Multimodal Sarcasm Detection

Jafar Vohra
University of New Haven, USA
jvohr1@unh.newhaven.edu

Abstract

Sarcasm is a complex and often ambiguous form of communication, which poses significant challenges to natural language processing (NLP) systems. Text-based models have succeeded in many NLP tasks, however, sarcasm detection often requires cues from other modalities such as tone, visual gestures, and context. This project focuses on Multimodal Sarcasm Detection, leveraging the MUStARD dataset to explore intermediate and early concatenation fusion strategies for combining textual, visual, and audio modalities. Two models are developed throughout the project. The first model includes a visual and textual intermediate fusion approach. The second model builds on the complexity of the previous model, utilizing an audio, visual, and textual early fusion approach. The experimental results demonstrate that while multimodal approaches improve upon random guessing, challenges such as dataset size and modality imbalances hinder optimal model performance. This report also explores alternative datasets, discusses fusion strategies, and proposes future research directions to further advance sarcasm detection.

1. Introduction

Sarcasm is a complex and often misunderstood form of communication. It poses unique challenges in natural language processing (NLP). For accurate detection, sarcasm often requires context, tone, and nonverbal cues. Traditional text-only models fall short when tackling such intricacies, leading to the utilization of multimodal sarcasm detection. Through the integration of textual, visual, and audio data, multimodal learning enhances the model's ability to capture subtle cues. This paper focuses on sarcasm detection through multimodal fusion techniques and compares their effectiveness. This report also identifies alternative datasets for multimodal sarcasm detection and offers direction for future research. This paper offers four key contributions from its work. First, it evaluates two multimodal sarcasm detection architectures using the MUStARD dataset. Next, it analyzes various multimodal fusion strategies, including concatenation and attention-based approaches. Further, it identifies the limitations within current datasets for multimodal sarcasm detection and explores alternative data sources not used in the modeling process, noting necessary areas of improvement. Finally, it discusses potential research directions in this growing field.

2. Multimodal Sarcasm Detection

A significant portion of prior research on sarcasm detection has focused on text classification, treating it as a supervised machine-learning problem. Early studies utilized traditional machine learning models, while more recent approaches have shifted to deep learning methods, such as transformers. However, sarcasm often extends beyond text. For instance, while neither the text nor a corresponding image alone may convey sarcasm, their combination

can express clear sarcastic intent. Similarly, spoken sarcasm frequently relies on facial expressions, gestures, tone, and vocal inflection, making multimodal data, such as text, audio, and video, essential for accurate sarcasm detection.

Recent efforts in sarcasm detection have focused on Multimodal Sarcasm Detection, which integrates data from diverse modalities such as text, audio, and visual cues. Leveraging different data modalities can create more robust and context-aware models to optimize model accuracy and performance. Several multimodal datasets have been curated, often sourced from social media platforms or television shows like sitcoms. However, creating large, high-quality, human-annotated datasets for MSD is particularly challenging. The subjectivity of sarcasm makes annotation complex, as the underlying incongruity may or may not be explicit. Despite these challenges, multimodal approaches represent a critical step forward, enabling systems to more accurately interpret sarcasm in real-world scenarios.

3. Visual & Textual Sarcasm Detection

Sarcasm can be expressed through the use of images accompanied by text. Sarcasm expressed through such a medium relies heavily on the incongruity between the visual and textual modality. The incongruity between the text and image data can be evaluated using an intermediate fusion strategy to establish a Text and Visual Multimodal Sarcasm Detector. The following sections describe the dataset collected and the approaches to visual-textual sarcasm detection.

3.1 Dataset

The MUStARD dataset was utilized for this study. The dataset is compiled from popular TV shows including *Friends*, *The Golden Girls*, *The Big Bang Theory*, and *Sarcasmaholics Anonymous*. The MUStARD dataset consists of audiovisual utterances annotated with sarcasm labels. Each utterance is accompanied by its context, providing additional information on the scenario where it occurs. It consists of 690 instances, equally split between sarcastic and non-sarcastic audiovisual clips. Each instance includes textual data including sarcastic utterances, surrounding context, speakers of both the sarcasm and the context of the sarcasm, and corresponding visual features.

3.2 Methodology

The first step is to import the essential libraries required for data manipulation, model training, and evaluation. Key libraries include PyTorch for model building and training, transformers for utilizing pre-trained BERT models for text embedding, and other utility libraries for handling data and visualization. To improve training speed and leverage hardware acceleration, the model is configured to use a CUDA-enabled GPU, if available. This step ensures that the model takes advantage of parallel computation for faster processing.

The textual data, stored in a JSON file, is loaded into a DataFrame. This data consists of text paired with sarcasm labels and is used as the primary input for the textual modality of the model. It is processed to create a list of texts and their corresponding labels. The visual features, stored in HDF5 files, contain embeddings or features extracted from images. These features provide the input for the visual modality of the model. They are loaded and preprocessed for normalization to ensure consistency and comparability across data points. The textual data is tokenized using the pre-trained BERT tokenizer. These tokenized texts are then passed through the BERT model to obtain embeddings.

The final dataset includes both the textual embeddings and visual features, paired with the sarcasm labels.

A feed-forward neural network (FFNN) is designed to process both textual and visual data. The model consists of separate fully connected layers for the text and visual data, followed by ReLU activation functions. These outputs are concatenated into a fusion layer, which is further processed by a classifier to predict sarcasm. The CrossEntropyLoss function is used for training, as it is suitable for both binary and multiclass classification tasks. The Adam optimizer is chosen due to its efficiency in optimizing deep neural networks. These components are crucial for minimizing the loss during the training process.

Textual data is padded to ensure uniform input length for the BERT model. The dataset is then split into training, validation, and test sets. This ensures that the model is trained on a diverse set of examples, with separate subsets for tuning and evaluation.

The model is trained over 20 epochs. During training, the model learns to map the textual and visual data to the appropriate sarcasm labels. The best-performing model, based on validation accuracy, is saved for later evaluation. After training, the best model is evaluated on the test set. Evaluation metrics such as accuracy, precision, recall, and F1 score are computed to assess the model's performance. Additionally, visualizations such as the ROC curve and confusion matrix are generated to better understand the model's performance.

3.3 Results

The model's performance was evaluated after 20 epochs. The following metrics were obtained:

- **Loss**: 8.9870
- **Training Accuracy**: 0.7578
- **Validation Accuracy**: 0.7670
- **Test Accuracy**: 0.6731

Additionally, the classification performance was assessed using:

- **Precision**: 0.7027
- **Recall**: 0.5306
- **F1 Score**: 0.6047
- **AUC-ROC**: 0.7369

The confusion matrix is as follows:

$$\begin{bmatrix} 44 & 11 \\ 23 & 26 \end{bmatrix}$$

The metrics indicate that while the model performs reasonably well in distinguishing sarcasm from non-sarcasm, there is significant potential for improvement, especially in increasing recall. Future work may involve fine-tuning the model, adjusting the class weights to address the class imbalance, or using more advanced architectures or fusion methods to better capture the subtleties of sarcasm in multimodal data. The AUC-ROC score suggests that the model is on the right track, but there is still room to enhance its performance, particularly in real-world applications where sarcasm is often expressed in more subtle or complex ways.

The model outperformed random guessing, however further optimization is required to improve its performance.

4. Audio, Visual & Textual Sarcasm Detection

In this modality, sarcasm is detected via audio and video recordings of dialogue accompanied by text captions. This type of sarcasm is very prevalent in sitcoms, TV shows, and stand-up comedy. All the datasets and methods developed to tackle this modality have focused on sitcoms and TV shows.

4.1 Dataset

The MUStARD dataset was used for this approach. For this model, in addition to textual and visual data, the audio features from the audiovisual clips were also incorporated.

4.2 Methodology

The first step involves installing and importing the required libraries and functions. If available, access to a CUDA-enabled GPU is set up to accelerate model training. This ensures that the computations are performed on the GPU for improved efficiency, particularly when dealing with large datasets and complex models.

The textual data is loaded from a JSON file, which contains the text along with labels indicating whether the text is sarcastic or not. This data serves as the input for the text modality of the model, providing essential contextual information for sarcasm detection. Visual features, stored in HDF5 files, are loaded and set up as dictionaries. These features correspond to images or visual contexts associated with each textual data point. These visual features will be used in conjunction with the textual data to help the model detect sarcasm more effectively. Pretrained BERT embeddings are loaded for the textual data. These embeddings, extracted from the last four layers of the BERT model, provide rich contextual representations of the text. BERT's ability to capture complex language patterns makes it ideal for understanding sarcasm and irony.

The features from all modalities are padded to the maximum length within each modality to ensure consistent input size. These padded features are then tensorized and flattened before being concatenated in an early fusion strategy. This creates a unified dataset with all features, along with the sarcasm labels. The dataset is split into training and testing subsets, ensuring that the model is evaluated on unseen data. For memory efficiency, the data is loaded in batches during training, which is particularly useful when dealing with large datasets.

A simple Feed Forward Neural Network (FFNN) is built, with ReLU activation functions applied between layers to introduce non-linearity. At the output layer, a SoftMax function is used to convert the raw scores into probabilities, suitable for both binary and multiclass classification. The model uses the CrossEntropyLoss as the loss function, which is appropriate for multi-class classification problems. The Adam optimizer is used to minimize the loss and update the model weights efficiently during training.

The model is trained over 5 epochs, where it learns to map the input features (text and visual) to the sarcasm labels. The model is evaluated at the end of each epoch, and the best model is retained based on performance. After training, the model is used to make predictions on the test set. These predictions are compared to the true labels to assess the model's accuracy and generalization ability. The model's performance is evaluated using standard metrics, including accuracy, precision, recall, and F1 score. In addition, visualizations such as the ROC curve and confusion matrix are generated to provide further insights into the model's classification performance.

4.3 Results

The model's performance was evaluated after 5 epochs. The following metrics were obtained:

- **Loss**: 0.7470
- **Accuracy**: 0.5290

Additionally, the classification performance was assessed using:

- **Precision**: 0.4960
- **Recall**: 0.9688
- **F1 Score**: 0.6561
- **AUC-ROC**: 0.5587
-

The confusion matrix is as follows:

$$\begin{bmatrix} 11 & 63 \\ 2 & 62 \end{bmatrix}$$

The model's performance after 5 epochs shows promise in detecting sarcasm, particularly with high recall for sarcastic instances. However, the low precision and accuracy indicate that the model is doing slightly better than simply predicting the majority class, needing further tuning, particularly to reduce false positives and improve its ability to distinguish between sarcasm and non-sarcasm. Future improvements could focus on better handling the class imbalance, refining the training procedure, or integrating more sophisticated feature engineering to address these challenges.

5. Multimodal Fusion Methods

This project delved deeply into integrating multiple modalities in sarcasm detection, to capture the intricate interactions between text, visual cues, and sometimes audio data. These modalities, when used in isolation, often fail to capture the full context and subtlety of sarcasm. Thus, integrating them through advanced fusion techniques becomes critical for achieving robust and accurate sarcasm detection models. Below, we elaborate on the key fusion strategies employed in this project and others that can be implemented to increase the performance of multimodal sarcasm detection models.

Attention mechanisms, like those derived from transformer architectures, excel at learning contextual relationships both within and across modalities. Self-attention mechanisms are used to capture dependencies within individual modalities, such as the temporal relationships in video or semantic relationships in text. By processing each modality independently, models can first learn the internal nuances before integrating cross-modal cues. Cross-attention layers can be leveraged to establish contextual dependencies between modalities. For instance, the model could identify how specific visual cues, like facial expressions, align with ironic phrases in the text, effectively fusing the two sources of information to detect sarcasm.

Concatenation is one of the simplest yet effective fusion techniques, where embeddings from different modalities are combined into a single unified representation. This approach was utilized in both early fusion and intermediate fusion strategies, allowing the model to integrate features at different stages of the pipeline. In early fusion, raw features from each modality were concatenated at the input stage, allowing the model to learn joint representations directly from the fused data. For instance, pixel-level features from images were concatenated with word embeddings from text and spectral features from audio. In intermediate fusion, embeddings from modality-specific encoders for text, and spectrogram-based encoders for audio were concatenated after being processed separately. This allows the model to learn modality-specific patterns before combining them into a shared

space for later tasks. While concatenation is computationally efficient and straightforward to implement, it does not explicitly model the interactions between modalities, requiring complementary techniques such as attention or bilinear pooling for enhanced performance.

Dot product methods were employed to capture pairwise interactions between feature vectors from different modalities. This approach, though conceptually simple, is mathematically powerful for modeling relationships between high-dimensional data. Feature Interaction Modeling takes the dot product of feature vectors from text and image modalities, allowing the model to directly quantify the alignment or similarity between their representations. A major limitation of dot product methods is the computational cost, especially when using high-dimensional data. To address this, dimensionality reduction techniques such as Principal Component Analysis (PCA) and learned bottleneck layers can be applied to reduce the size of embeddings before computing the dot product. This optimization ensures that the computational overhead remains manageable without sacrificing accuracy.

6. Alternative Datasets to Explore

Incorporating diverse datasets is critical for understanding the variability in sarcasm expression across different contexts, cultures, and modalities. By analyzing multiple datasets, researchers can gain deeper insights into how sarcasm manifests in different scenarios and improve the robustness and generalizability of sarcasm detection models. Below, we provide an in-depth overview of prominent datasets used
SarcNet is a dataset explicitly designed to support sarcasm detection across multiple languages and modalities, with a focus on image-text pairs. The dataset includes 3,335 samples in both English and Chinese, totaling over 10,000 labeled instances. This approach provides an excellent resource for studying how sarcasm varies across linguistic and cultural boundaries. Separate annotations for unimodal and multimodal sarcasm detection allow researchers to evaluate the performance of models across different input types.

The MMSD dataset is one of the most widely used resources for multimodal sarcasm detection, offering a rich mix of text and image data. With approximately 24,600 rows of data, MMSD provides a balanced mix of sarcastic and non-sarcastic instances. Contextual information is included to enhance model training, enabling researchers to consider situational cues that are often crucial for sarcasm detection. MMSD's balanced representation of different data types makes it suitable for training models that rely on multimodal inputs. The dataset has been extensively used for benchmarking various fusion methods, including attention mechanisms and hybrid approaches.

MMSD 2.0 is an enhanced version of the original MMSD dataset, addressing previous limitations and incorporating new features to better reflect real-world scenarios. Like MMSD, it contains approximately 24,600 rows of data, ensuring continuity and comparability with earlier studies. More detailed sarcasm-specific labels that account for varying degrees of sarcasm. MMSD 2.0 incorporates data collected from more recent sources, ensuring relevance and applicability in modern social media contexts. Greater effort has been made to balance the dataset across modalities, ensuring equal representation of text, images, and multimodal combinations.
While existing datasets like SarcNet, MMSD, and MMSD 2.0 provide a strong foundation, further enhancements can significantly improve their utility for model development and

evaluation. Finer-grained labels for sarcasm types and contextual cues can improve model understanding. Ensuring equal representation of text-only, image-only, and multimodal data reduces bias and improves generalizability. Regularly incorporating recent social media data, including content from emerging platforms like short-form videos, ensures dataset relevance. Expanding to diverse linguistic and cultural contexts and including annotator metadata can account for cultural nuances in sarcasm interpretation.

Exploring datasets like SarcNet, MMSD, and MMSD 2.0 provides valuable insights into the variability of sarcasm expression across modalities, languages, and cultures. However, continued efforts to enhance these datasets, through improved annotations, balanced modalities, real-time updates, and cross-cultural inclusivity, are essential for advancing the field of multimodal sarcasm detection. By addressing these challenges, researchers can develop more robust, accurate, and context-aware models capable of understanding the complex and dynamic nature of sarcasm in real-world settings.

7. Future Directions of Research

This study demonstrates the potential of multimodal learning for sarcasm detection by integrating textual, auditory, and visual cues. While the models showed promise, challenges such as limited dataset size, unbalanced modalities, and suboptimal fusion methods limited their performance. Future work should focus on exploring cultural and linguistic variations in sarcasm, developing robust datasets with real-time, cross-cultural annotations, and integrating sarcasm detection with other NLP tasks like sentiment analysis and emotion recognition.

Sarcasm varies widely across cultures and languages, influenced by differences in communication styles and cultural norms. For instance, some cultures use overt tones or wordplay to express sarcasm, while others rely on context or subtle cues. Language-specific features like idioms or cultural references also impact how sarcasm is perceived. To improve sarcasm detection models, it's essential to develop systems that adapt to these variations. Creating multilingual and cross-cultural models requires datasets that capture diverse forms of sarcasm, accounting for cultural nuances and expressions. Such models would improve accuracy and make sarcasm detection more applicable in global settings, from social media to customer service.

Implementing sarcasm detection in real-time on social media platforms and communication tools can help moderate content and improve user interaction. Sarcasm often leads to misunderstandings, and detecting it can reduce conflicts, enhance content filtering, and improve engagement. In customer service, for instance, identifying sarcastic remarks can lead to more appropriate responses, preventing possible miscommunication. Integrating sarcasm detection into platforms can also help contextualize sarcastic comments, improving content recommendations and user experience.

Sarcasm detection can enhance other NLP tasks like sentiment analysis, emotion detection, and humor recognition. Sarcasm-aware models can correct sentiment misclassifications by accurately capturing the underlying tone of sarcastic remarks. Additionally, sarcasm often involves emotions like frustration or amusement, so integrating sarcasm with emotion detection can provide a fuller understanding of the text. Combining sarcasm detection with humor recognition can also improve humor detection models. This integrated approach would result in more accurate, contextually aware systems that

better understand text, improving applications like content moderation, customer service, and social media analysis.

By addressing these areas, researchers can improve the accuracy and applicability of multimodal sarcasm detection systems.

8. Conclusion

In conclusion, Multimodal Sarcasm Detection was conducted using the MUStARD dataset and two multimodal approaches. The first Textual and Visual Sarcasm Detection Model includes a visual and textual intermediate fusion approach. This model demonstrated accuracy beyond that of random guessing, however further optimization is required to improve its performance. The second Textual, Audio, and Visual Sarcasm Detection model builds on the complexity of the previous model, utilizing an early fusion approach. The model struggled to achieve meaningful performance, likely due to the small dataset and the limitations of early fusion by concatenation. Sarcasm is a complex form of communication that poses significant challenges to natural language processing systems. For future research in sarcasm detection, robust datasets with real-time, cross-cultural annotations should be developed and multimodal fusion techniques should be enhanced.

References

Bharti, Santosh Kumar, et al. "Multimodal sarcasm detection: a deep learning approach." Wireless Communications and Mobile Computing 2022.1 (2022): 1653696.

Castro, Santiago, et al. "Towards multimodal sarcasm detection (an _obviously_ perfect paper)." arXiv preprint arXiv:1906.01815 (2019).

Farabi, Shafkat, et al. "A Survey of Multimodal Sarcasm Detection." arXiv preprint arXiv:2410.18882 (2024).

Qin, Libo, et al. "MMSD2. 0: towards a reliable multi-modal sarcasm detection system." *arXiv preprint arXiv:2307.07135* (2023).

Tang, Binghao, et al. "Leveraging Generative Large Language Models with Visual Instruction and Demonstration Retrieval for Multimodal Sarcasm Detection." *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*. 2024.

“Top 10 Multimodal Models.” *Encord*, encord.com/blog/top-multimodal-models/. Accessed 8 Dec. 2024.