Evaluating the Cross-Domain Robustness of Sarcasm Detection Models: From Twitter to News Headlines

Wonwoo Choi
University of Maryland
wchoi12@umd.edu

Abstract

Sarcasm challenges sentiment analysis by inverting literal meaning, often leading to misclassifications. While transformer-based models perform well on sarcasm detection in social media, their robustness under domain shift remains unclear. This study evaluates a RoBERTa-based sarcasm detector trained on Twitter and tested on news headlines from HuffPost and The Onion. The model achieves high in-domain accuracy (95%) but drops to near chance performance (47%) on headlines, revealing strong domain sensitivity and systematic overprediction of sarcasm, particularly for neutral text. Few-shot fine-tuning with 50 headlines improves accuracy to 62%, indicating partial mitigation but limited transferability across domains.

1 Introduction

Oxford Language defines sarcasm as the use of irony to mock or convey contempt. An example of sarcasm is “Great, another Monday!” which appears to be positive on the surface but actually conveys negativity. Sarcasm has shaped how audiences interpret information, critique social norms, and engage with political and social issues. While mainstream news outlets, such as HuffPost, aims to report factual events, platforms like The Onion intentionally use humor, irony, and exaggeration to deliver current events. This strategy produces headlines resemble real news on the surface but differ in tone, sentiment, and communicative intent. For both human readers and computational systems, sarcastic headlines invert literal meaning in ways that are difficult to detect using surface-level lexical cues.

Understanding sarcasm is essential for accurate sentiment analysis. Traditional sentiment analysis models misclassify sarcastic text because positive or neutral lexical patterns mask the underlying negative sentiment. Misinterpreting sarcastic headlines can have a negative impact on downstream tasks, such as media analysis, opinion mining, and misinformation detection. Prior work shows that sarcasm disrupts sentiment modeling across domains, including business-oriented customer reviews (Tan et al., 2023) and political discourse on social media (McClain et al., 2024). Detecting sarcasm is crucial for accurate computational analysis and clear interpretation as factual reporting and satire become increasingly mixed in public communication.

This project investigates how sarcasm influences sentiment in news headlines by comparing two contrasting sources:

- **HuffPost** (a mainstream news outlet that publishes factual, non-sarcastic headlines.)
- **The Onion** (a satirical news website that intentionally produces humorous, exaggerated, and sarcastic versions of current events)

Using transformer-based models, this study aims to **(1) classify whether a headline is sarcastic** and **(2) identify how news platforms express sentiment toward political issues and current events.** By applying sarcasm detection and sentiment analysis to these parallel headline domains, the study examines both the linguistic features that signal sarcasm and the challenges sarcasm poses for modern sentiment classifiers.

2 Background

2.1 Sarcasm Detection

Understanding sentiment in news media is essential for analyzing how political issues and current events are framed for the public. However, there is still a persistent challenge for sentiment analysis for sarcasms, because it reverses the literal meaning of the text (Slim et al., 2024; Davidson et al., 2024). Prior study highlights that sentiment models constantly misclassify sarcastic content, leading to inaccurate conclusions (Razali et al., 2017).

Since news headlines frequently contain irony and sarcasm to critique political actors, detecting sarcasm becomes crucial for any downstream sentiment analysis pipeline.

One computational approach for sarcasm detection was rule-based. Riloff et al. (2013) detected sarcasm by using a rule-based classifier that looked for a positive verb and a negative situation. If found, sarcasm was detected and everything did not matter. However, the entire system in this paper is built on the assumption that sarcasm arises from the contrast of positive sentiment and a negative situation. This ignores many other forms of sarcasm, such as rhetorical questions ("Oh really? That's great"), understatement ("What a great surprise"), hyperbole ("Best day ever"), and contextual sarcasm that relies on world knowledge. This limits generalizability, especially to news headlines, where sarcasm rarely follows the stated structure.

Another approach used by researchers to detect sarcasm was a statistical approach. Tsur et al. (2010) introduce a semi-supervised algorithm to detect sarcasm in Amazon product reviews. The authors utilized kNN-style classification, classifying new sentences by similarity to labeled pattern vectors. The kNN style classification used in this paper is especially sensitive to pattern sparsity and irrelevant high-dimensional feature expansions. The authors noted that classification failed to distinguish between "This book was really good until page 2!" (sarcastic) and "This book was really good until page 430!" (non-sarcastic), demonstrating the weakness of shallow representations.

Finally, the most popular approach nowadays is using deep learning. Potamias et al. (2020) explore a transformer-based approach to irony and sarcasm detection, utilizing RCNN-RoBERTa, a pre-trained RoBERTa model. Their proposed model outperformed models like ELMo, XLnet, BERT-Cased and Uncased, and RoBERTa. However, the authors fail to provide a clear theoretical motivation for why sequential RCNN layers improve over RoBERTa alone, and it remains unclear whether improvements come from the RCNN or simply from increased model capacity.

2.2 Sentiment Analysis and Opinion Mining

There is a huge and growing volume of opinion-rich content online (reviews, blogs, forums), and people and increasingly depend on these opinions for decisions (Pang and Lee, 2008). The media plays an important role in shaping people's views and thoughts, which calls for the need foror appropriate opinion extraction of media platforms on political issues and current events (Agarwal et al., 2016). Sentiment analysis has become a popular method for analysis of social media discourse as sentiment scores reflect an objective means of assessing the mood of social media users, consumers, and the public at large (Puschmann and Powell, 2018), allowing text opinions to be converted to categorized labels, which can be aggregated, visualized, and modeled statistically.

Linguistic Inquiry and Word Count (LIWC) categories are derived from decades of psycholinguistic research, enabling direct interpretation in terms of psychological states (Santos and Vieira, 2017; Tausczik and Pennebaker, 2010). LIWC counts category words without modeling word order or context (Tausczik and Pennebaker, 2010), making it computationally inexpensive and suitable for large-scale sentiment-oriented analyses.

Although LIWC is an attractive tool for sentiment analysis, it has important limitations, such as ignoring discourse structure and contextual incongruity (Sajadi et al., 2025). Valence Aware Dictionary and sEntiment Reasoner (VADER) is another sentiment analysis method that retains LIWC's advantages, is more sensitive to social media sentiment, and generalizes better than LIWC across domains (Hutto and Gilbert, 2014).However, VADER still struggles with contextual understanding, particularly in cases involving complex linguistic phenomena or domain shift (Garg et al., 2023).

In recent years, transformer-based approaches to sentiment analysis have become increasingly popular, as they outperform traditional lexicon-based and deep learning methods in capturing nuanced sentiments (Davoodi and Mezei, 2022; Gowda et al., 2025). Specifically, pre-trained models on GoEmotions, instead of just polarity, provide fine-grained emotion classification. This allows the models to capture subtle, mixed emotions within text (Wang et al., 2024) that LIWC or VADER reduces to a single polarity number.

2.3 Present Study

Prior works have successfully demonstrated identification of sarcasm, but limitations remain regarding generalizability beyond the training dataset. Most models being evaluated only on social media text, raises questions about whether they can detect sarcasm in structurally different settings such as news headlines. This limitation becomes particularly consequential for downstream tasks such as sentiment analysis and opinion mining. When news headlines employ sarcasm to critique political actors or frame current events, sentiment polarity is intentionally inverted, and conventional sentiment models frequently misinterpret these cues. Without robust sarcasm detection, sentiment analysis pipelines can yield misleading assessments of media tone and political framing.

Given this gap, the present study examines the extent to which transformer-based sarcasm detection model, trained on Twitter data, can generalize to sarcastic news headlines. By evaluating the model under domain shift, this work provides empirical evidence about the reliability of sarcasm-aware sentiment analysis for opinion mining in news media, where accurate interpretation of tone and stance is essential.

3 Data

3.1 Training Data

For this project, I trained and evaluated the sarcasm detection model using a publicly available sarcasm dataset hosted on Hugging Face and collected by Malik (2021) from Twitter. An example of the dataset is provided in Appendix A (Table A1). The dataset consists of short tweets labeled for sarcasm and is well suited for fine-tuning transformer-based classification models such as RoBERTa.

The dataset contains 5,000 examples, each represented by two fields:

- **Tweet**: a short tweet
- **Sarcasm (yes/no)**: a textual sarcasm indicator (“yes” for sarcastic, “no” for non-sarcastic)

Because the sarcasm label was originally provided as a string, the only preprocessing step required before modeling was converting these string labels into a binary numerical format:

- **yes** → 1
- **no** → 0

49.6% of the dataset was labeled as ‘no’ and 50.4% as ‘yes,’ indicating that the dataset is effectively balanced across the two classes.

No additional cleaning, such as stopword removal or punctuation stripping, was necessary as the dataset was already in clean form, and RoBERTa operates directly on raw text through subword tokenization.

3.2 Evaluation Data

News Headlines Dataset for Sarcasm Detection (Misra and Grover, 2021; Misra and Arora, 2023), publicly available on Kaggle, was used for model evaluation. An example of the dataset is provided in Appendix A (Table A2). Unlike prior work, such as Tan et al. (2023), which uses this dataset as a primary training source, the dataset was used to test the model’s performance on sarcastic versus non-sarcastic news headlines. The News Headlines dataset for Sarcasm Detection is collected from two news websites:

- **HuffPost**: literal news headlines
- **The Onion**: satirical news headlines

The news headlines dataset contains 26,709 rows, 14,985 not sarcastic headlines, and 11,724 sarcastic headlines as ground truth labels, each represented by three fields:

- **article_link**: URL link to the original article
- **headlines**: a textual news headline
- **is_sarcastic**: a binary label (1 = sarcastic, 0 = non-sarcastic)

No other cleaning process was required other than dropping the article links because they provide no linguistic information relevant to sarcasm detection.

4 Methods

4.1 Preprocessing

Minimal preprocessing was required because RoBERTa operates directly on raw text using subword tokenization. The only preprocessing step applied was converting the sarcasm labels in the training dataset from string format (“yes”/”no”) to a binary numerical format (1/0). No additional normalization, such as lowercasing, stopword removal, or punctuation stripping, was needed, as transformer-based models do not require these steps and can remove useful features relevant to detecting sarcasm. The dataset was randomly divided into training and test splits using scikit-learn train_test_split() using an 80/20 ratio with a random state of 42 to ensure reproducibility.

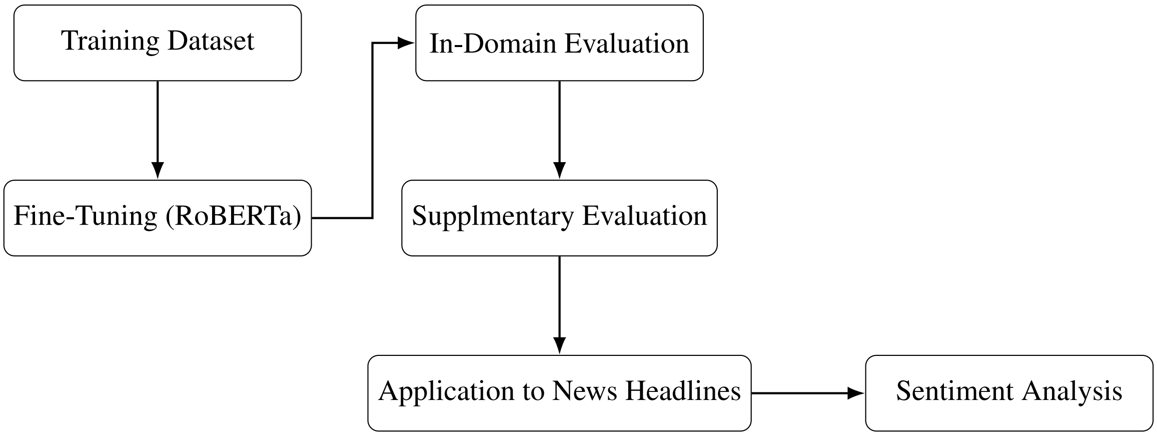


Figure 1: Methodology pipeline for training, evaluating, and applying the sarcasm detection model for downstream sentiment analysis.

For the evaluation dataset, article URLs were discarded as they provide no linguistic information relevant to sarcasm detection.

4.2 Model Architecture

The sarcasm detection model is based on RoBERTa-base, a 12-layer transformer architecture pretrained on large-scale English corpora. RoBERTa-base consists of:

- 12 transformer encoder layers
- 12 self-attention heads per layer
- 768-dimensional hidden representations
- Byte-Pair Encoding (BPE) tokenizer

RoBERTa computes contextualized token embeddings through multi-head self-attention, allowing the model to capture the semantic incongruity characteristic of sarcastic text.

A task-specific classification head was automatically added by calling `AutoModelForSequenceClassification`, which consists of a dropout layer ($p = 0.1$), followed by a linear projection that maps the final classification embedding to two output logits. A softmax function is applied during inference to convert logits into class probabilities.

4.3 Training Procedure

Fine-tuning was conducted using the Hugging Face Transformer library. The model was trained on the Twitter sarcasm dataset with the following hyperparameters:

- Learning rate = 2e-5
- Batch size = 16
- Epochs = 3
- Weight decay = 0.01
- Loss function = Cross-entropy loss

Training was performed in a GPU-enabled environment to accelerate the computation, and a fixed random seed of 42 was used for reproducibility.

4.4 Supplementary Evaluation

Before conducting the out-of-domain evaluation on news headlines, a supplementary robustness check using 30 GPT-generated examples was conducted. Prompt and GPT-generated examples are provided in Appendix B (Table 1). These examples included sarcastic, non-sarcastic, and intentionally challenging borderline cases. This synthetic evaluation served as an internal sanity check to verify that the model behaved consistently on difficult or ambiguous inputs prior to testing its generalization to a real-world out-of-domain dataset.

4.5 Cross Domain Evaluation

The goal of this study is to evaluate whether a sarcasm detection model trained solely on Twitter data can generalize to a distinct domain: news headlines. This requires a strict out-of-domain (OOD) evaluation, where the model is evaluated on text that differs from the training domain in:

- Writing styles: informal and conversational vs. formal and concise
- Sarcasm expression: explicit vs. implicit
- Linguistic cues: social media markers vs. editorialized phrasing

The model trained on Twitter was directly applied to the News Headline dataset described in the Data section above, without any additional fine-tuning or adaptation. This evaluates whether the transformer-based sarcasm detection model retains its performance when exposed to data from a different domain, which is important for real-world applications.

4.6 Sentiment Analysis

To examine the emotional tones of news headlines and assess the stance of news platforms on social issues and current events, a transformer-based emotion classification model was applied to the headline dataset. Sentiment was inferred using the pretrained RoBERTa-based model by Lowe (2024), a state-of-the-art classifier fine-tuned on Google's GoEmotion corpus of 58,000 Reddit comments spanning 28 distinct emotions plus neutral category .

Headlines were processed in batches for computational efficiency. For each headline, the model provided a probability distribution over all GoEmotions labels, and the emotions with the highest confidence were selected as the predicted category.

The original goal of applying emotion classification to news headlines was to explore the emotional patterns expressed in sarcastic and non-sarcastic news headlines, with the intention of examining how sarcasm might shape sentiment towards social or political issues. However, because the cross-domain evaluation revealed significant misclassification when the sarcasm detector was applied to the news headlines (as discussed in the Results section), a secondary analysis was conducted to investigate how emotion categories relate to sarcasm misclassification. Specifically, the emotion distribution of ground truth labels and predicted labels was compared to identify which emotional expressions were most associated with prediction errors.

5 Results

5.1 In-Domain Performance

To evaluate performance on data similar to the training distribution, the model was tested on the 20% of the Twitter sarcasm dataset.

The model performs well on 20% of the Twitter dataset, with an overall accuracy of 0.95. This indicates that RoBERTa successfully learned to distinguish between sarcastic and non-sarcastic tweets within the training domain. The precision is high for both classes, 0.91 for not_sarcastic, and 1 for sarcastic tweets. This means when the model predicts sarcasm, it is almost always correct, and there are very few false positives. There is a slight imbalance in recall between classes. The model identified all non-sarcastic tweets correctly, while it missed some sarcastic tweets (false negatives) with 0.89 recall. This shows that sarcastic examples are slightly harder for the model to detect than non-sarcastic ones.

Class	**Precision**	**Recall**	**F1-score**	**Support**
not_sarcastic	0.91	1.00	0.96	537
sarcastic	1.00	0.89	0.94	463
Accuracy		0.95 (N = 1000)		
Macro Avg	0.96	0.95	0.95	1000
Weighted Avg	0.95	0.95	0.95	1000

Table 1: Classification report for the RoBERTa sarcasm detection model evaluated on the 20% Twitter test set.

Overall, the accuracy is very high, indicating that the model fits the Twitter dataset well and captures strong sarcasm patterns common in social media.

5.2 GPT-generated Examples

The model performed well on the 30 GPT-generated out-of-domain examples. It correctly classified all the non-sarcastic sentences and 9 out of 10 sarcastic sentences with high confidence, indicating that RoBERTa generalizes well to clean, well-formed text that expresses sarcasm through explicit lexical cues. Even among the "hard" ambiguous cases, the model correctly identified most instances of indirect sarcasm such as "Well, that presentation was… something," or "He's a real genius, isn't he?" The only error occurred in cases where sarcasm was extremely subtle or required pragmatic inference, such as "I'm sure this will end well."

These results suggest that the model is robust to linguistic variations of sarcasm when the examples resemble the stylistic patterns of the training domain.

5.3 Cross Domain Evaluation

The out-of-domain evaluation results (Table 2) show that the fine-tuned RoBERTa sarcasm detector struggles to generalize from Twitter data to news headlines (see Appendix D for sample misclassified headlines). The classifier achieves an overall accuracy of 47.23%, which is close to random chance given the balanced label distribution. Performance notably differs across classes; for non-sarcastic headlines, precision is 0.54 but recall is 0.39, indicating that the model frequently fails to identify true non-sarcastic cases and tends to misclassify them as sarcastic. On the other hand, for sarcastic headlines, the model attains a higher recall (0.58), but lower precision (0.43). This hints that the model has built an overprediction bias towards sarcasm. The macro-averaged F1-score of 0.47 further confirms that neither class is modeled well. These results collectively demonstrate that the model trained on Twitter sarcasm do not transfer effectively to the linguistic and stylistic patterns of news, highlighting the significance of domain shift between social media and journalistic text.

Class	Precision	Recall	F1-score	Support
not_sarcastic	0.54	0.39	0.45	14985
sarcastic	0.43	0.58	0.49	11724
Accuracy		0.47 (N = 26709)		
Macro Avg	0.48	0.48	0.47	26709
Weighted Avg	0.49	0.47	0.47	26709

Table 2: Classification report for the RoBERTa sarcasm detection model evaluated on the news headline dataset.

The confusion matrix further reveals that the model shows a strong bias toward predicting sarcasm when applied to news headlines. Out of the 14,985 non-sarcastic headlines, only 5,813 were correctly classified (true negatives), while 9,172 were incorrectly labeled as sarcastic (false positives). This indicates significant over-prediction of sarcasm in the new domain.

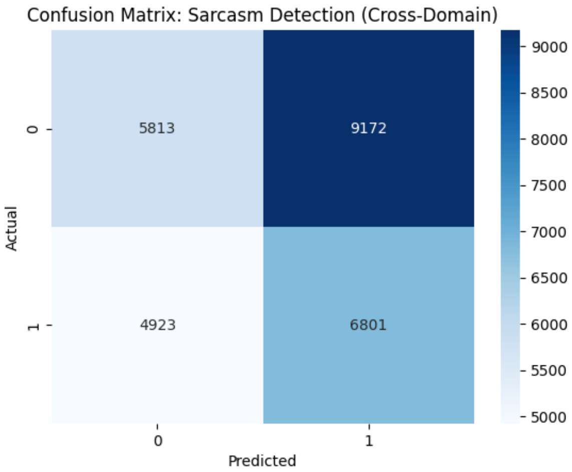


Figure 2: Confusion matrix for sarcasm detection on the News Headlines dataset.

For sarcastic headlines, the model correctly identified 6,801 cases (true positives) but failed on 4,923 cases (false negatives), reflecting a limited ability to detect satire or implicit sarcasm expressed in news headlines.

Overall, these patterns demonstrate that the Twitter-trained model struggled to generalize to the headline domain, where sarcasm cues are subtler, less structural, and less explicit compared to social media. The high false-positive rate suggests that indicators of sarcasm learned from Twitter were incorrectly triggered by news headlines indicating that Twitter-specific sarcasm cues do not generalize to news headlines.

5.4 Headline Sentiment Analysis

Looking at Figure C2 (Appendix C), most news headlines were annotated with a neutral emotional label. The model misclassified 3,068 neutral headlines as sarcastic, representing the single largest source of false positives. This indicates that the model struggles to recognize lack of emotion, a characteristic typical of news headlines but uncommon in Twitter data, where sarcasm is often accompanied by strong affective cues. To examine whether this pattern persisted across other emotions, neutral headlines were removed from the analysis. Figure 3 presents the distribution of emotion labels across ground-truth sarcasm categories and RoBERTa's predictions. Enlarged figure is provided in Appendix C (Figure C2). A clear trend emerges: the Twitter-trained model systematically overpredicts sarcasm for several emotional categories. The most pronounced discrepancy appears in the curiosity category, where only 58 headlines were truly sarcastic, yet the model predicted 509 as sarcastic. Similar inflation patterns occur for emotions such as approval, disapproval, admiration, and annoyance, each showing a substantial increase in predicted sarcastic cases relative to ground truth. These findings suggest that the model relies heavily on emotional cues learned from Twitter, where such patterns signal sarcasm, but these cues do not generalize to the stylistic norms of news media.

6 Discussion

The model's poor generalization to news headlines appears to stem from a substantial domain mismatch between the training and evaluation datasets. Twitter sarcasm often includes explicit

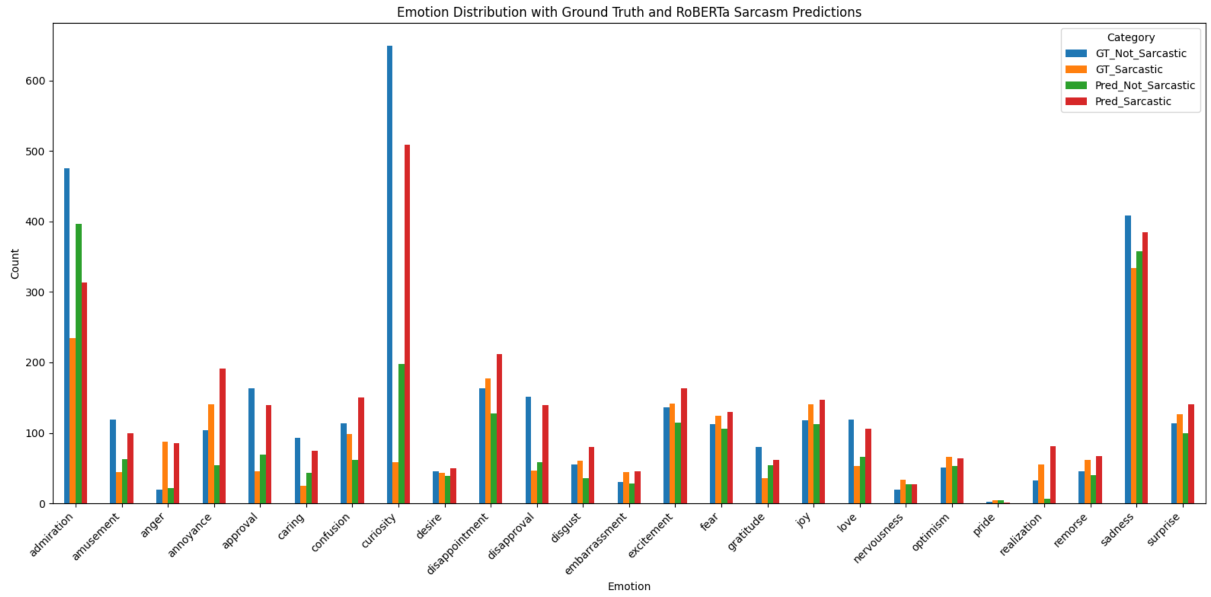


Figure 3: Emotion distributions across ground-truth and predicted sarcasm labels.

Metric	Baseline RoBERTa	Few-Shot Fine-Tuned	Improvement
Accuracy	0.47	0.62	+0.15
Precision (Macro)	0.47	0.61	+0.14
Recall (Macro)	0.47	0.61	+0.14
F1-Score (Macro)	0.47	0.61	+0.14
Non-sarcastic F1	0.49	0.67	+0.18
Sarcastic F1	0.45	0.56	+0.11

Table 3: Comparison of baseline RoBERTa model and few-shot fine-tuned model on out-of-domain news headline sarcasm detection.

cues such as exaggeration, conversational markers, or rhetorical constructions that RoBERTa learns effectively during fine-tuning. In contrast, sarcastic news headlines rely on subtle, stylistic, or context-dependent cues, rather than lexical indicators. As a result, the model overfits to Twitter-specific sarcasm patterns and incorrectly predicts a large proportion of headlines as sarcastic. This problem is reflected clearly in the emotion distribution analysis: the model marked more than 12,000 neutral headlines as sarcastic and substantially inflated sarcastic predictions for emotions such as curiosity, admiration, annoyance, and approval. These errors suggest that RoBERTa relies heavily on emotional cues that reliably signal sarcasm on Twitter but do not function the same way in journalistic writing.

In addition, RoBERTa lacks the knowledge of the world and contextual reasoning. Sarcastic news headlines often reference political events, policy debates, or broader social issues, requiring background knowledge to interpret correctly. Because the model lacks an understanding of these contexts, it continues to rely on shallow lexical heuristics rather than conceptual reasoning, which produces large clusters of false positives. Conversely, when RoBERTa is fine-tuned directly on the News Headlines dataset, the opposite pattern emerges: the model becomes overly conservative and classifies most GPT-generated examples as non-sarcastic. This demonstrates that each dataset encodes sarcasm through fundamentally different mechanisms: Twitter through explicit, conversational sarcasm cues and news headlines through more formalized, context-dependent satire. Consequently, models trained on one domain fail to generalize to the other, either by overpredicting sarcasm or underpredicting it. These results reinforce that sarcasm detection is highly domain sensitive and that sarcasm cues do not transfer cleanly across genres or platforms.

To examine whether even minimal domain-specific supervision could mitigate this mismatch, I applied few-shot fine-tuning. Few-shot fine-tuning with 50 headlines (25 per class) resulted in a modest improvement in out-of-domain performance (accuracy increased from 47.23% to 62%) **(Table 3)**. The implementation details of few-shot fine-tuning are omitted for brevity, as the experiment served primarily as a supplementary diagnostic to evaluate whether limited in-domain supervision could mitigate the domain shift. Recall for non-sarcastic headlines improved slightly, suggesting that the additional domain-aligned samples helped the model better calibrate its decision boundary. However, the gains remained limited because the few-shot dataset provides only a weak supervisory signal relative to the large domain shift. With just 50 labeled examples, the model receives only minimal exposure to the linguistic structures and contextual cues that characterize sarcastic headlines, making it insufficient to reshape the Twitter-trained representations in a meaningful way. This indicates that while few-shot fine-tuning can nudge the model toward the target domain, substantially larger or more diverse in-domain datasets are required to achieve robust transfer.

7 Conclusion

This study evaluated the cross-domain robustness of a transformer-based sarcasm detection model trained on Twitter and tested on news headlines. While the model achieved strong performance on in-domain Twitter data (95% accuracy), its performance dropped to near chance levels when applied to sarcastic and non-sarcastic news headlines (47% accuracy), indicating a substantial domain shift. As discussed above, error patterns, such as systematic overprediction of sarcasm, demonstrates that the model relied heavily on Twitter-specific sarcasm cues that do is not present in concise and editorial stylistic headlines. These findings reinforces that sarcasm is highly domain sensitive and that surface level lexical cues learned from social media are not sufficient for generalizing in headline datasets.

Few-shot fine-tuning moderately improved performance (to 62%), partially mitigating the cross-domain degradation. This improvement suggests that the performance drop is indeed driven by domain mismatch, and that even small amounts of in-domain supervision can help recalibrate the model. However, the limited gains also indicate that more robust transfer will require substantially larger or more diverse headline-specific datasets, or alternative modeling approaches that incorporate contextual or world knowledge.

Future work should explore more robust domain adaptation strategies to improve sarcasm detection across heterogeneous text genres. Incorporating larger and more diverse headline-specific training data could help models learn the subtle and context-dependent cues characteristic of journalistic sarcasm. Additionally, multi-domain or multi-task learning frameworks that jointly model sarcasm and sentiment across social media and news domains may improve generalization. Integrating external knowledge sources, such as political or world-event context, could further enhance interpretation of satirical headlines. Finally, evaluating larger language models and prompt-based approaches may provide insights into whether implicit reasoning and contextual understanding can mitigate domain sensitivity in sarcasm detection.

Limitations

This study has several limitations that should be considered when interpreting the findings.

First, this study did not incorporate contextual or world-knowledge reasoning, which is often necessary for interpreting satire or political irony in news headlines. Sarcastic headlines from sources like The Onion frequently rely on external events, cultural references, or political knowledge that transformer-based classifiers cannot infer from text alone. Without grounding in real-world context, the model relies on shallow lexical heuristics, contributing to false positives and false negatives.

Second, the few-shot fine-tuning experiment, while showing moderate improvement (Table 3), used only 50 labeled headlines. This limited sample is insufficient to represent the full stylistic and semantic range of sarcastic news writing. While few-shot results suggest the model can adapt with minimal supervision, the extent of improvement is constrained by the scarcity and narrowness of the added data.

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.

Davidson, Dr.D.Ravindran, and R.Anne Pratheeba. 2024. Enhancing sentiment analysis accuracy amidst sarcasm challenges with aspect-based machine learning for detection.

Laleh Davoodi and József Mezei. 2022. A comparative study of machine learning models for sentiment analysis: Customer reviews of e-commerce platforms. In *Proceedings of the 35th Bled eConference*, 27. AIS Electronic Library (AISeL).

Simran Garg, Devang Chaturvedi, and Tanya Jain. 2023. Sentiment Analysis: Techniques, Limitations, and Case Studies in Data Extraction and Classification.

Kumar Puttaswamy Gowda, Rabins Porwal, Cindhe Ramesh, Shashank Shekhar Tiwari, Kriti Srivastava, R. Rambabu, and S Govinda Rao. 2025. Transformers in Sentiment Analysis: A Paradigm Shift in Deep Learning Research. *Journal of Information Systems Engineering and Management*, 10(5s):262–280.

C. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.

Sam Lowe. 2024. roberta-base-go_emotions (revision 58b6c5b).

Nikesh Malik. 2021. Sarcasm detection dataset. Accessed: 2025-02-10.

Colleen McClain, Monica Anderson, and Risa Gelles-Watnick. 2024. How Americans Navigate Politics on TikTok, X, Facebook and Instagram.

Rishabh Misra and Prahal Arora. 2023. Sarcasm detection using news headlines dataset. *AI Open*, 4:13–18.

Rishabh Misra and Jigyasa Grover. 2021. *Sculpting Data for ML: The first act of Machine Learning*.

Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. *Foundations and Trends® in Information Retrieval*, 2(1–2):1–135.

Rolandos Alexandros Potamias, Georgios Siolas, and Andreas Georgios Stafylopatis. 2020. A transformer-based approach to irony and sarcasm detection. *Neural Computing and Applications*, 32(23):17309–17320.

Cornelius Puschmann and Alison Powell. 2018. Turning words into consumer preferences: How sentiment analysis is framed in research and the news media. *Social Media + Society*, 4(3).

Md Saifullah Razali, Alfian Abdul Halin, Noris Mohd Norowi, and Shyamala C. Doraisamy. 2017. The importance of multimodality in sarcasm detection for sentiment analysis. In *2017 IEEE 15th Student Conference on Research and Development (SCOReD)*, pages 56–60, Putrajaya. IEEE.

Ellen Riloff, Ashequl Qadir, Prafulla Surve, Lalindra De Silva, Nathan Gilbert, and Ruihong Huang. 2013. Sarcasm as Contrast between a Positive Sentiment and Negative Situation. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 704–714, Seattle, Washington, USA. Association for Computational Linguistics.

Amirali Sajadi, Kostadin Damevski, and Preetha Chatterjee. 2025. Psycholinguistic analyses in software engineering text: A systematic literature review. *Preprint*, arXiv:2503.05992.

Henrique Santos and Renata Vieira. 2017. PLN-PUCRS at EmoInt-2017: Psycholinguistic features for emotion intensity prediction in tweets. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 189–192, Copenhagen, Denmark. Association for Computational Linguistics.

Salwa Slim, Amal Aboutabl, and Ahmed Yacoub. 2024. A survey of sentiment analysis and sarcasm detection: Challenges, techniques, and trends. *International journal of electrical and computer engineering systems*, 15:69–78.

Yik Yang Tan, Chee-Onn Chow, Jeevan Kanesan, Joon Huang Chuah, and YongLiang Lim. 2023. Sentiment Analysis and Sarcasm Detection using Deep Multi-Task Learning. *Wireless Personal Communications*, 129(3):2213–2237.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54.

Oren Tsur, Dmitry Davidov, and Ari Rappoport. 2010. ICWSM — A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. *Proceedings of the International AAAI Conference on Web and Social Media*, 4(1):162–169.

Kaipeng Wang, Zhi Jing, Yongye Su, and Yikun Han. 2024. Large language models on fine-grained emotion detection dataset with data augmentation and transfer learning. *Preprint*, arXiv:2403.06108.

A Dataset Details

Text	Sarcastic
I genuinely love musicians.	no
I find vegetarians to be wonderful.	no
Because artists really make everything better.	yes
Can’t wait for more of the game.	yes
Oh great, another artists. Just what I needed.	yes

Table A1: Example sentences from the sarcasm annotation dataset.

Headline	Sarcastic
former versace store clerk sues over secret ’b...	0
the ’roseanne’ revival catches up to our thorn...	0
mom starting to fear son’s web series closest ...	1
boehner just wants wife to listen, not come up...	1
j.k. rowling wishes snape happy birthday in th...	0

Table A2: Example headlines from the sarcasm detection dataset (article links omitted for simplicity).

B GPT-generated Examples

Prompt: Create short examples of easily distinguishable non-sarcastic examples, easily distinguishable sarcastic examples, and hard to detect sarcasm examples, 10 of each.

Category	Example Sentence
Non-Sarcastic	I'm excited to start my new job next week.
	The weather is beautiful today.
	I finally finished my project on time.
	She did a great job organizing the event.
	I love spending weekends with my family.
	The teacher explained the concept very clearly.
	I really appreciate your help.
	The restaurant served delicious food.
	My flight arrived earlier than expected.
	I enjoy reading books in the evening.
Sarcastic	Oh great, another meeting that could've been an email.
	Yeah, because waking up at 5 AM is exactly what I wanted.
	Fantastic, my phone died right before the interview. Perfect timing.
	I absolutely love waiting in line for an hour.*
	Amazing, another update that breaks everything.
	Sure, I'll totally drop everything just to fix your mistake.
	Wow, what a surprise, the Wi-Fi is terrible again.
	Yeah, because that's exactly how physics works.
	Perfect, rain on the one day I planned a picnic.
	Oh look, the printer stopped working again. Shocking.
Hard Case	Well, that presentation was… something.
	Nice job. Really. Amazing.
	He's a real genius, isn't he?
	Guess I should've expected that from him.
	Love it when people talk over me.
	What a productive day. I did absolutely nothing.
	Oh awesome, they changed the deadline again.
	Yeah, because that's never happened before.
	I'm sure this will end well.*
	Great, now I get to redo all my work.

Table B1: Example sentences generated by ChatGPT 5.1 for non-sarcastic, sarcastic, and hard-to-classify cases. * represents sentences that was misclassified by the model.

C Figure

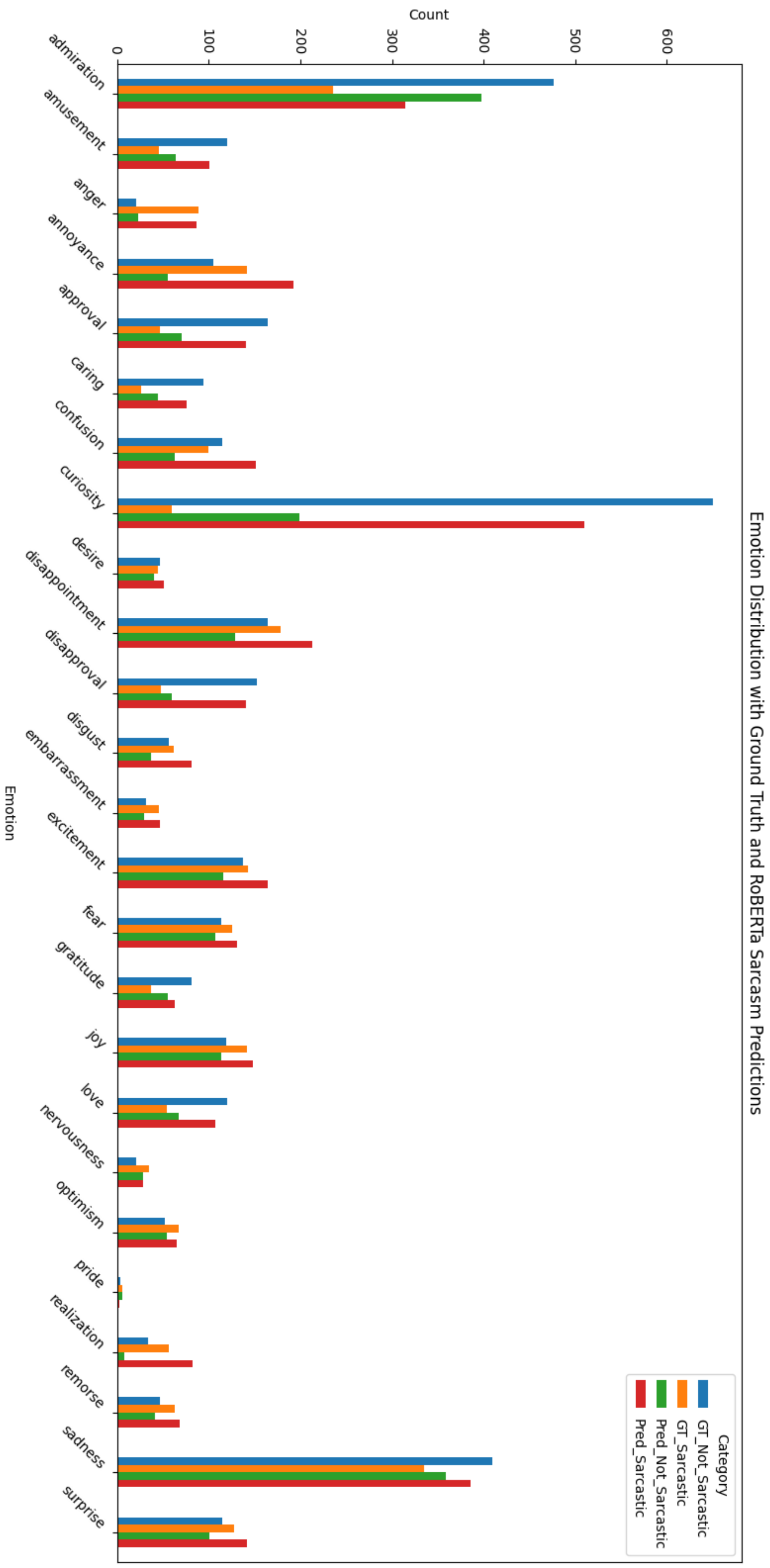
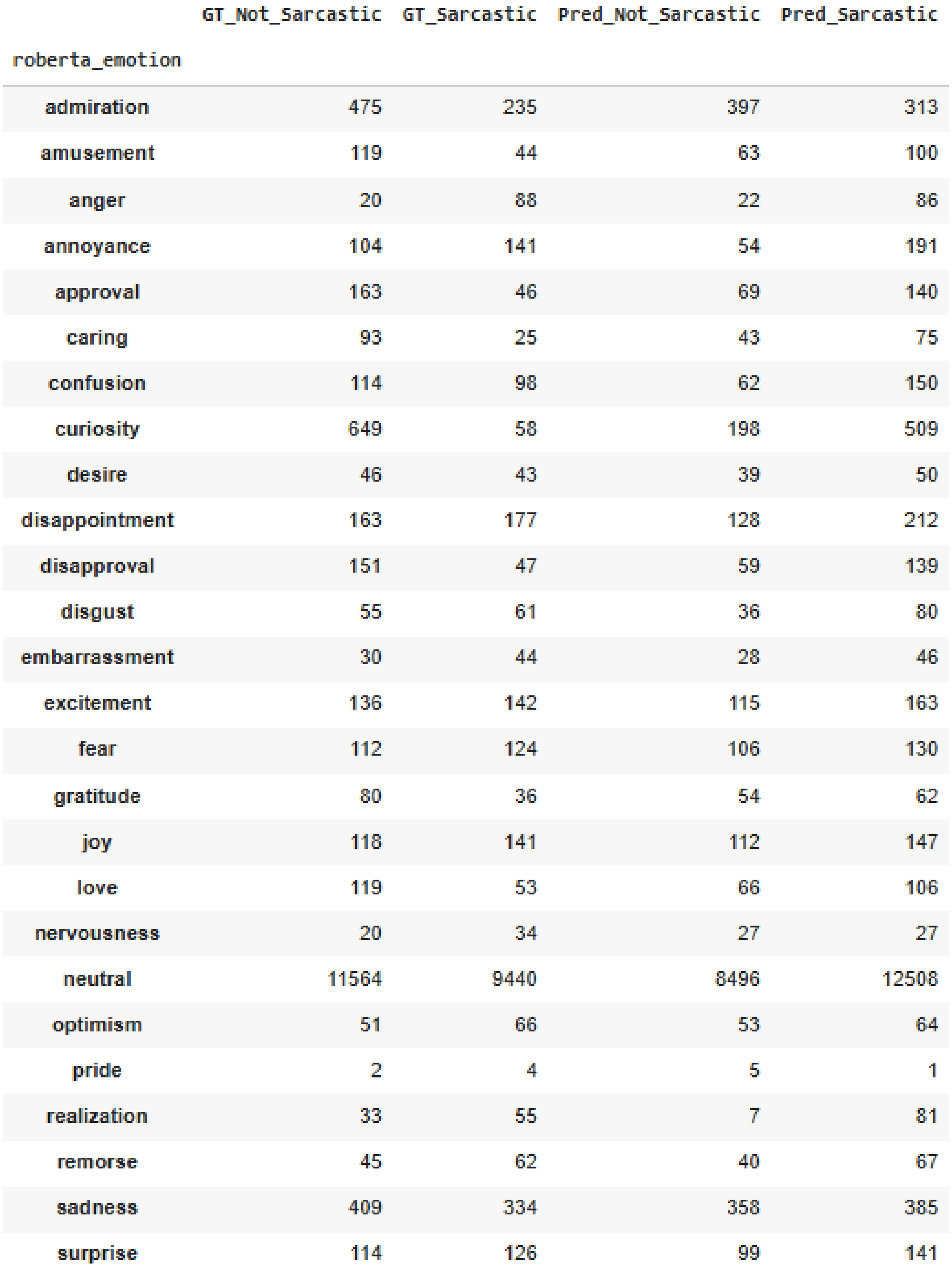


Figure C1: Emotion distributions across ground-truth and predicted sarcasm labels.



	GT_Not_Sarcastic	GT_Sarcastic	Pred_Not_Sarcastic	Pred_Sarcastic
roberta_emotion				
admiration	475	235	397	313
amusement	119	44	63	100
anger	20	88	22	86
annoyance	104	141	54	191
approval	163	46	69	140
caring	93	25	43	75
confusion	114	98	62	150
curiosity	649	58	198	509
desire	46	43	39	50
disappointment	163	177	128	212
disapproval	151	47	59	139
disgust	55	61	36	80
embarrassment	30	44	28	46
excitement	136	142	115	163
fear	112	124	106	130
gratitude	80	36	54	62
joy	118	141	112	147
love	119	53	66	106
nervousness	20	34	27	27
neutral	11564	9440	8496	12508
optimism	51	66	53	64
pride	2	4	5	1
realization	33	55	7	81
remorse	45	62	40	67
sadness	409	334	358	385
surprise	114	126	99	141

Figure C2: Exact counts of emotional distribution across ground-truth and predicted sarcasm labels.

D Error Analysis

Index	Headline	True	Pred
0	former versace store clerk sues over secret ‘black code’ for minority shoppers	0	1
1	the ‘roseanne’ revival catches up to our thorny political mood, for better and worse	0	1
4	j.k. rowling wishes snape happy birthday in the most magical way	0	1
9	friday’s morning email: inside trump’s presser for the ages	0	1
10	airline passengers tackle man who rushes cockpit in bomb threat	0	1

Table D1: Example misclassified headlines (article links omitted for simplicity).