Explainable Transformer Models for Sarcasm Detection on a News Headline Dataset

Chris Haleas
Indiana University
chaleas@iu.edu

Grayson Pettit
Indiana University
grpettit@iu.edu

1 Background & Motivation

In our project, we will explore the explainability of transformer-based neural networks for sentence-level sarcasm detection on a news headline dataset.

Inspiration for our project first came from our 'Computation and Linguistic Analysis' class material. In Chapters 6 and 7 of Jurafsky and Martin's *Speech and Language Processing*, we were introduced to the concepts of vector semantics, embeddings, and neural networks (2025). We learned that word meanings can be represented as points in a multidimensional space, and that similar meanings cluster together within this space, with their semantic relationships captured through embedding techniques to reveal deeper patterns in language. Those vectors in multidimensional space are learned by neural networks. The neural networks can then make predictions based on those vectors, passing input through layers of nodes, ultimately producing a single output prediction. Transformer-based neural networks are a class of architectures that are especially effective for language tasks, which is why we chose to explore them for our project.

Sarcasm is a form of speech in which an utterance is intended to convey a meaning opposite to what is literally said, typically to mock or criticize someone or something. In other words, sarcasm involves a mismatch between what is said and what is actually meant. Transformer-based neural networks may struggle with sarcasm because it effectively inverts meaning, and the trigger for this inversion is often based on implicit context. Negating a sentence by inserting a clear token such as "not" or an "un-" morpheme is relatively easy for a model to detect, but incorporating sentence-wide and sentence-external context introduces significant complexity. For this reason, sarcasm detection is a challenging task in computational linguistics, which is why we chose to explore it specifically.

Our study aims to compare how well different transformer models handle sarcasm detection, with a focus on identifying which words contribute most to their predictions. We investigate whether transformer embeddings capture the relationships between words within a sentence and can detect sarcasm when there is a mismatch between the literal meaning and the intended meaning. Many papers have utilized transformers for sarcasm detection previously. For example, Wen and Rezapour introduced a transformer model for sarcasm detection, which was tested on multiple sarcasm datasets and achieved accuracies of up to 98 percent (2025). We will conduct similar experiments by training different transformer models on a news headlines dataset to compare their accuracy. We will take this a step further by applying explainable artificial intelligence to analyze the sarcasm detection decisions of our transformer models, a method that, to our knowledge, has been largely overlooked in previous studies on transformers for sarcasm detection. It is important to understand whether language models like transformers capture the true meaning of sentences through their complex networks of neural layers, rather than simply classifying based on patterns or prior examples, so that we can determine whether they truly grasp language nuances, which we will investigate through explainability. Visani explains that LIME is a powerful tool for explainability in machine learning models by running input variations through the model and fitting a simple linear model to approximate the complex model's behavior (2020). This approach will help us successfully identify which linguistic features each transformer prioritizes when making sarcasm detection decisions.

2 Key Technological Focuses

The key technologies in our research are explainable artificial intelligence and transformer-based neural networks.

2.1 Explainability

Explainability in AI, or XAI, refers to methods that make the internal decision-making processes of neural networks transparent and interpretable. Rather than treating models as black boxes, explainability allows developers, researchers, and even non-technical individuals to understand which features or patterns most influence a model's predictions. If a developer can identify the most critical factors affecting outputs, they can more effectively evaluate and adjust the model. his can be especially helpful in fields where it can improve human outcomes, such as mitigating bias in criminal justice or explaining previously unknown diagnostic markers in healthcare (Burkart and Huber, 2021). Beyond these applications, explainability can improve model performance across any domain by revealing which aspects of the training data, such as gaps, biases, or over-representations, lead to unexpected predictions.

For our research, we use the Local Interpretable Model-agnostic Explanations (LIME) package. LIME reconstructs the predictions of machine learning models by building an interpretable model around the outputs of the original model (Ribeiro et al., 2016). It creates many variations of the original input and feeds them through the model to observe changes in predictions, assigning higher weights to samples that remain most similar to the original input. Using these weighted samples, LIME trains a linear model that approximates how the original model behaves in the local neighborhood of that specific input. The coefficients of this linear model indicate which words contribute most to the model's prediction, allowing us to interpret the factors driving the model's decisions.

2.2 Transformer-Based Neural Networks

Transformer models are a type of neural network architecture that capture context through self-attention mechanisms. This mechanism allows the model to consider all the words in a sentence and determine which ones are most relevant for understanding each word.

Given that the semantics of sarcasm are highly context-dependent, it can be difficult for models to capture the meaning of an entire sentence. Traditional models such as recurrent neural networks process text unidirectionally, which limits their ability to incorporate broader context. In contrast, transformer models have demonstrated strong performance on sarcasm detection benchmarks. For example, Wen and Rezapour evaluated a transformer-based model on Twitter sarcasm datasets, Sarcasm Corpus V2 Dialogues, and SARC 2.0, achieving accuracies above 83% across all datasets and reaching up to 98% accuracy on the Twitter dataset (2025).

The specific models that we implemented for this research were all variants of the Bidirectional Encoder Representations from Transformers model (Devlin et al., 2019). **BERT**, as it is more commonly called, comes with self-supervised pretraining on an English corpus. An important part of this pretraining is that it uses masked language modeling, concealing 15% of the input words and predicting them when running the full sentence. Masked words are replaced with [MASK] 80% of the time, a random token 10% of the time, and left unchanged 10% of the time. This forces the model to learn bidirectionally rather than in reading order like recurrent neural networks. BERT was also designed with next-sentence prediction (NSP) in mind, a binary task that trains the model to detect whether two sentences follow each other, although NSP did not have any relevance to our headline dataset.

Our second model, **RoBERTa**, the "Robustly optimized BERT approach", is a fork of BERT with more extensive pretraining and tuning (Liu et al., 2019), trained on roughly ten times more data with batch sizes eight times larger, dynamic masking, and no NPS loss, making it a larger, more powerful model.

Our third model under investigation is **DistilBERT**, a lightweight adaptation of BERT (Sanh et al., 2020). The distillation process involved designing a separate model to match the output of BERT directly, not mimic its internal neural network. By doing this, they tailored a model that is less complex and less resource-intensive than the original, while still maintaining output quality. DistilBERT is

smaller in size by 40%, while being 60% faster and keeping 97% of its language understanding capabilities.

3 Dataset

We chose a dataset of online news article titles for our work. The specific dataset that we chose for our comparison was "raquiba/Sarcasm_News_Headline" from Huggingface (Sultana, 2024). The dataset is publicly available, though its licensing terms are not explicitly stated. It is made up of sarcastic news headlines from the Onion as well as non-sarcastic headlines from the Huffington Post. It is comprised of 55,328 news headlines, with 29,970 non-sarcastic news headlines and 25,358 sarcastic news headlines. Each entry of the dataset includes a binary label indicating non-sarcastic (0) or sarcastic (1), the news headline text, and a link to the original article. For our purposes, we used only the headline text and the sarcasm label. Some examples of entries in the dataset are shown in Figure 1.

is_sarcastic int64 0 1	headline string · lengths 7 926	article_link string · lengths 44 272
1	thirtysomething scientists unveil doomsday clock of hair loss	https://www.theonion.com/thirtysomething-scientists-unveil-doomsday-clock-of-hai-1819586205
0	dem rep. totally nails why congress is falling short on gender, racial equality	https://www.huffingtonpost.com/entry/donna-edwards-inequality_us_57455f7fe4b055bb1170b207
0	eat your veggies: 9 deliciously different recipes	https://www.huffingtonpost.com/entry/eat-your-veggies-9-delici_b_8899742.html
1	inclement weather prevents liar from getting to work	https://local.theonion.com/inclement-weather-prevents-liar-from-getting-to-work-1819576031
1	mother comes pretty close to using word 'streaming' correctly	https://www.theonion.com/mother-comes-pretty-close-to-using-word-streaming-cor-1819575546
0	my white inheritance	https://www.huffingtonpost.com/entry/my-white-inheritance_us_59230747e4b07617ae4cbe1a

Figure 1: raquiba/Sarcasm_News_Headline dataset

We opted to use a dataset of news headlines to reduce reliance on contextual information. Many sarcasm datasets utilize data from social media platforms such as Twitter or Reddit, but these are often drawn from comment threads or ongoing conversations, where sarcasm depends heavily on external context. In contrast, news headlines must stand alone as single phrases, needing minimal external context even when sarcasm is present.

4 Evaluation Criteria

The following outlines our experimental methodology:

1. Load Sarcasm News Headline dataset from HuggingFace and split 80/20 train-test with seed 42
2. Fine-tune pretrained BERT base uncased model on binary sarcasm classification task
3. Train for 3 epochs with batch size 16, learning rate 2e-5, and weight decay 0.01
4. Evaluate model performance using accuracy, precision, recall, F1 score, and confusion matrix
5. Apply LIME to 1,000 randomly sampled test headlines, generating explanations with 10 features and 1,000 perturbed samples each
6. Aggregate signed word weights across all LIME explanations separately for sarcastic and non-sarcastic predicted classes
7. Filter out stopwords and visualize top 20 words by weight for each class using horizontal bar charts
8. Generate LIME explanations for 9 individual headlines: 3 sarcastic news (Babylon Bee), 3 non-sarcastic news (Fox News), and 3 general sarcastic statements

For our evaluation criteria, as highlighted in our methodology above, we considered both quantitative and qualitative measures of model performance. Quantitatively, we used confusion matrices and calculated accuracy, precision, recall, and the F1 score. Accuracy measures the proportion of correct predictions and is defined as:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP represents true positives (correctly identified sarcastic headlines), TN represents true negatives (correctly identified non-sarcastic headlines), FP represents false positives, and FN represents false negatives. Precision measures the proportion of predicted sarcastic headlines that are truly sarcastic: Precision measures the proportion of predicted sarcastic headlines that are truly sarcastic:

$$Precision = \frac{TP}{TP + FP}$$

Recall measures the proportion of truly sarcastic headlines that are correctly identified:

$$Recall = \frac{TP}{TP + FN}$$

The F1 score is the harmonic mean of precision and recall:

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

Qualitatively, we analyzed LIME's word-level explanations to understand which linguistic features the model relied upon for its predictions. LIME assigns signed weight scores to individual words, indicating both the magnitude and direction (toward sarcastic or non-sarcastic) of each word's contribution to the model's classification decision. We visualized these weights as horizontal bar charts, allowing us to identify the most influential words for each class and assess whether the model captures meaningful semantic patterns associated with sarcasm.

Finally, our last qualitative evaluation criterion involved evaluating each model on nine alternative sarcasm examples: three sarcastic news headlines from The Babylon Bee, three non-sarcastic news headlines from Fox News, and three common sarcastic phrases. For each example, we used LIME to analyze word-level importance and interpret the models' predictions. The goal of this evaluation is to assess how well the fine-tuned models generalize to other sources of sarcasm.

5 Baselines

For baselines to compare all three models, it was challenging to find a study using a sarcastic headline dataset. Wen and Rezapour's custom RoBERTa model was trained on social media and dialogue sarcasm data and achieved between 83% and 98% accuracy in those domains (2025). Although it was not tested on headlines, it provides a reasonable comparison since it is also BERT-based. Additionally, we use the performance of the original BERT model on our dataset as a standard for comparison, as the other models tested are enhanced versions of this base model.

6 Results & Discussion

6.1 Confusion Matrices

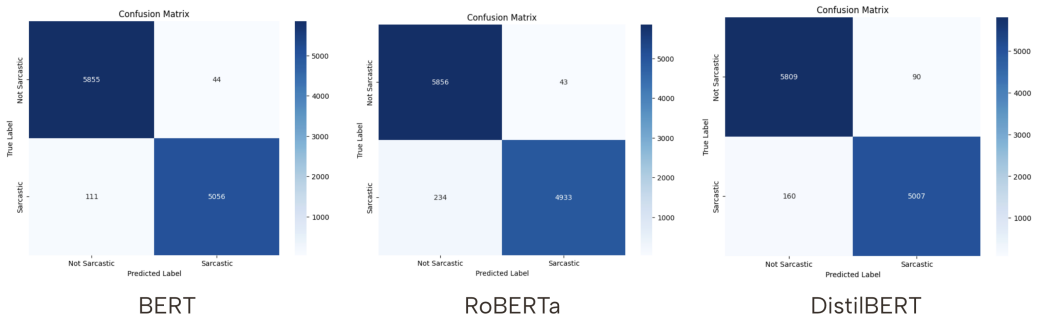

Figure 2: Confusion Matrices for each transformer model

Figure 2 shows the confusion matrices for the three transformer models, BERT, RoBERTa, and DistilBERT, visualizing their classification performance. BERT achieved 3,055 true negatives, 44 false positives, 111 false negatives, and 5,056 true positives, showing strong overall performance with relatively balanced errors.

RoBERTa had 3,856 true negatives, 43 false positives, 234 false negatives, and 4,933 true positives. This model generated less false positives than BERT, suggesting it is better at avoiding incorrectly labeling non-sarcastic headlines as sarcastic. However, it had a higher number of false negatives, meaning it missed more actual instances of sarcasm.

DistilBERT achieved the highest true negative count at 5,809, with 90 false positives, 160 false negatives, and 5,007 true positives. The model's large true negative count shows strong performance in predicting non-sarcastic headlines, though it had slightly more false positives than BERT. Its false negative count is moderate, showing it maintains reasonable sarcasm detection while prioritizing correct predictions of non-sarcastic instances.

6.2 Evaluation Metrics

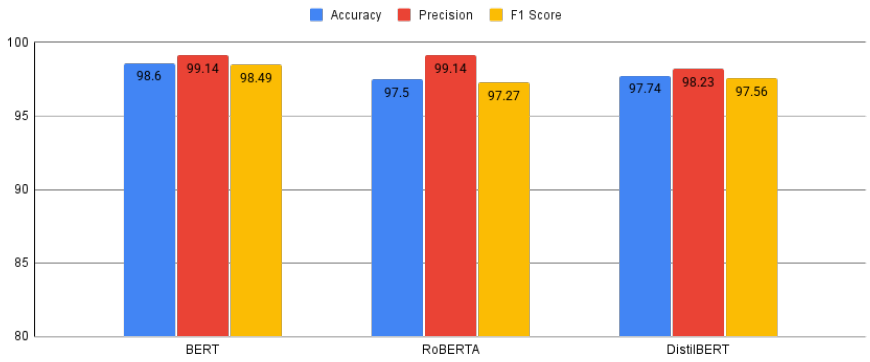

Figure 3: Bar chart comparing accuracy, precision, and F1 scores for the BERT, RoBERTa, and DistilBERT transformer models tested on a sarcasm headline dataset.

For accuracy, BERT scored highest with 98.60%, followed by DistilBERT with 97.74% and RoBERTa with 97.50%. In precision, BERT and RoBERTa both were at 99.14%, with DistilBERT at 98.23%. BERT's F1 score was the highest, at 98.49%; DistilBERT's F1 was 97.56% and RoBERTa's was 97.27%. Figure 3 visualizes these metrics as a bar chart.

Compared to the Wen and Rezapour study's custom RoBERTa, all of our fine-tuned models performed roughly as well as their model's highest reported accuracy of 98 percent, though that result was achieved on a different dataset. This shows that BERT-based models can achieve strong sarcasm detection performance even when adapted to our news headline dataset.

Compared to the base BERT model, the variant models performed at roughly the same level, with scores within about one percent of the original model's performance.

6.3 Large-Scale LIME Results

Moving on to the LIME results for our competing models, the outcomes were again very similar. Figure 4 shows the top 20 words influencing each prediction for each model. All three models shared the same top four most influential words for sarcastic predictions (man, new, report, and area), although in different orders. This similarity in sarcastic influential words may be due to all sarcastic data coming from a single source. The Onion follows a more standardized style, mimicking classic newsprint headlines. For example, an Onion headline such as "Area man could use the overtime anyway" demonstrates this standardized, generalized style, avoiding specific names while mocking common situations.

The Huffington Post, on the other hand, publishes in a much more diverse range of styles, from top 10 lists to op-eds to political commentary. This lack of a standardized style results in greater variation in the top 20 weighted words for the non-sarcastic category. However, a consistent pattern is the use of real proper nouns rather than generic ones. The word with the highest weight across all three models was "Trump," which indicates that the models are distinguishing between concrete proper nouns found in actual news and the generic placeholder subjects (e.g., "man," "woman," "area") characteristic of satirical headlines. "Trump" was so strongly associated with non-sarcastic content that even DistilBERT assigned it a negative weight in its top 20 words. Other common words across the models that pushed predictions toward the non-sarcastic class included "GOP," "Hillary," and "Ukraine."

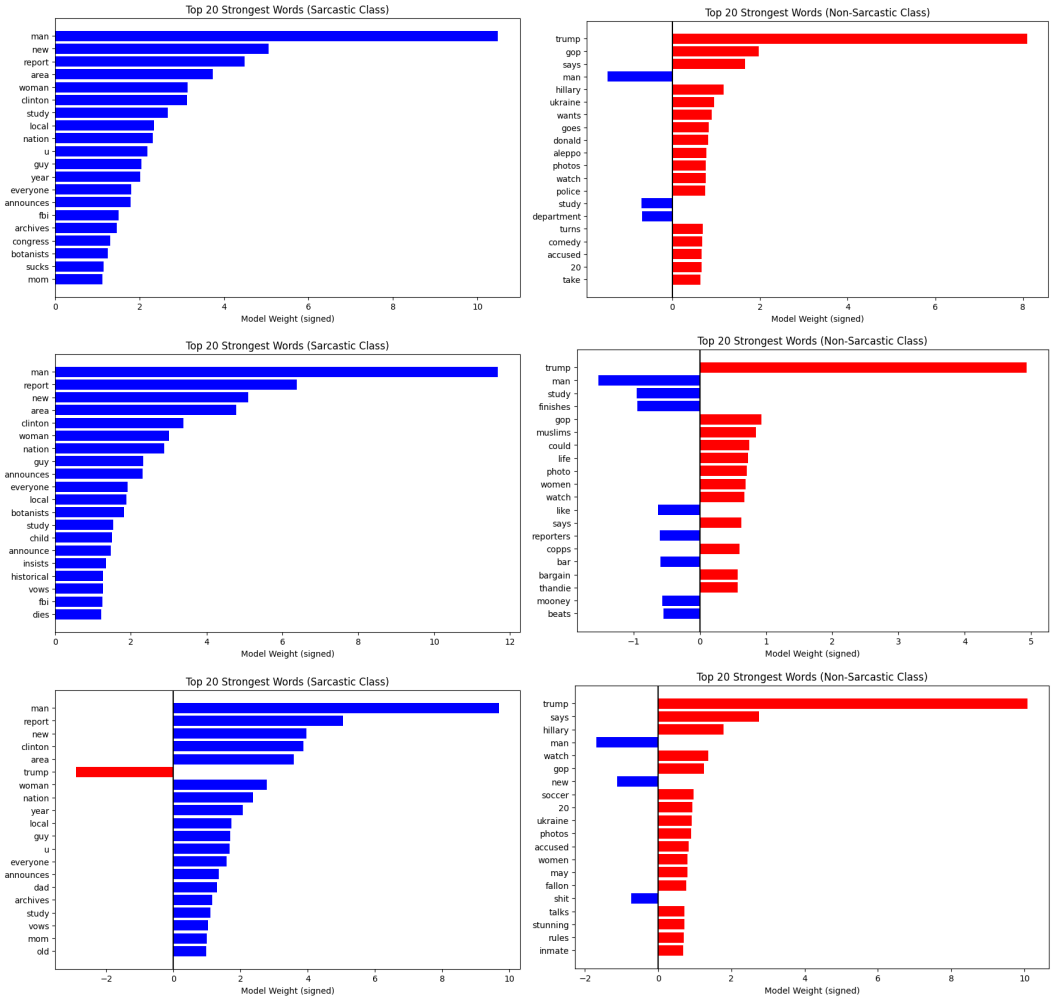

Figure 4: Bar charts showing LIME's top 20 most influential words for BERT, RoBERTa, and DistilBERT transformer models tested on a sarcasm headline dataset (from top to bottom: BERT, RoBERTa, DistilBERT).

6.4 Standalone LIME Tests

Our results should be interpreted with caution. The extremely high accuracies and overlapping top-weighted words suggest that the models may have been identifying source-specific writing styles rather than sarcasm itself, since the sarcastic and non-sarcastic headlines were drawn from only one source each. Unfortunately, we could not find any other sarcasm headline dataset without the same limitation, as most draw their sarcastic titles from The Onion. Additionally, there may be some political bias in our dataset, given that both The Onion and HuffPost are left-leaning outlets and that political names frequently appear among the influential words.

To test these potential limitations, we conducted nine standalone experiments using headlines from sources other than The Onion or HuffPost. We selected three headlines each from two conservative sources: the Babylon Bee for sarcastic titles and Fox News for non-sarcastic titles. We also tested three standalone common sarcastic phrases to evaluate whether the models could generalize beyond news headlines.

	BERT	RoBERTa	DistilBERT
"Trump Worried Reports Of Venezuelan Oil Tanker Seizure Have Ruined The Surprise Of Melania's Christmas Gift" via *the Babylon Bee*	Sarcastic Top contributing words: Worried (+0.50, sarcastic), The (−0.20, non-sarcastic), Surprise (−0.16, non-sarcastic), Trump (−0.11, non-sarcastic), Reports (+0.10, sarcastic)	Sarcastic Top contributing words: Worried (+0.57, sarcastic), Trump (+0.45, sarcastic), The (−0.23, non-sarcastic), Have (+0.14, sarcastic), Ruined (+0.10, sarcastic)	Sarcastic Top contributing words: Worried (+0.43, sarcastic), The (−0.31, non-sarcastic), Ruined (+0.16, sarcastic), Seizure (+0.15, sarcastic), Venezuelan (+0.13, sarcastic)
"California Family Still Waiting For Permit To Build Gingerbread House" via *the Babylon Bee*	Sarcastic Top contributing words: Family (+0.17, sarcastic), Still (+0.16, sarcastic), California (+0.12, sarcastic), Build (+0.08, sarcastic), To (+0.07, sarcastic)	Sarcastic Top contributing words: Family (+0.25, sarcastic), Still (+0.15, sarcastic), Gingerbread (+0.11, sarcastic), California (+0.10, sarcastic), Build (+0.09, sarcastic)	Sarcastic Top contributing words: Family (+0.15, sarcastic), Still (+0.13, sarcastic), Gingerbread (+0.08, sarcastic), House (+0.07, sarcastic), California (+0.07, sarcastic)
"Democrats Accuse Luke Skywalker Of War Crimes For Using More Than One Proton Torpedo Against Death Star" via *the Babylon Bee*	Sarcastic Top contributing words: Democrats (+0.35, sarcastic), Than (+0.26, sarcastic), Of (+0.15, sarcastic), Accuse (+0.11, sarcastic), For (+0.10, sarcastic)	Sarcastic Top contributing words: Torpedo (+0.18, sarcastic), Democrats (+0.16, sarcastic), For (+0.15, sarcastic), Using (+0.13, sarcastic), Than (+0.09, sarcastic)	Sarcastic Top contributing words: Democrats (+0.08, sarcastic), Accuse (+0.07, sarcastic), Of (+0.06, sarcastic), Using (+0.06, sarcastic), One (+0.05, sarcastic)
"Maduro trapped with few retaliation options after Trump admin seizes Venezuelan oil tanker" via *Fox News*	Sarcastic Top contributing words: few (−0.37, non-sarcastic), Trump (−0.27, non-sarcastic), seizes (−0.21, non-sarcastic), after (−0.20, non-sarcastic), Maduro (−0.15, non-sarcastic)	Not sarcastic Top contributing words: trapped (−0.35, non-sarcastic), Trump (−0.29, non-sarcastic), seizes (−0.29, non-sarcastic), Maduro (−0.23, non-sarcastic), options (−0.23, non-sarcastic)	Sarcastic Top contributing words: few (+0.49, sarcastic), Trump (+0.37, sarcastic), trapped (+0.35, sarcastic), Maduro (+0.23, sarcastic), after (+0.21, sarcastic)
"Trump greenlights Nvidia AI chip exports to China, touts 25% US share" via *Fox News*	Not sarcastic Top contributing words: Trump (−0.15, non-sarcastic), China (−0.12, non-sarcastic), to (−0.08, non-sarcastic), Nvidia (−0.08, non-sarcastic), US (−0.05, non-sarcastic)	Not sarcastic Top contributing words: China (−0.48, non-sarcastic), exports (−0.22, non-sarcastic), 25 (−0.18, non-sarcastic), Trump (−0.08, non-sarcastic), to (−0.08, non-sarcastic)	Not sarcastic Top contributing words: China (−0.52, non-sarcastic), to (−0.22, non-sarcastic), Trump (−0.17, non-sarcastic), Nvidia (−0.10, non-sarcastic), exports (−0.08, non-sarcastic)
"Kamala Harris Harris thanks supporters for 'standing up for democracy' at annual DNC meeting" via *Fox News*	Not sarcastic Top contributing words: Harris (−0.09, non-sarcastic), for (−0.08, non-sarcastic), supporters (−0.07, non-sarcastic), meeting (−0.06, non-sarcastic), annual (−0.05, non-sarcastic)	Not sarcastic Top contributing words: democracy (−0.10, non-sarcastic), thanks (−0.09, non-sarcastic), annual (−0.09, non-sarcastic), Harris (−0.08, non-sarcastic), at (−0.04, non-sarcastic)	Not sarcastic Top contributing words: for (−0.07, non-sarcastic), at (−0.07, non-sarcastic), annual (−0.07, non-sarcastic), Harris (−0.05, non-sarcastic), Kamala (−0.03, non-sarcastic)
"I love when things don't work exactly as promised." (Common sarcastic comment)	Not sarcastic Top contributing words: things (−0.12, non-sarcastic), exactly (−0.09, non-sarcastic), as (−0.08, non-sarcastic), promised (−0.07, non-sarcastic), love (−0.07, non-sarcastic)	Not sarcastic Top contributing words: love (−0.00, non-sarcastic), things (−0.00, non-sarcastic), I (−0.00, non-sarcastic), when (−0.00, non-sarcastic), don (−0.00, non-sarcastic)	Not sarcastic Top contributing words: when (−0.00, non-sarcastic), things (−0.00, non-sarcastic), don (−0.00, non-sarcastic), exactly (−0.00, non-sarcastic), as (−0.00, non-sarcastic)
"A flawless solution to a problem that didn't need to exist." (Common sarcastic comment)	Not sarcastic Top contributing words: A (−0.31, non-sarcastic), a (−0.15, non-sarcastic), to (−0.10, non-sarcastic), that (−0.09, non-sarcastic), didn't (−0.05, non-sarcastic)	Not sarcastic Top contributing words: a (−0.03, non-sarcastic), A (−0.03, non-sarcastic), need (−0.02, non-sarcastic), solution (−0.01, non-sarcastic), problem (−0.01, non-sarcastic)	Not sarcastic Top contributing words: A (−0.11, non-sarcastic), a (−0.08, non-sarcastic), solution (−0.08, non-sarcastic), flawless (−0.04, non-sarcastic), that (−0.03, non-sarcastic)
"Nothing like a productive waste of time." (Common sarcastic comment)	Not sarcastic Top contributing words: of (−0.32, non-sarcastic), productive (−0.30, non-sarcastic), Nothing (−0.26, non-sarcastic), like (−0.23, non-sarcastic), a (−0.17, non-sarcastic)	Not sarcastic Top contributing words: like (−0.09, non-sarcastic), time (−0.07, non-sarcastic), Nothing (−0.04, non-sarcastic), a (−0.03, non-sarcastic), waste (−0.02, non-sarcastic)	Not sarcastic Top contributing words: of (−0.33, non-sarcastic), productive (−0.32, non-sarcastic), Nothing (−0.32, non-sarcastic), like (−0.20, non-sarcastic), time (−0.18, non-sarcastic)

Figure 5: Table of nine standalone examples showcasing tests on conservative sarcastic sources (The Babylon Bee), non-sarcastic sources (Fox News), and common sarcastic phrases, displaying each model's prediction and the words that contributed most to the predictions according to LIME.

5 shows that all three models generalize reasonably well to new, conservative news headlines. RoBERTa correctly predicted all headline labels, while BERT and DistilBERT misclassified only one Fox News headline ("Maduro trapped with few retaliation options after Trump admin seizes Venezuelan oil tanker") as sarcastic. However, none of the models were able to correctly classify any common sarcastic phrases. This demonstrates that while the models can generalize across news headline sources, they struggle with sarcasm outside of news contexts.

7 Conclusion

The three BERT-based transformer models performed very similarly in detecting sarcasm in the raquiba/Sarcasm_News_Headline dataset, achieving over 97% across all metrics. However, when models exhibit such comparable efficacy, it is important to look beyond raw performance. We conclude that DistilBERT is the best option of the three, as it achieves nearly the same performance while requiring significantly less computational power, running faster, and being more lightweight. While all three models are relatively efficient, DistilBERT stands out due to its streamlined, distilled architecture.

For our exploration of explainability, we found considerable overlap in the influential features across all three models. The top weighted words for sarcastic predictions were nearly identical, while non-sarcastic predictions showed greater variation, especially beyond the top 5 words. This pattern suggests that the shared BERT-based architecture influences which words the models finds most important when making predictions. The similarity in sarcastic feature importance shows that all three models learned to recognize similar structural patterns in satirical news headlines, while the greater variance in non-sarcastic feature weights shows the differing writing styles present in real news sources. When tested on inputs outside our dataset, including headlines from the Babylon Bee, Fox News, and common sarcastic phrases, our fine-tuned models generalized reasonably well to new news sources but struggled significantly with sarcastic phrases that were not in a headline format.

There are potential issues with our dataset that may have skewed the results. Using a single source for each sarcasm category makes it difficult to separate publication style from actual sarcasm detection. Transformer models inherently exploit the structure of the data they are trained on and look for similar patterns during testing to achieve the highest accuracy. In our case, the transformers may have learned to distinguish between sources based on the structures of satirical sites or the presence of specific proper nouns in real news, rather than learning the deeper linguistic characteristics that define sarcasm. If we were to update and rerun this study, we would aim to create a more diverse headline dataset, drawing from multiple sources with varied writing styles and political orientations. This would allow the models to generalize better beyond the sources they were trained on, rather than exploiting the structures they learned. To generalize beyond news headlines, we could also utilize a more diverse corpus of sarcastic text. This would ensure that the models are not merely learning to classify sarcasm from news headlines, but could be applied to broader tasks.

8 Statement of Contributions

We completed roughly one half of the written portion each. For the model implementations with LIME, we originally planned to each run half of the models. Grayson was unable to get hardware acceleration working on his computer (the builds would take >15 hours each), so Chris ran all of the implementations. AI usage on this assignment was limited to minor grammar and spelling checks via Overleaf's Writefull language model tool.

9 Code Repository

https://github.com/chrisiu/sarcasm-detection

References

N. Burkart and M. F. Huber. A Survey on the Explainability of Supervised Machine Learning. *Journal of Artificial Intelligence Research*, 70:245–317, Jan. 2021. ISSN 1076-9757. doi: 10.1613/jair.1.12228. URL https://www.jair.org/index.php/jair/article/view/12228.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

D. Jurafsky and J. H. Martin. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition, with Language Models*. 3rd edition, 2025. URL https://web.stanford.edu/~jurafsky/slp3/.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach, July 2019. URL http://arxiv.org/abs/1907.11692. arXiv:1907.11692 [cs].

M. T. Ribeiro, S. Singh, and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier, Aug. 2016. URL http://arxiv.org/abs/1602.04938. arXiv:1602.04938 [cs].

V. Sanh, L. Debut, J. Chaumond, and T. Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, Mar. 2020. URL http://arxiv.org/abs/1910.01108. arXiv:1910.01108 [cs].

R. Sultana. raquiba/Sarcasm_news_headline, Aug. 2024. URL https://huggingface.co/datasets/raquiba/Sarcasm_News_Headline.

G. Visani. LIME: explain Machine Learning predictions, Dec. 2020. URL https://medium.com/data-science/lime-explain-machine-learning-predictions-af8f18189bfe.

X. Wen and R. Rezapour. A Transformer and Prototype-based Interpretable Model for Contextual Sarcasm Detection, Mar. 2025. URL http://arxiv.org/abs/2503.11838. arXiv:2503.11838 [cs].