the crowdsourced English SEN. Interestingly, on the English SEN annotated by researchers, NC outperforms both fine-tuned BERT models and GPT. However, on STONE– the prescriptively annotated dataset – BERTić surpasses all other models by a significant margin. We argue this performance difference might stem from using different annotation paradigms. The best-performing LLMs seem to grasp the descriptive paradigm well, performing TSA closest to annotators. On the other hand, the performance gap observed in LLMs on STONE may stem from the prompts' vagueness and lack of alignment with its prescriptiveness – a question we explore next.

Level	**Description**
1	Concise, exploring the fundamental concepts of sentiment and targeted sentiment.
2	Includes a definition of targeted sentiment specifically within the framework of news headlines.
3	Provided with concise guidelines.
4	Comprehensive instructions provided as guidelines, excluding examples.
5	Comprehensive instructions presented as guidelines, including examples and brief explanations.
6	Comprehensive instructions provided exactly as they were presented to the annotators.

Table 2: Short descriptions of prompt prescriptiveness levels (cf. Appendix B.3 for full prompts)

4.2 Level of Prompt Prescriptiveness

We utilize the STONE dataset and its annotator guidelines to create six prompts of increasing prescriptiveness level, with each subsequent level incorporating additional information from the guidelines. Table 2 outlines these six levels (cf. Appendix A for full prompts). Our goal is to assess the ability of LLMs to follow instructions as accurately as human annotators and to determine the most effective level of prompt prescriptiveness.

Table 3 shows the results. We observe variance in performance across all levels for all models. GPT-4 consistently outperforms other models across all levels, with GPT-4 and Neural Chat reaching their performance peaks at level 4 (detailed instructions formatted as guidelines without examples) and GPT 3.5 performing best at level 3 (concise guidelines). The performance drop seen at levels 5 and 6, the only ones with few-shot examples, may be due to the sensitivity regarding the selection and ordering of examples, a phenomenon observed in few-shot prompting (Lu et al., 2022; Chang and Jia, 2023). The increasing accuracy from levels 1 to 4 suggests that more prescriptive instructions positively impact LLM performance. Despite their overall lower performance, Llama 3, Gemma, and Phi-3 significantly improve at levels 5 and 6 (few-shot prompts). This difference in performance could be due to instruction tuning, which may have reduced sensitivity to few-shot configuration and improved context following.

4.3 Uncertainty Quantification

Given the inherent subjectivity of TSA and leveraging the stochastic nature of predictions generated

Level	NC	GPT 3.5	GPT 4	Llama 3	Gemma	Phi-3
1	59.8	60.1	65.9	53.5	48.4	43.5
2	61.2	58.3	64.3	50.9	48.8	40.6
3	61.5	**65.7**	69.9	52.9	55.8	44.2
4	**63.1**	64.0	**70.2**	51.9	53.1	43.6
5	60.5	63.0	66.8	60.6	59.3	**49.4**
6	62.5	64.5	68.2	**61.9**	**59.4**	46.3

Table 3: F1 scores for levels of prompt prescriptiveness

by LLMs, we explore how LLMs can model human label variation and whether this varies across levels of prompt prescriptiveness. Using STONE, we approach this question from two angles: (1) examining the relationship between LLMs' predictive and calibration accuracies and (2) investigating if the uncertainty of LLM predictions aligns with inter-annotator disagreement.

We use three uncertainty quantification methods: self-consistency sampling, distribution prompting, and verbal confidence assessment. *Self-consistency sampling* (SCS) (Xiong et al., 2023) leverages the inherent stochasticity of LLMs, influenced additionally by internal parameters such as temperature. For each headline, we prompt the same model six times and accumulate the responses to mimic the distribution of six annotator responses, setting the temperature to 0.7 for all models (cf. Appendix A.4 for details). The second method, which we refer to as *distribution prompting* (DP), prompts the model to explicitly predict how six annotators would label the targeted sentiment, directly resulting in a distribution of positive, neutral, and negative responses. Lastly, the *verbal confidence assessment* (VCA) method (Xiong et al., 2023) prompts the LLM to produce three predictions for each headline, representing each sentiment class, along with a confidence score ranging from 0 to 100. For the complete set of prompts used in each method, refer to Table 9 in Appendix B.2.

In addition to evaluating the model's prediction accuracy, we also consider model calibration. Calibration evaluates the alignment between a model's expressed confidence and its actual accuracy: ideally, predictions with a 70% confidence should be accurate 70% of the time. To analyze the calibration error, we consider only the labels with the highest confidence score for each headline and calculate the expected calibration error (ECE), computed as the average discrepancy between model confidence and observed accuracy. Model predictions are divided into m quantile-scaled bins B_i, with m set to 10 for this analysis. For each bin, we calculate both the average accuracy $\mathrm{acc}(B_i)$ and the average confidence $\mathrm{conf}(B_i)$. The Expected Calibration Error (ECE) is then derived as the weighted sum of the absolute differences between these averages, with weights proportional to the bin size n:

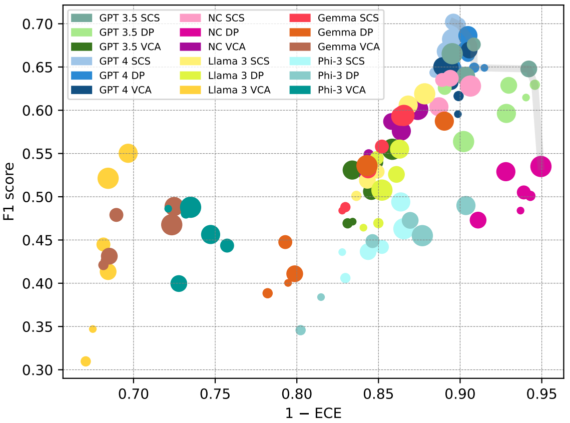

Figure 1: Comparison of F1 scores and calibration accuracy for various uncertainty quantification methods and across levels of prompt prescriptiveness (indicated by dot size). The gray lines indicate the Pareto front.

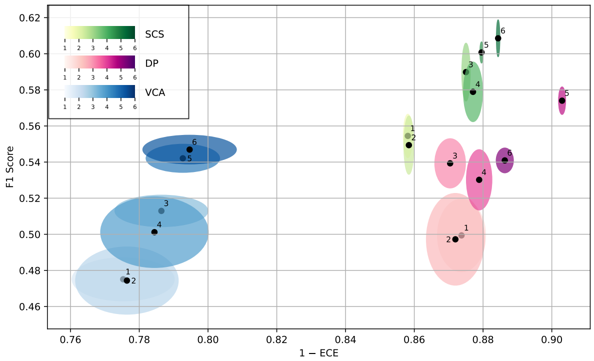

Figure 2: F1 scores and calibration accuracy averaged over all models across different uncertainty quantification methods and prescriptiveness levels (shaded ellipses indicate covariances)

$$\mathrm{ECE} = \sum_{i=1}^{m} \frac{|B_i|}{n} \left|\mathrm{acc}(B_i) - \mathrm{conf}(B_i)\right|. \qquad (1)$$

Figure 1 compares the predictive accuracy (F1 score) with calibration accuracy, defined as $1 - \mathrm{ECE}$, evaluated for each model. Figure 2 provides an overview of both metrics averaged across all models (cf. Tables 6 and 7 in Appendix A.7 for comprehensive data across all models). GPT-4 stands out as the best model, with the highest F1 scores and sound calibration, stable across different levels of prescriptiveness and uncertainty quantification methods. In comparison, the other models'

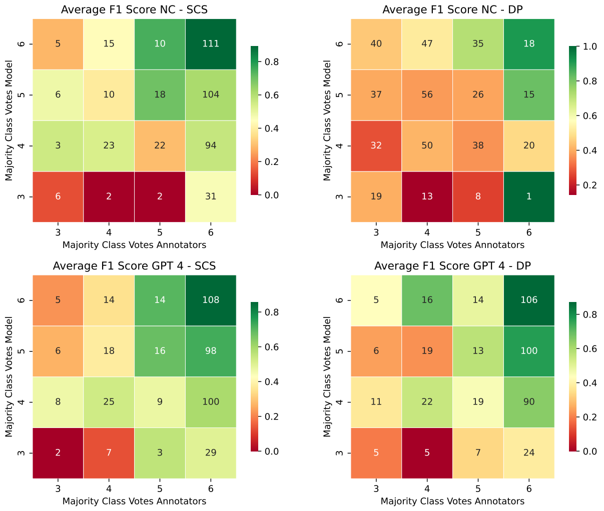

Figure 3: F1 score per majority vote bins for annotators (X) and model (Y) for SCS vs. DP for NC and GPT-4

performance is significantly affected by the uncertainty quantification method employed. Considering the averaged results, F1 scores are higher for SCS compared to DP and VCA, This aligns with expectations, as the models' predictions are evaluated solely against the gold labels. Average calibration accuracy is generally high (above 0.75) across models and uncertainty methods. Higher prescriptiveness levels show an increasing trend in predictive and calibration accuracy, with the optimal level varying by uncertainty method (Level 6 for SCS and VCA, and Level 5 for DP). This suggests that prescriptive annotation guidelines can enhance LLM performance for prescriptive datasets.

Besides quantifying uncertainty, SCS and DP can model human label variation, implicitly (SCS) or explicitly (DP). We compare these label distributions to human label variation. Figure 3 shows heatmaps of average F1 scores for the two best-performing open- and closed-source models, GPT-4 and NC. The axes represent the majority vote per instance by annotators and model. The highest F1 score is achieved when both the annotators' votes and the models' prediction are unanimous (6 votes). The lowest F1 scores are generally achieved for instances with less agreement within annotators or model votes. For GPT-4, DP performs similarly to SCS, whereas for NC, there is a significant performance drop and dispersion of model votes in bins, signaling the model is not grasping the concept. This suggests that SCS is a better choice for modeling label distribution across models.

5 Conclusion

Building on parallels with annotation paradigms for subjective tasks, we investigated the performance of LLM in-context learning for targeted sentiment analysis on news headlines. Our findings indicate that predictive accuracy increases with prompt prescriptiveness, though the optimal level varies by model, and only some models benefit from few-shot prompting. Calibration generally improves with prompt prescriptiveness, and self-consistency sampling aligns best with human label variation.

Limitations and Risks

Limitations. We find several limitations in this work. Firstly, our choice of LLMs is restricted. This is primarily due to computing and budget constraints. We are aware that a more expansive collection of models is necessary for a more comprehensive overview of LLM performance, along with open-source models larger than 8B parameters. Additionally, we prompted both GPT models using batches of data, which impacted performance during initial tests, but did not warrant the high costs of repeating the prompt for each individual instance.

Secondly, the aspect of varying prescriptiveness in prompts was only evaluated on one dataset, STONE. To our knowledge, there are currently no publicly available datasets on TSA in news headlines annotated with detailed guidelines. Furthermore, since the dataset in focus is in Croatian, it is unclear whether a difference in performance is due to the difference in the ability for sentiment analysis or the general understanding of the language and its cultural and political background, both essential for the task.

Finally, while evaluating the effect of prompt prescriptiveness level, the six levels were chosen arbitrarily so that they resemble a logical step-up in detail level. This number and method of prompt generation can differ based on the task at hand and annotation guidelines.

Risks. The risks in our work are mostly connected with the risks associated with sentiment analysis. Automatically evaluating sentiment might promote exclusion towards certain entities. As we performed no masking of entities, internal model biases could affect the classification.

References

Apoorv Agarwal, Vivek Sharma, Geeta Sikka, and Renu Dhir. 2016. Opinion mining of news headlines using SentiWordNet. In *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, pages 1–5.

Meysam Alizadeh, Maël Kubli, Zeynab Samei, Shirin Dehghani, Juan Diego Bermeo, Maria Korobeynikova, and Fabrizio Gilardi. 2023. Open-source large language models outperform crowd workers and approach chatgpt in text-annotation tasks.

Faheem Aslam, Tahir Mumtaz Awan, Jabir Hussain Syed, Aisha Kashif, and Mahwish Parveen. 2020. Sentiments and emotions evoked by news headlines of coronavirus disease (covid-19) outbreak. *Humanities and Social Sciences Communications*, 7(1).

Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik van der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. 2013. Sentiment Analysis in the News. ArXiv:1309.6202 [cs].

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

Ana Barić, Laura Majer, David Dukić, Marijana Grbeša-zenzerović, and Jan Snajder. 2023. Target Two Birds With One SToNe: Entity-Level Sentiment and Tone Analysis in Croatian News Headlines. In *Proceedings of the 9th Workshop on Slavic Natural Language Processing 2023 (SlavicNLP 2023)*, pages 78–85, Dubrovnik, Croatia. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2023. Spanish Pre-trained BERT Model and Evaluation Data. ArXiv:2308.02976 [cs].

Ting-Yun Chang and Robin Jia. 2023. Data curation alone can stabilize in-context learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8123–8144, Toronto, Canada. Association for Computational Linguistics.

Stanislav Chumakov, Anton Kovantsev, and Anatoliy Surikov. 2023. Generative approach to Aspect Based Sentiment Analysis with GPT Language Models. *Procedia Computer Science*, 229:284–293.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 Task 5: Fine-Grained Sentiment Analysis on Financial Microblogs and News. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Sławomir Dadas, Michał Perełkiewicz, and Rafał Poświata. 2020. Pre-training polish transformer-based language models at scale. In *Artificial Intelligence and Soft Computing*, pages 301–314. Springer International Publishing.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Evan Dufraisse, Adrian Popescu, Julien Tourille, Armelle Brun, and Jerome Deshayes. 2023. MAD-TSC: A Multilingual Aligned News Dataset for Target-dependent Sentiment Classification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8286–8305, Toronto, Canada. Association for Computational Linguistics.

Felix Hamborg and Karsten Donnay. 2021. NewsMTSC: A Dataset for (Multi-)Target-dependent Sentiment Classification in Political News Articles. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1663–1675, Online. Association for Computational Linguistics.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. ArXiv:1911.03064 [cs].

Kalyani Joshi, Bharathi N, and Jyothi Rao. 2016. Stock trend prediction using news sentiment analysis. *International Journal of Computer Science and Information Technology*, 8:67–76.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Nikola Ljubešić and Davor Lauc. 2021. BERTi\'c – The Transformer Language Model for Bosnian, Croatian, Montenegrin and Serbian. ArXiv:2104.09243 [cs].

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

M. Melih Mutlu and Arzucan Özgür. 2022. A Dataset and BERT-based Models for Targeted Sentiment Analysis on Turkish Texts. ArXiv:2205.04185 [cs].

László Nemes and Attila Kiss. 2021. Prediction of stock values changes using sentiment analysis of stock news headlines. *Journal of Information and Telecommunication*, 5(3):375–394.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, ..., and Barret Zoph. 2023. Gpt-4 technical report.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative AI requires validation. *CoRR*, abs/2306.00176.

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

Matthew Renze and Erhan Guven. 2024. The Effect of Sampling Temperature on Problem Solving in Large Language Models. ArXiv:2402.05201 [cs].

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining Confidence Elicitation and Sample-based Methods for Uncertainty Quantification in Misinformation Mitigation. ArXiv:2401.08694 [cs].

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

David Rozado, Ruth Hughes, and Jamin Halberstadt. 2022. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with Transformer language models. *PLOS ONE*, 17(10):e0276367.

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.

Holli Semetko and Patti Valkenburg. 2000. Framing european politics: A content analysis of press and television news. *Journal of Communication*, 50:93 – 109.

Josef Steinberger, Polina Lenkova, Mijail Kabadjov, Ralf Steinberger, and Erik van der Goot. 2011. Multilingual Entity-Centered Sentiment Analysis Evaluated by Parallel Corpora. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 770–775, Hissar, Bulgaria. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. ArXiv:2305.14975 [cs].

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhengxuan Wu and Desmond C. Ong. 2021. Context-Guided BERT for Targeted Aspect-Based Sentiment Analysis. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(16):14094–14102.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. ArXiv:2306.13063 [cs].

Huibing Zhang, Junchao Dong, Liang Min, and Peng Bi. 2020. A BERT Fine-tuning Model for Targeted Sentiment Analysis of Chinese Online Course Reviews. *International Journal on Artificial Intelligence Tools*, 29(07n08):2040018.

A Appendix

A.1 Additional Information on Datasets

In Table 4 the dataset sizes alongside the respective class counts are shown. For the STONE dataset (Barić et al., 2023) and the Spanish dataset (ES) (Cañete et al., 2023), we used the split for train, validation, and test sets given by the respective authors. For the variations of the SEN dataset, we used a 60/20/20 split generated using the sci-kit learn library with a fixed random seed of 42.

A.2 Optimization of BERT* Hyperparameters

For the BERT* models, we performed a grid search for hyperparameter optimization. We varied the learning rates, batch sizes, and number of epochs.

				SEN		
		STONE	**ES**	**en_amt**	**en_r**	**pl**
train	all	1614	1371	806	662	688
	pos	463	548	163	102	162
	neutr	810	434	314	355	308
	neg	341	389	329	205	218
valid	all	231	459	269	220	230
	pos	59	173	50	30	55
	neutr	120	167	89	118	101
	neg	52	119	130	72	74
test	all	462	609	269	220	230
	pos	122	241	54	45	50
	neutr	231	166	106	115	99
	neg	109	202	109	60	81

Table 4: Dataset sizes and sentiment counts used in experiments.

			SEN		
	STONE	**ES**	en_amt	en_r	pl
learning rate	1e-5	1e-5	2e-5	2e-5	3e-5
batch size	16	16	16	64	32
num of epochs	4	5	3	3	5

Table 5: Optimal hyperparameters determined for each dataset: for STONE, the results are obtained using the BERTić model; for ES, we used the BETO model; for SEN_en_amt and SEN_en_r, RoBERTa-base is utilized; and for SEN_pl, Polish-RoBERTa-base-v2 is employed.

The grid search covered the following hyperparameter values:

$$\begin{aligned}&\text{learning rate} : \{\text{5e-5, 3e-5, 2e-5, 1e-5, 5e-6}\}\\&\text{batch size} : \{16, 32, 64, 128, 256\}\\&\text{number of epochs} : \{1, 2, 3, 4, 5\}\end{aligned}$$

The optimal hyperparameters are summarized in Table 5.

A.3 Additional Information on Models

In our experiments, we employed the following LLMs:

Neural Chat [1] (7B): A fine-tuned model based on Mistral[2] with good coverage of domain and language.

Llama 3 instruct (8B) [3] : Instruction-tuned models fine-tuned and optimized for dialogue/chat use cases that outperform many of the available open-source chat models on common benchmarks.

Phi-3 Mini instruct [4] (3.8B): Phi-3 Mini is a lightweight, state-of-the-art open model by Microsoft[5], trained with a focus on high-quality and reasoning dense properties.

Gemma [6] (8.5B): Gemma is a lightweight, state-of-the-art open model built by Google DeepMind.[7]

GPT-4 Turbo [8] (560B): Latest generation OpenAI[9] model in time of running our experiments. We used the `gpt-4-1106-preview` model.

GPT-3.5 Turbo[10] (175B): Released in 2023, faster and more affordable OpenAI model. We used the `gpt-3.5-turbo-0125` model.

A.4 Setting the LLM Temperature Hyperparameter

Even though the optimal sampling temperature for problem-solving tasks is 0.0, as it maximizes reproducibility without compromising accuracy, LLMs showed relatively stable problem-solving performance across temperatures from 0.0 to 1.0, regardless of the LLM, prompt-engineering technique, or problem domain (Renze and Guven, 2024). For the purposes of uncertainty quantification and calibration assessment, we opted for a temperature of 0.7, maintaining stable performance while leveraging the stochastic properties of LLMs.

A.5 Additional Information on GPU Usage

We utilized a total of 201 hours of GPU resources. Specifically, 14 hours were allocated for obtaining results for optimal models and hyperparameters for BERT-based models. Additionally, 38 hours were dedicated to GPT 3.5 Turbo, 76 hours to GPT 4 Turbo, 62 hours to Neural Chat inference, and 11 hours to Mistral. Neural Chat and Mistral were run locally, while the GPT models were executed using the OpenAI Platform [11].

A.6 Additional Information on Used Toolkits

For tokenizing data to obtain results on BERT-based models, we utilized the PyTorch Transformers library[12].

A.7 Complete results

In this section, we present the complete results for all levels of prescriptiveness detail across all methods of uncertainty quantification. In Table 6 F1 scores are provided for all models and levels, and in Table 7 ECE is given per level and model. Figure 1 shows it graphically.

[11] https://platform.openai.com/docs/introduction

	NC			GPT 3.5			GPT 4			Llama 3			Gemma			Phi-3		
	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA
1	60.3	48.4	52.4	60.9	61.5	47.1	66.0	64.9	59.5	53.4	46.4	34.7	48.4	40.0	42.7	43.6	38.4	48.6
2	62.0	50.1	54.9	61.4	**62.9**	46.9	64.4	64.9	61.6	50.1	46.9	30.9	48.8	38.8	42.1	40.6	34.6	48.1
3	63.5	50.5	53.9	**67.6**	62.5	53.9	69.9	66.4	63.2	52.9	54.5	44.4	55.8	44.7	47.9	44.2	44.9	44.4
4	**63.7**	47.3	58.7	64.8	62.8	50.6	**70.2**	66.9	**66.9**	51.9	52.6	41.3	53.1	41.1	43.1	43.6	47.2	39.9
5	60.4	52.9	57.6	63.9	59.6	53.1	66.8	**68.6**	65.0	60.6	**55.5**	**55.0**	59.3	**58.7**	**48.9**	**49.4**	**48.9**	45.6
6	62.8	**53.5**	**60.1**	66.5	56.3	**55.4**	68.2	64.7	64.9	**61.9**	50.7	52.1	**59.4**	53.5	46.7	46.3	45.5	**48.8**

Table 6: F1 scores for levels of detail in prompt and uncertainty quantification metrics.

	NC			GPT 3.5			GPT 4			Llama 3			Gemma			Phi-3		
	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA	SCS	DP	VCA
1	13.1	6.3	15.9	11.3	6.0	16.6	11.1	**8.5**	10.1	15.3	15.9	32.5	17.2	20.5	31.8	17.2	18.5	27.9
2	12.1	5.7	15.6	10.9	**5.4**	16.9	11.6	9.1	10.1	16.4	15.0	32.9	17.0	21.8	31.9	17.0	19.8	26.8
3	11.1	6.1	15.3	9.2	10.9	15.1	**10.1**	9.5	10.5	15.1	15.1	31.8	14.8	20.7	31.1	14.8	15.4	**24.3**
4	10.6	8.9	14.2	**5.8**	7.0	15.4	10.4	9.7	**9.4**	15.7	13.9	31.6	15.6	20.1	31.5	15.6	13.1	27.2
5	11.3	7.2	13.6	9.7	7.2	16.6	10.8	9.5	11.1	13.2	**13.7**	**30.3**	13.6	**11.0**	**27.5**	13.6	**9.6**	25.3
6	**9.4**	**5.1**	**12.6**	10.5	9.8	**14.2**	10.5	10.6	10.7	**12.2**	14.8	31.6	**13.4**	15.7	27.7	**13.4**	12.3	26.5

Table 7: Expected calibration error (ECE) for levels of detail in prompt and uncertainty quantification metrics.

	Prompt
System	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.
User	Classify targeted sentiment towards entity {*entity*} in the following news headline: {*headline*}

Table 8: System and user prompt for zero-shot basic TSA on LLMs

B Prompt Catalogue

B.1 Prompts for Basic LLM Inference

Prompts for basic zero-shot TSA on LLMs are provided in Table 8. The system prompt establishes the task, and the user prompt provides the headline and target entity of the headline to be evaluated. The system prompt is aligned with Level 1 prescriptiveness in Table 10, and the user prompt corresponds to the Self-consistency Sampling (SCS) method of uncertainty quantification in Table 9.

B.2 Uncertainty quantification methods

Table 9 provides the user prompts for all the uncertainty quantification methods.

B.3 Prompts by Prescriptiveness Level

In this section, the complete prompts system and user prompts are given in tables 10, 11, 12, 13 and 14. The yellow highlight shows an expansion in text and information compared to the previous level.

	Prompt
SCS	Classify targeted sentiment towards entity {*entity*} in the following news headline: {*headline*}
DP	Your task is to imagine you are representing 6 different people detecting the targeted sentiment in Croatian news headlines, each following the given guidelines. For a headline and an entity, you need to return detected targeted sentiment for each of the 6 voters. Detect targeted sentiment for entity ’*entity*’ in headline: ’*headline*’. Possible sentiment classes are positive, neutral and negative. Please return the answer in JSON format like: ["targeted sentiment 1":"class 1" "targeted sentiment 2":"class 2" "targeted sentiment 3":"class 3" "targeted sentiment 4":"class 4" "targeted sentiment 5":"class 5" "targeted sentiment 6":"class 6"]
VCA	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. Following the given guidelines, please return the confidence for detection of each class. Detect targeted sentiment for entity {*entity*} in headline: {*headline*}. Possible sentiment classes are positive, neutral and negative. Please return the confidence for each class in format like: ["confidence positive class", "confidence neutral class" ,"confidence negative class"]

Table 9: User prompt used for inference on the STONE dataset accross methods for uncertainty quantification.

Level	Prompt
1	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.
2	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. Targeted sentiment involves understanding the author’s intention to evoke emotion towards a target entity, considering the deliberate choice in conveying news and recognizing that the same information can be presented in various ways, with the understanding that such intentional choices aid in detecting the targeted sentiment. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.
3	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. Targeted sentiment is the emotional stance the author aims to convey specifically towards a mentioned entity. It involves interpreting the intention behind the author’s choice of language and tone when discussing the target entity. The sentiment is not only influenced by the conveyed information but also by the author’s subjective evaluation and emotional coloring of the entity. Actions associated with the entity play a role in determining the sentiment, with negative actions implying a negative quality and, consequently, a negative sentiment. Distinguishing between negative actions and negative occurrences is crucial, as negative occurrences towards the entity don’t color the entity. In headlines featuring a quote, the entity authoring the quote is attributed neutral sentiment as they are merely conveying an opinion. The overall goal of the author, whether it be praise or criticism, is considered in cases of headlines with a mix of positive and negative views. In summary, targeted sentiment is the nuanced emotional evaluation directed specifically at a particular entity within the context of news reporting. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.

Table 10: System prompts used for inference on the STONE dataset.

Level	Prompt
4	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. Guidelines for Targeted Sentiment Annotation: 1. Detecting Sentiment through Author's Intent and News Presentation: Evaluate the intended sentiment towards an entity by analyzing the emotions the author aims to evoke and recognizing that news can be conveyed in multiple ways, with the chosen manner of conveyance serving a purpose and aiding in targeted sentiment detection. 2. Impact of Entity Actions: Acknowledge that entity actions influence sentiment, with negative actions implying negative quality. However, distinguish between negative actions undertaken by the entity and negative occurrences directed towards the entity that do not inherently portray the entity in a negative light. 3. Neutrality of Quoting Authors: In headlines featuring quotes, two types of entities are involved: the statement's author and the entities mentioned in the quote. If the target entities in the headline are the authors of the statement, the sentiment towards them typically leans towards neutrality because, in this scenario, they serve as conveyors of an opinion rather than direct subjects of sentiment. 4. Overall Authorial Goal: Consider the author's overall goal, whether it involves praise or criticism, especially in mixed-view headlines. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.
5	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. Guidelines for Targeted Sentiment Annotation: 1. Detecting Sentiment through Author's Intent and News Presentation: Evaluate the intended sentiment towards an entity by analyzing the emotions the author aims to evoke and recognizing that news can be conveyed in multiple ways, with the chosen manner of conveyance serving a purpose and aiding in targeted sentiment detection. Examples Illustrating Sentiment towards Entity Solin: Headline: 'SRAMOTA USolinuse djeca nemaju gdje liječiti, roditelji očajni' Targeted Sentiment: Negative Explanation: The author criticizes Solin, suggesting a disgraceful situation where children lack medical care, portraying a negative sentiment. Headline: 'U Solinu radi samo jedna pedijatrica, roditelji traže hitno rješenje' Targeted Sentiment: Negative Explanation: The negative sentiment persists as the author emphasizes the shortage of pediatricians in Solin, prompting urgent solutions according to parents. Headline: 'U Solinu nastupio nedostatak liječničkog kadra' Targeted Sentiment: Neutral Explanation: The sentiment is neutral here as the author focuses on conveying information about the shortage of medical staff without explicitly criticizing the responsible institutions.

Table 11: System prompts used for inference on the STONE dataset.

Level	Prompt
5	2. Impact of Entity Actions: Recognize that entity actions play a role in shaping sentiment, particularly with negative actions like murder and theft suggesting a negative quality. However, distinguish between negative actions where the entity is the perpetrator and negative occurrences where the entity is the recipient. Acknowledge that in the case of negative occurrences, the entity cannot be held responsible for the consequences but may be in a negative situation as a result, implying neutrality in the assessment. Headlines with negative quality of entities linked to their actions: a) Examples of linking entity quality to actions: Headline: 'Bivša tehnološka direktorica Elizabeth Holmes osuđena na 11 godina zatvora' Entity: Elizabeth Holmes Targeted Sentiment: Negative Explanation: Negative sentiment is assigned to Elizabeth Holmes based on her negative actions. Headline: 'Zbog ubojstva srpskih civila sudit će se Đuri Brodarcu, bivšem Sanaderovom savjetniku' Entity: Đuro Brodarac Targeted Sentiment: Negative Explanation: Negative sentiment is assigned to Đuro Brodarac due to his association with a serious crime. b) Examples of negative occurences towards the entity. Headline: 'Potres u Indoneziji: Poginulo najmanje 46 ljudi, ozlijeđenih oko 700' Entity: Indonezija Targeted Sentiment: Neutral Explanation: Neutral sentiment is assigned to Indonesia as the entity is a recipient of a negative occurrence. Headline: 'Horor u Mogadišuu: U terorističkom napadu na hotel 10 mrtvih, ozlijeđen i somalijski ministar' Entity: Mogadišu Targeted Sentiment: Neutral Explanation: Similar to the previous example, neutral sentiment is assigned to Mogadishu as it is a recipient of a negative occurrence. 3. Neutrality of Quoting Authors: Define sentiment towards the entity by considering the author's stance in a statement, whether the author is the headline writer or the individual quoted. When conveying someone's sentiment in a quote, transfer that sentiment to the mentioned entity. In headlines quoting individuals, recognize two entity types: the statement's author and the entities mentioned in the quote. If the target entities in the headline are the authors of the statement, the sentiment is typically neutral since they serve as conveyors of an opinion. Examples of Handling Quotes in Headlines: Headline: 'Milanović: Žao mi je što sam podržao Bidena' Entity: Milanović Targeted Sentiment: Neutral Entity: Biden Targeted Sentiment: Negative Explanation: Neutral sentiment is assigned to Milanović, who is conveying an opinion, while negative sentiment is assigned to Biden based on the conveyed sentiment. Headline: 'Gotovac: Ako sam ja politički antitalent, onda je tom antitalentu išlo bolje nego Grbinu' Entity: Gotovac Targeted Sentiment: Positive Entity: Grbin Targeted Sentiment: Negative Explanation: Positive sentiment is assigned to Gotovac, who comments on himself, while negative sentiment is assigned to Grbin based on the conveyed sentiment. Headline: 'Anka Mrak Taritaš: Tužna sam i razočarana situacijom u Zagrebu. Tomašević ne bi dobio dobru ocjenu' Entity: Anka Mrak Taritaš Targeted Sentiment: Neutral Entity: Tomašević Targeted Sentiment: Negative Explanation: Neutral sentiment is assigned to Anka Mrak Taritaš, the quoted individual, while negative sentiment is assigned to Tomašević based on the conveyed sentiment.

Table 12: System prompts used for inference on the STONE dataset.

Level	Prompt
5	4. Overall Authorial Goal: Consider the author’s overall goal, whether it involves praise or criticism, especially in mixed-view headlines. Example of a Combined Statement (Combination of Positive and Negative Views) Headline: ’Vanna je definitivno promijenila stil naglavačke i dosadne kombinacije zamijenila onima koje prate trendove’ Entity: Vanna Targeted Sentiment: Positive Explanation: A positive sentiment is attributed to Vanna because the author’s intention is to praise the improvement in her style, despite simultaneously criticizing her previous dressing style. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.
6	You are a helpful assistant who performs targeted sentiment classification in Croatian news headlines. Here are some guidelines for detecting targeted sentiment in news headlines: To determine sentiment towards an entity, we consider the kind of emotion the statement’s author intended to evoke regarding the target entity, that is, how the author intended to "color" that entity. To aid in discerning the intended sentiment towards the entity, one can consider the fact that the same piece of news can always be conveyed in multiple ways. The chosen manner of conveying a piece of news is selected with a purpose, and understanding that intention can be utilized for targeted sentiment detection. An example of various ways of reporting the same news about entity Solin: Headline: ’SRAMOTA USolinuse djeca nemaju gdje liječiti, roditelji očajni’ Targeted Sentiment: Negative Explanation: Negative sentiment is attributed to Solin due to the author’s intention to criticize the institution for the shortage of pediatricians. Headline: ’U Solinu radi samo jedna pedijatrica, roditelji traže hitno rješenje’ Targeted Sentiment: Negative Explanation: Similar negative sentiment is conveyed towards Solin by criticizing the shortage of medical staff. Headline: ’U Solinu nastupio nedostatak liječničkog kadra’ Targeted Sentiment: Neutral Explanation: Neutral sentiment is assigned as the author’s intention is to convey information without criticizing the responsible institutions. When detecting targeted sentiment, we can assign a quality to the target entity as an aid in determining sentiment, based on the emotion the statement’s author associates with it. The quality of the entity is linked to the actions of that entity, which can be either negative or positive. Negative actions of the entity, such as murder, theft, and other illegal or socially unacceptable activities like insults, are attributed to the quality of that entity. Negative actions signify a negative quality of the entity, implying a negative sentiment. The same approach will be applied in cases of positive actions of the entity, indicating a positive sentiment towards the entity. It is necessary to distinguish between the negative actions of an entity and negative occurrences towards the entity. In the case of negative actions by the entity, the entity is the perpetrator and therefore responsible for that action. In the case of negative occurrences towards the entity, the entity is the recipient of the negative action and cannot be held responsible for the consequences of the action, although it may be in a negative situation as a result. Examples of linking entity quality to actions: Headline: ’Bivša tehnološka direktorica Elizabeth Holmes osuđena na 11 godina zatvora’ Entity: Elizabeth Holmes Targeted Sentiment: Negative Explanation: Negative sentiment is assigned to Elizabeth Holmes based on her negative actions.

Table 13: System prompts used for inference on the STONE dataset.

Level	Prompt
	Headline: ’Zbog ubojstva srpskih civila sudit će se Ðuri Brodarcu, bivšem Sanaderovom savjetniku’ Entity: Ðuro Brodarac Targeted Sentiment: Negative Explanation: Negative sentiment is assigned to Ðuro Brodarac due to his association with a serious crime. Examples of negative occurences towards the entity. Headline: ’Potres u Indoneziji: Poginulo najmanje 46 ljudi, ozlijeđenih oko 700’ Entity: Indonezija Targeted Sentiment: Neutral Explanation: Neutral sentiment is assigned to Indonesia as the entity is a recipient of a negative occurrence. Headline: ’Horor u Mogadišuu: U terorističkom napadu na hotel 10 mrtvih, ozlijeđen i somalijski ministar’ Entity: Mogadišu Targeted Sentiment: Neutral Explanation: Similar to the previous example, neutral sentiment is assigned to Mogadishu as it is a recipient of a negative occurrence. We define sentiment towards the entity as the author’s stance towards the target entity in a statement. The statement’s author can be the person who wrote the article headline or the author whose quote is conveyed in the form of the article headline. When conveying someone’s negative/positive sentiment in a quote or paraphrase, that sentiment is transferred to the entity. In headlines conveying someone’s quote, there are two types of entities - the statement’s author and the entities mentioned in the quote. If the target entities in the headline are the authors of the statement, the sentiment towards them will usually be neutral because, in this case, they are just conveyors of an opinion. An exception is the following example with entity Gotovac, where the statement’s author comments on himself, and the expressed sentiment is then transferred to the author himself.
6	Examples of Handling Quotes in Headlines: Headline: ’Milanović: Žao mi je što sam podržao Bidena’ Entity: Milanović Targeted Sentiment: Neutral Entity: Biden Targeted Sentiment: Negative Explanation: Neutral sentiment is assigned to Milanović, who is conveying an opinion, while negative sentiment is assigned to Biden based on the conveyed sentiment. Headline: ’Gotovac: Ako sam ja politički antitalent, onda je tom antitalentu išlo bolje nego Grbinu’ Entity: Gotovac Targeted Sentiment: Positive Entity: Grbin Targeted Sentiment: Negative Explanation: Positive sentiment is assigned to Gotovac, who comments on himself, while negative sentiment is assigned to Grbin based on the conveyed sentiment. Headline: ’Anka Mrak Taritaš: Tužna sam i razočarana situacijom u Zagrebu. Tomašević ne bi dobio dobru ocjenu’ Entity: Anka Mrak Taritaš Targeted Sentiment: Neutral Entity: Tomašević Targeted Sentiment: Negative Explanation: Neutral sentiment is assigned to Anka Mrak Taritaš, the quoted individual, while negative sentiment is assigned to Tomašević based on the conveyed sentiment.

Table 14: System prompts used for inference on the STONE dataset.

Level	Prompt
6	In the case of a headline containing a combination of positive and negative views towards the entity, the final goal of the author towards the entity is considered, i.e., whether the author aimed for praise or criticism. Example of a Combined Statement (Combination of Positive and Negative Views): Headline: ’Vanna je definitivno promijenila stil naglavačke i dosadne kombinacije zamijenila onima koje prate trendove’ Entity: Vanna Targeted Sentiment: Positive Explanation: Positive sentiment is assigned to Vanna as the author’s intention is to praise the improvement in her style despite also criticizing her previous dressing choices. The available sentiment classes are positive, neutral, and negative. For each given headline, identify the targeted sentiment class towards the entity.

Table 15: System prompts used for inference on the STONE dataset.