Few-Shot Prompting, Full-Scale Confusion: Evaluating Large Language Models for Humor Detection in Croatian Tweets

Petra Bago[1], **Nikola Bakarić**[2]
[1]Faculty of Humanities and Social Sciences, University of Zagreb
[2]University of Applied Sciences Velika Gorica
pbago@ffzg.hr, nbakaric@vvg.hr

Abstract

Humor detection in low-resource languages is hampered by cultural nuance and subjective annotation. We test two large language models, GPT-4 and Gemini 2.5 Flash, on labeling humor in 6,000 Croatian tweets with expert gold labels generated through a rigorous annotation pipeline. LLM–human agreement (κ = 0.28) matches human–human agreement (κ = 0.27), while LLM–LLM agreement is substantially higher (κ = 0.63). Although concordance with expert adjudication is lower, additional metrics imply that the models equal a second human annotator while working far faster and at negligible cost. These findings suggest, even with simple prompting, LLMs can efficiently bootstrap subjective datasets and serve as practical annotation assistants in linguistically under-represented settings.

1 Introduction

Humor detection remains a subtle challenge in natural language processing (NLP), due to its dependence on cultural norms, context, and individual interpretation. While computational humor has traditionally been approached through rule-based systems or crowd-sourced annotations, recent advances in large language models (LLMs) have enabled a new paradigm: automated annotation using pre-trained models.

Prompt-based learning. Surveys by Liu et al. (2023) and Zhao et al. (2023) show that carefully engineered prompts let LLMs perform zero- or few-shot classification in low-resource settings, albeit with lingering problems of bias, prompt sensitivity, and framing effects. Philippy et al. (2025) extend the idea cross-lingually, finding that soft prompts transfer knowledge across languages but still need adaptation under distribution shift. Fields et al. (2024) caution that bigger models are not always more accurate or economical, and Chae & Davidson (2023) report that fine-tuned smaller models can rival GPT-4-scale systems when adequate labeled data exist—an important trade-off for multilingual or resource-limited work.

LLMs as annotators. Ding et al. (2023) and He et al. (2024) evaluate GPT-3.5 as a stand-in for human crowdworkers; with explain-then-annotate prompting, the model matches or outperforms human accuracy at lower cost, though it still overgeneralizes and handles edge cases poorly. Gilardi et al. (2023) reinforce these findings: ChatGPT surpassed both crowd and expert annotators on several text-classification benchmarks, yet struggled on semantically rich categories—humor among them—where cultural grounding is crucial, a limitation echoed in the stance-detection study of Chae & Davidson (2023).

Despite this progress, most work is English-centric, leaving smaller languages underexplored. We close that gap by testing GPT-4 and Gemini 2.5 Flash as humor annotators for Croatian tweets. Using a 6,000-tweet corpus with adjudicated gold labels, we measure LLM–human agreement and practical efficiency in a low-resource, high-subjectivity setting.

The remainder of the paper proceeds as follows: Section 2 surveys computational humor and LLM-based annotation; Section 3 introduces the dataset; Section 4 explains the experimental design; Section 5 presents results; and Section 6 discusses limitations and future work.

2 Related Work

Computational humor research spans three axes: recognition, generation, and scoring. Early work by Mihalcea & Strapparava (2005) cast recognition as binary classification on hand-selected one-liners, proving feasibility yet omitting context, audience, and annotation uncertainty.

Proceedings of the 10th Workshop on Slavic Natural Language Processing (Slavic NLP 2025), pages 9–16
July 31, 2025 ©2025 Association for Computational Linguistics

Social-media corpora then became central. Castro et al. (2018) compiled 27,000 Spanish tweets, labeling humor presence and graded funniness through crowdsourcing; their moderate agreement (α = 0.57) underscored humor's subjectivity and showed how sampling choices affect balance and bias.

Disagreement itself has since been scrutinized: analyzing German conversational laughter, Ludusan (2024) reported uniformly low consistency across functional tags and compared adjudication with majority voting—insights transferable to humor labeling.

Researchers have also moved beyond monolingual text. Bedi et al. (2023) released MaSaC, a multimodal Hindi-English code-mix dataset for sarcasm and humor; hierarchical attention over acoustic and textual cues improved detection, highlighting that humorous intent is multimodal, contextual, and language-specific.

On the generation side, Mirowski et al. (2024) conducted workshops with professional comedians and found that ChatGPT still produces bland or stereotypical jokes, revealing limited cultural grounding despite impressive fluency.

Collectively, these studies expose three enduring challenges: annotator disagreement, scarcity of non-English multimodal resources, and unclear LLM capability in humor understanding versus annotation. Our work tackles all three by testing GPT-4 and Gemini 2.5 Flash as rapid, low-cost annotators of Croatian humor tweets and comparing their labels against adjudicated human gold.

3 Dataset

3.1 Source Corpus

We build on Twitter-HBS 1.0 (Ljubešić & Rupnik 2022), a crawl of 63,160 Twitter users labelled as Bosnian, Croatian, Montenegrin or Serbian. From that collection, Živičnjak (2025) had already produced a Croatian-only slice: 28,129 tweets after deduplication. Tweets shorter than four tokens or consisting solely of URLs/photos were discarded, yielding 11,929 unique tweets. A simple random sample of 6,000 tweets was drawn for manual annotation, naming the result HRumor 1.0: Corpus of Croatian Humorous Tweets.

3.2 Annotation Protocol

The 6,000 tweets were divided into six 1,000-tweet batches. Each batch was labelled by two native Croatian speakers (12 annotators total) who followed concise written guidelines with examples. Using a spreadsheet, they assigned one of two mutually exclusive labels: **HUMOROUS** – tweet is intended to be funny for a general Croatian audience; **NOT HUMOROUS** – all other cases.

Annotators worked blindly and independently. When the pair disagreed, a single expert adjudicator (13-th annotator) inspected both labels and issued a binding decision. Thus every tweet has three annotations, but only the adjudicated label is treated as gold.

3.3 Label distribution and agreement

The corpus is skewed toward the negative class: **994 HUMOROUS** (16.57 %) vs. **5,006 NOT HUMOROUS** (83.43 %). Similar imbalance appears in earlier Twitter humor corpora—22.5 % (Holton & Lewis 2011), 7.0 % (Mendiburo-Seguel et al. 2022) and 1.4 % (Vázquez 2016).

Blind annotator pairs achieved Cohen's κ = 0.26 ± 0.07 (min 0.21, max 0.41), "fair" on the Landis & Koch (1977) scale. Agreement with the adjudicator is markedly higher: κ = 0.69 ± 0.05 for the stronger annotator, κ = 0.50 ± 0.12 for the weaker, implying that many conflicts are borderline judgments rather than errors. Because κ decreases as class prevalence becomes skewed (Pontius Jr & Millones 2011), the ≈17 % positive rate lowers absolute values. Aggregate figures are reported here; detailed model comparisons follow in § 5.

HRumor 1.0 is planned for public release in the near future, with licensing details to be determined. The remainder of this paper tests whether GPT-4 and Gemini 2.5 Flash can serve as fast, low-cost annotators for such subjective, low-resource data.

4 Experiment

We tested humor annotation in a low-resource setting by prompting **GPT-4** (OpenAI API) and **Gemini 2.5 Flash** (Google API) to label 6,000 Croatian tweets. All LLM annotations (GPT-4 and Gemini 2.5 Flash) were performed on the identical set of 6,000 tweets labeled by human annotators, enabling a direct comparison of human and automated annotations. Each tweet already bore two independent human labels plus third-expert adjudication, forming a high-quality gold standard for comparison.

4.1 Setup

A custom Python pipeline sent tweets to both APIs in batches of five through a single English prompt that remained unchanged across calls, guaranteeing consistency. The prompt declared the system role, stated the humor-classification task, and supplied four fixed few-shot Croatian examples (see Appendix A). Models could return only **HUMOROUS**, **NOT HUMOROUS**, or **MAYBE HUMOROUS** and had to produce four tab-separated fields—tweet ID, ≤ 8-word Croatian rationale, label, and 0–100 confidence. Our use of a third MAYBE HUMOROUS label mirrors the human annotation protocol and serves dual purposes: ensuring methodological consistency and enabling the identification of tweets that may warrant deeper qualitative analysis in future work.

Translation was forbidden: with **step-back prompting** (Boonstra 2025), the model had to print its brief Croatian rationale before committing to a label, encouraging more deliberate, less generic decisions.

Both systems ran at **temperature 0** for deterministic output. Batch token ceilings were 1,500 for GPT-4 and 5,000 for Gemini 2.5 Flash, ample for the prompt plus structured reply. This configuration yielded fully parsable outputs without manual intervention while preserving the rich rationale information needed for future error analysis.

4.2 Labeling and Normalization

Models could output **HUMOROUS**, **NOT HUMOROUS**, or **MAYBE HUMOROUS**. Under the MAYBE label, GPT labeled 847 tweets (14.12 %), whereas Gemini labeled 595 tweets (9.92 %). To match the final binary human scheme we post-processed predictions: tweets tagged MAYBE with confidence ≥ 50 were reassigned to HUMOROUS, whereas all other cases—including MAYBE < 50—became NOT HUMOROUS.

The full model output was written in a tab-separated format, with one line per tweet, allowing automatic parsing and alignment with gold annotations. No manual intervention was required at any stage of model interaction or output extraction.

This design enables direct, statistics-ready comparison of human and LLM annotations. Because humor is intrinsically subjective and even human agreement is limited, we do not attempt a fine-grained error taxonomy in this study.

5 Results

We gauged LLM–human alignment with Cohen's κ (Cohen 1960), the standard chance-corrected agreement statistic. By factoring in chance agreement, κ yields a conservative baseline, yet it is sensitive to skewed class distributions (Pontius Jr & Millones 2011). Because only 16.57 % of our tweets are HUMOROUS, κ may under-estimate true concordance. We therefore also computed prevalence-adjusted bias-adjusted kappa (PABAK) (Khraisha et al. 2024). PABAK preserves the same ranking: strong agreements look even stronger, weak ones slightly lower. For clarity and consistency with prior work, we focus on reporting Cohen's κ in the main text.

Table 1 summarizes agreement scores across all relevant annotator and model pairings, averaged across the six subcorpora. For each pairing, we report the mean (with SD), median, minimum, maximum, and range of κ values, detailed table found in Appendix B.

5.1 Agreement Patterns

We observe five main patterns in the agreement scores. **(1) Human–Human Agreement.** Agreement between primary annotators (A vs. B) is relatively low (κ = 0.27 on average), consistent with prior findings on the subjectivity of humor classification. **(2) Human–Adjudicator Agreement.** Agreement between each annotator and the adjudicated gold label is substantially higher (mean κ = 0.59), confirming the value of expert adjudication in borderline or ambiguous cases. **(3) LLM–Human Agreement.** Agreement between LLMs and human annotators (mean κ = 0.28) is slightly higher than human–human agreement (by ~1 percentage point), which suggests that the models can stand in for a human in binary humor classification. **(4) LLM-Adjudicator Agreement.** The lowest observed agreement is between LLMs and the adjudicator (mean κ = 0.20). **(5) LLM-LLM Agreement.** Interestingly, LLMs agree strongly with each other (mean κ = 0.63), suggesting that while they may develop a shared labeling strategy, it does not consistently align with human intuition.

Annotator and model pairings	Mean (SD)	Median	Min	Max	Range
A & B	0.2649 (±0.0744)	0.2396	0.2080	0.4115	0.2035
Adjudicator & (A / B)	0.5948 (±0.1379)	0.6025	0.3238	0.8216	0.4977
LLM & (A / B)	0.2758 (±0.1243)	0.2839	0.1109	0.5195	0.4087
LLM & Adjudicator	0.2049 (±0.0466)	0.2104	0.1250	0.2586	0.1337
LLM & LLM	0.6322 (±0.0254)	0.6258	0.5964	0.6697	0.0733

Table 1: Summarized Cohen's kappa agreement scores averaged across six subcorpora

Croatian examples (with English translation) of agreement across all annotators (human and LLM) can be found in Appendix E.

In addition to annotator agreement metrics, we compared the annotation results of each LLM (as predicted class) to the human annotators and adjudicator respectively (as actual class) using a confusion matrix. While considering the skewness of data in favor of the NON HUMOROUS class, we observed fair values of the F1 score (0.61 on average) for all combinations of annotators A and B with both LLMs. The F1 score for adjudicator and both LLMs is lower and stands at 0.53 for GPT-4 and 0.54 for Gemini 2.5 Flash. Nevertheless, F1 scores and fair confusion matrix agreement for true positive and true negative cases between the two human annotators (A and B) and the LLMs leads us to believe that the models could shoulder a significant part of annotation efforts in this and similar scenarios. Classification results for all annotators, adjudicator and LLMs can be found in Appendix C, while the full table of F1 scores and confusion matrices for all evaluated combinations is available in Appendix D.

For illustrative purposes, we present below one representative tweet annotated independently by both GPT-4 and Gemini 2.5 Flash. Although the original model outputs provided only the Tweet ID, explanation, label, and confidence score, here we include the tweet text along with its English translation to improve readability and understanding.

Example Tweet ID & tweet:

5676 To je i ona rekla.*

**[That's what she said.]*

GPT-4 output:

5676 Nejasan kontekst, moguća dvoznačnost**
MAYBE HUMOROUS 50

***[Unclear context, possible ambiguity.]*

Gemini 2.5 Flash output:

5676 Klasična šala s dvostrukim značenjem.***
HUMOROUS 95

**** A classic double-meaning joke.*

To gauge practicality, we logged runtime and cost. GPT-4 (OpenAI API) processed the 6,000 tweets in 3 h 28 min for USD 48.57. Gemini 2.5 Flash completed the same set in 2 h 25 min, and, with larger batches, 1 h 27 min; running under a free-tier quota, it incurred zero cost.

These results highlight both the potential and the limitations of LLMs as humor annotators and form the basis for further analysis in the following section.

Overall, LLM–human agreement matches human–human reliability, yet both models align less with the adjudicated gold. Their mutual κ is much higher, indicating a shared, consistent labeling strategy even though they sometimes diverge from nuanced human judgments. While our findings demonstrate high internal consistency among LLMs, their lower agreement with adjudicated labels could also indicate subjective biases inherent to any single adjudicator. Future studies should explore multiple adjudicators or larger annotator groups to verify whether this discrepancy persists.

6 Conclusion and Future Work

Our results show that large language models can reach inter-annotator agreement comparable to humans in humor classification. GPT-4 and Gemini 2.5 Flash exhibit strong mutual consistency (κ = 0.63), even in a subjective, culturally grounded task like humor recognition, yet align less with the adjudicated gold. Interestingly, LLM–human κ = 0.28 nearly matches human–human κ = 0.27, indicating the models can act as a second annotator. This parity does not imply that LLMs understand humor; they may exploit surface cues rather than deep pragmatic insight, agreeing with humans for different reasons. While LLMs display strong mutual consistency, their divergence from expert adjudication may stem from fundamentally different processing mechanisms. Explaining this divergence calls for in-depth research into LLM internals, beyond the scope of this study.

Nonetheless, the practical upside is clear: speed, negligible marginal cost, and reproducible labels make LLMs attractive in low-resource or large-scale settings—for dataset bootstrapping, cost reduction, or pre-labeling before expert review.

Several extensions merit investigation. Applying the same pipeline to other platforms or text genres would test generalizability. Broader model and prompt exploration, including chain-of-thought prompting, could reveal whether agreement improves with alternative framing. The current prompting followed human annotation instructions closely. However, we hypothesize that LLM-specific prompting, acknowledging their distinct interpretative mechanisms, could improve performance and merits systematic exploration. Testing additional Slavic languages would probe cross-lingual robustness.

Further analytical work may also prove valuable. Qualitative inspection of disagreement cases could pinpoint humor types that mislead LLMs. The MAYBE HUMOROUS label, which we retained from the human guidelines, provides a pool of ambiguous cases that can support such qualitative analysis. Studying how confidence correlates with reliability may inform thresholding strategies. Finally, simulating multi-annotator crowds via LLM ensembles could approximate majority labels at scale. Such lines of work will clarify both the limits and the promise of LLMs as annotation assistants in culturally nuanced, inherently subjective tasks. Given the observed divergence between LLMs and the single expert adjudicator, future research should specifically examine the potential impact of adjudicator bias by employing multiple adjudicators or exploring crowdsourced adjudication methods. While this study treated humor as a binary class for clarity and feasibility, future iterations will aim to model humor more granularly, including diverse types and gradations of humorous intent.

Acknowledgments

We are grateful to student Klara Živičnjak for granting us early access to HRumor 1.0 prior to its public release.

This work was supported in part by the Croatian Science Foundation under the project number HRZZ-IP-2022-10-7697 (MWE-Cro: Multiword Expressions in Croatian – Lexicological, Computational Linguistic and Glottodidactic Approach).

Limitations

Several limitations apply. Humor is subjective; human annotators reached only modest agreement, capping attainable scores. Our binary labels oversimplify humor and may not generalize beyond Croatian tweets. We used a single English prompt with fixed few-shot examples and tested no alternative prompts or model versions. Evaluation focused on Cohen's κ; we omitted qualitative error analysis and other metrics. Finally, LLM behavior can shift across releases, so results hold only for the specific GPT-4 and Gemini 2.5 Flash APIs employed. High LLM–LLM agreement might reflect pattern replication rather than genuine understanding. Investigating the nature of this agreement lies outside our scope, as we approach LLMs as functional rather than cognitive systems.

Ethics Statement

All tweets were collected from the publicly available corpora Twitter-HBS 1.0. We intend to redistribute our annotated corpus once HRumor 1.0 is officially released, under an appropriate license to be determined. The corpus contains public social-media content that may include offensive language, but no sensitive personal data. Annotators consented to having their anonymized decisions published, and no demographic information about them is stored.

References

Manjot Bedi, Shivani Kumar, Md Shad Akhtar, and Tanmoy Chakraborty. 2023. Multi-Modal Sarcasm Detection and Humor Classification in Code-Mixed Conversations. IEEE Transactions on Affective Computing, 14(2):1363–1375.

Lee Boonstra. 2025. Prompt engineering.

Santiago Castro, Luis Chiruzzo, Aiala Rosá, Diego Garat, and Guillermo Moncecchi. 2018. A Crowd-Annotated Spanish Corpus for Humor Analysis. arXiv:1710.00477 [cs].

Youngjin Chae and Thomas Davidson. 2023. Large Language Models for Text Classification: From Zero-Shot Learning to Instruction-Tuning.

Jacob Cohen. 1960. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37–46.

Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing.

2023. Is GPT-3 a Good Data Annotator? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11173–11195, Toronto, Canada. Association for Computational Linguistics.

John Fields, Kevin Chovanec, and Praveen Madiraju. 2024. A Survey of Text Classification With Transformers: How Wide? How Large? How Long? How Accurate? How Expensive? How Safe? IEEE Access, 12:6518–6531.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120. arXiv:2303.15056 [cs].

Xingwei He, Zhenghao Lin, Yeyun Gong, A-Long Jin, Hang Zhang, Chen Lin, Jian Jiao, Siu Ming Yiu, Nan Duan, and Weizhu Chen. 2024. AnnoLLM: Making Large Language Models to Be Better Crowdsourced Annotators. In Yi Yang, Aida Davani, Avi Sil, and Anoop Kumar, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 165–190, Mexico City, Mexico. Association for Computational Linguistics.

Avery E Holton and Seth Lewis. 2011. Journalists, social media, and the use of humor on Twitter. *Electronic Journal of Communication*, 21.

J. Richard Landis and Gary G. Koch. 1977. An Application of Hierarchical Kappa-type Statistics in the Assessment of Majority Agreement among Multiple Observers. *Biometrics*, 33(2):363–374.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, 55(9):1–35.

Nikola Ljubešić and Peter Rupnik. 2022. The Twitter user dataset for discriminating between Bosnian, Croatian, Montenegrin and Serbian Twitter-HBS 1.0. *https://www.clarin.si/info/k-centre/*. Accepted: 2022-01-27T18:55:44Z.

Bogdan Ludusan. 2024. Obtaining Agreement for Conversational Laughter Function Annotation. In *Laughter and Other Non-Verbal Vocalisations Workshop 2024*, pages 10–12. ISCA.

Andrés Mendiburo-Seguel, Stéphanie Alenda, Thomas E. Ford, Andrew R. Olah, Patricio D. Navia, and Catalina Argüello-Gutiérrez. 2022. #funnypoliticians: How Do Political Figures Use Humor on Twitter? *Frontiers in Sociology*, 7.

Rada Mihalcea and Carlo Strapparava. 2005. Making computers laugh: investigations in automatic humor recognition. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing - HLT '05*, pages 531–538, Vancouver, British Columbia, Canada. Association for Computational Linguistics.

Piotr Mirowski, Juliette Love, Kory Mathewson, and Shakir Mohamed. 2024. A Robot Walks into a Bar: Can Language Models Serve as Creativity SupportTools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1622–1636, Rio de Janeiro Brazil. ACM.

Fred Philippy, Siwen Guo, Cedric Lothritz, Jacques Klein, and Tegawendé F. Bissyandé. 2025. Enhancing Small Language Models for Cross-Lingual Generalized Zero-Shot Classification with Soft Prompt Tuning.

Robert Gilmore Pontius Jr and Marco Millones. 2011. Death to Kappa: birth of quantity disagreement and allocation disagreement for accuracy assessment. *International Journal of Remote Sensing*, 32(15):4407–4429.

María Simarro Vázquez. 2016. Mecanismos de humor verbal en Twitter. *Caracteres: estudios culturales y críticos de la esfera digital*, 5(2):32–57.

Biao Zhao, Weiqiang Jin, Yu Zhang, Subin Huang, and Guang Yang. 2023. Prompt learning for metonymy resolution: Enhancing performance with internal prior knowledge of pre-trained language models. *Knowledge-Based Systems*, 279:110928.

Klara Živičnjak. 2025. *Karakteristike humora na hrvatskom X-u*. M.A. thesis, Filozofski fakultet u Zagrebu.

A Appendices

Appendix A Prompt Text

"You are a language expert annotating Croatian tweets. Your task is to classify each tweet as either HUMOROUS, NOT HUMOROUS, or MAYBE HUMOROUS.

Use MAYBE HUMOROUS if the intention is unclear, or if the tweet contains subtle irony that might not be perceived as humorous by all readers.

Humor can be expressed through wordplay, parody, satire, jokes, puns, exaggeration, or unexpected associations. Special attention should be paid to irony and sarcasm (where the meaning is opposite to what is said), which are frequent on social media.

Context matters: some tweets may refer to current events or pop culture. Use your best judgment to determine if the tweet was intended to be funny.

Briefly explain the reasoning behind your label in 1 short sentence (maximum 8 words).

After the label, provide a confidence score from 1 to 100, where 100 means "very confident".

Respond using this format:
Tweet ID: [ID]
Tweet: [text]
Explanation: [short sentence, max 8 words]
Label: [HUMOROUS / NOT HUMOROUS / MAYBE HUMOROUS]
Confidence: [1–100]

Tweets and explanations are in Croatian. Do not translate.

Below are 4 examples:

Tweet ID: A
Tweet: nikad nisi prestar za crtice
Explanation: zapažanje bez šale ili ironije
Label: NOT HUMOROUS

Tweet ID: B
Tweet: Ovo je bio jedan zaista lijep petak. #friends #carpediem
Explanation: iskrena objava, bez humora
Label: NOT HUMOROUS

Tweet ID: C
Tweet: punica - antidepresivi, zanimljiv tijek misli heh
Explanation: ironično povezivanje punice i lijekova
Label: HUMOROUS

Tweet ID: D
Tweet: Mile Fontana je omogućio Zagrebčanima rashlađivanje u svojim fontanama #Bandića za batmana!
Explanation: sarkastičan ton i hiperbola o fontanama
Label: HUMOROUS

Now classify the following tweets:”

Appendix B Cohen’s kappa agreement scores averaged across six subcorpora

Annotator and model pairings	Mean (SD)	Median	Min	Max	Range
A/B	0.2649 (±0.0744)	0.2396	0.2080	0.4115	0.2035
A/Adjudicator	0.6907 (±0.1380)	0.7208	0.4314	0.8216	0.3901
A/GPT	0.2832 (±0.1155)	0.3206	0.1154	0.4075	0.2921
A/Gemini	0.2891 (±0.0988)	0.3290	0.1456	0.3757	0.2301
B/Adjudicator	0.4988 (±0.1379)	0.4841	0.3238	0.6902	0.3664
B/GPT	0.2710 (±0.1359)	0.2485	0.1109	0.5195	0.4087
B/Gemini	0.2600 (±0.1469)	0.2374	0.1231	0.5127	0.3895
Adjudicator/GPT	0.2000 (±0.0484)	0.2046	0.1250	0.2586	0.1337
Adjudicator/Gemini	0.2098 (±0.0449)	0.2162	0.1264	0.2538	0.1275
GPT/ Gemini	0.6322 (±0.0254)	0.6258	0.5964	0.6697	0.0733

Appendix C Annotation results for 6000 tweets

	HUMOROUS	NOT HUMOROUS
Annotator A	1443 (24.05 %)	4557 (75.95%)
Annotator B	2040 (34.00 %)	3960 (66.00 %)
Adjudicator	993 (16.55%)	5007 (83.45 %)
GPT	3477 (57.95 %)	2523 (42.05 %)
Gemini	3293 (54.88 %)	2707 (45.12 %)

Appendix D Confusion matrices and F1 scores of human annotators and adjudicator vs. LLMs

Annotator A vs. GPT		
F1 score = 0.60	GPT HUMOROUS	GPT NOT HUMOROUS
Annotator A HUMOROUS	1288	155
Annotator A NOT HUMOROUS	2189	2368
Annotator B vs. GPT		
F1 score = 0.61	GPT HUMOROUS	GPT NOT HUMOROUS
Annotator B HUMOROUS	1604	436
Annotator B NOT HUMOROUS	1873	2087
Adjudicator vs. GPT		
F1 score = 0.53	GPT HUMOROUS	GPT NOT HUMOROUS
Adjudicator HUMOROUS	906	87
Adjudicator NOT HUMOROUS	2571	2436
Annotator A vs. Gemini		
F1 score = 0.61	Gemini HUMOROUS	Gemini NOT HUMOROUS
Annotator A HUMOROUS	1245	198
Annotator A NOT HUMOROUS	2048	2509

Annotator B vs. Gemini		
F1 score = 0.61	Gemini HUMOROUS	Gemini NOT HUMOROUS
Annotator B HUMOROUS	1515	525
Annotator B NOT HUMOROUS	1778	2182
Adjudicator vs. Gemini		
F1 score = 0.54	Gemini HUMOROUS	Gemini NOT HUMOROUS
Adjudicator HUMOROUS	878	115
Adjudicator NOT HUMOROUS	2415	2592

Appendix E Examples of agreement across annotators

Tweet (Original)	**Tweet (English)**	**A**	**B**	**Adj.**	**GPT**	**Gemini**
LoL RT @mala_planeta: Kad umres, ti ne znas da si umro i nije ti tesko. Tesko je drugima. Isto je ako si GLUP!	LoL RT @mala_planeta: When you die, you don't know you're dead and you don't feel bad. Other people feel bad. It is the same when you are STUPID!	Yes	Yes	Yes	Yes	Yes
Uzeo lak za kosu da ubijem pauka. Još je živ, al mu je frizura spektakularna	Got hairspray to kill a spider. It's still alive, but its hair is spectacular.	Yes	Yes	Yes	Yes	Yes
Dođe mi da plačem kad vidim koliko posla imam.	I want to cry when I see how much work I have.	No	No	No	No	No
Miley Cyrus Wrecking ball meni vrhunska stvar, da se ne lažemo.	Miley Cyrus Wrecking ball is a great song, let's be honest.	No	No	No	No	No
Teorije zavjere. Teorije zavjere everywhere.	Conspiracy theories. Conspiracy theories everywhere.	No	No	No	Yes	Yes
Neprijateljima treba neprekidno opraštati, jer je upravo to ono što ih najviše ljuti. #zivot	One should constantly forgive one's enemies because that's what upsets them the most. #life	No	No	No	Yes	Yes
Stigla novogodišnja čestitka u vidu update 4.3 za Note2.	New Year's card in the form of the 4.3 update for Note2 has arrived.	No	No	No	Yes	Yes
Jesmo mi bili već? #Olympics #ZOI #Sochi	Were we already on? #Olympics #ZOI #Sochi	No	No	No	Yes	No
I tak, sjedimo u birtiji, a stol do živa domaća glazba iz Irske. Cool. #karlovac #434rođendan	There we were, sitting in a pub, and the next table had live homemade Irish music. Cool. #karlovac #434rođendan	No	No	No	No	Yes
To je i ona rekla. ,	That's what she said	Yes	Yes	Yes	No	Yes
od stoljeća sedmog tu žive budale	Idiots have been living here since the seventh century	Yes	Yes	Yes	Yes	No
Sretna NG meni i mojima.	Happy New Year to me and mine	Yes	No	Yes	No	No