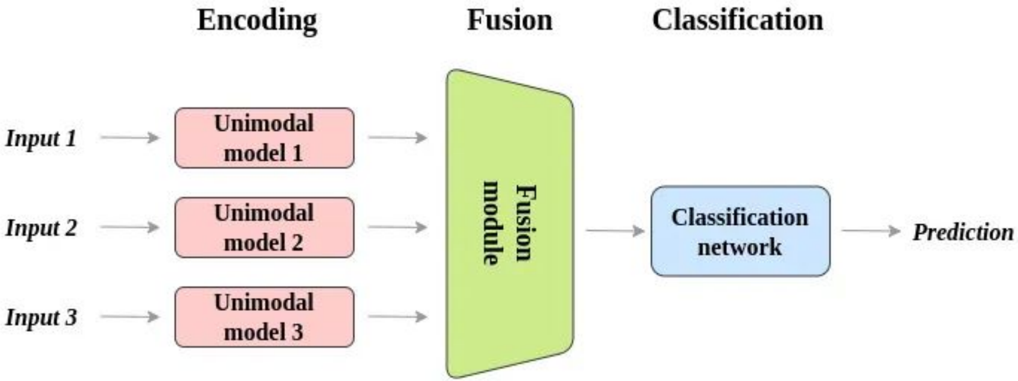
NLP Term Project

Multimodal Sarcasm Detection

Jafar Vohra

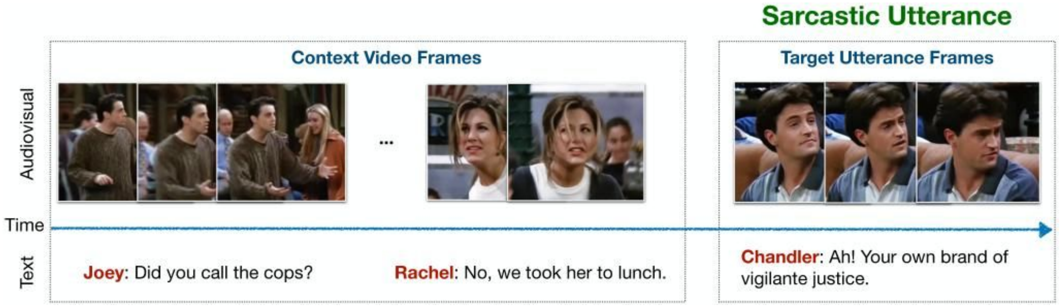
What is Multimodal Learning?

- Training models to understand and work with multiple types of data
  - Text
  - Image
  - Audio
- Different types of data correspond to various natural languages



Statement of Objectives

- Utilize audio, image, and / or text data to detect the presence of sarcasm in the MUStARD dataset using Python
- Explore approaches to fusion mechanisms in multimodal modeling
- Propose alternative datasets for Multimodal Sarcasm Detection
- Suggest future research opportunities in Multimodal Sarcasm Detection


Sarcastic Utterance
Context Video Frames
Target Utterance Frames
Audiovisual
...
Time
Text
Joey: Did you call the cops?
Rachel: No, we took her to lunch.
Chandler: Ah! Your own brand of vigilante justice.

Statement of Value

- Decoding Complex Communication
  - Sarcasm often depends on both text, audio, and visual cues, making it difficult for traditional text-only models to interpret accurately
- Real-World Applications
  - Enhances sentiment analysis and customer feedback interpretation, preventing misinterpretation of sarcastic tones in social media, reviews, and customer support
  - Supports content moderation by accurately identifying sarcasm to avoid unnecessary censorship on platforms
- Establishing Benchmarks
  - Addresses the lack of robust datasets, providing standardized benchmarks for future research on multimodal data and context-dependent sarcasm
- Research Gaps
  - Investigates challenges like unbalanced modalities (text/image imbalance) and cross-modal attention techniques to improve multimodal models

Relevant Work Review (Citations)

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

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

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

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.

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

Intended Approach

Algorithms / Models

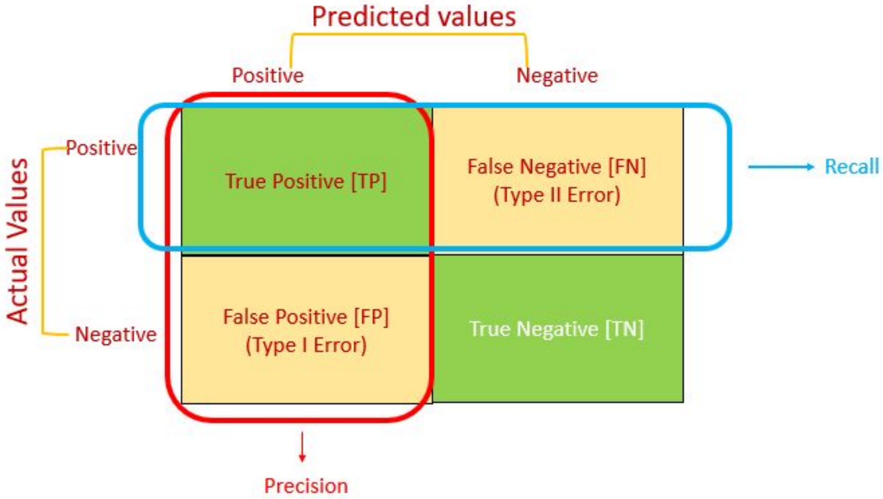
- Text and Visual MultiModal Sarcasm Detector with Strategic Intermediate Fusion
- Text, Audio, and Visual Sarcasm Detection Model with Early Concatenation

Tools

- PyTorch
- Pandas
- NumPy
- JSON
- HP5Y
- Scikit-learn
- Matplotlib
- Seaborn
- Cuda GPU

Evaluation Methodology

- Relevant Metrics
  - Accuracy
  - Precision (Type I Error)
  - Recall (Type II Error)
  - F1 Score
  - Classification Report
  - AUC-ROC Curve
  - Confusion Matrix
- Considerations
  - Model Complexity
  - Class Balance
  - Overfitting



	Positive	Negative
Positive	True Positive [TP]	False Negative [FN] (Type II Error)
Negative	False Positive [FP] (Type I Error)	True Negative [TN]

Code Demo

Multimodal Fusion Methods

1. Attention-based Methods
    a. Uses transformer architecture to convert embeddings into a query-key-value structure.
    b. Initially improved language models; now used in computer vision and generative AI.
    c. Enables context-aware processing by understanding relationships between embeddings.
2. Concatenation
    a. Merges multiple embeddings into a single feature representation.
    b. Combines textual and visual embeddings for a consolidated multimodal feature.
    c. Useful in intermediate fusion strategies.
3. Dot Product
    a. Element-wise multiplication of feature vectors from different modalities.
    b. Captures interactions and correlations between modalities.
    c. Effective for low-dimensional vectors; high-dimensional vectors may require extensive computational power and miss critical nuances.

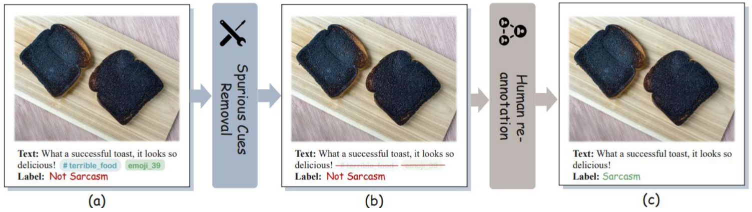
Alternative Data for Multimodal Sarcasm Detection

Datasets

- SarcNet
  - English and Chinese image-text pairs (3,335)
  - Annotations for unimodal and multimodal data (10,000)
- MMSD
  - English Tweet Text and Image (24,600)
- MMSD2.0
  - Spurious cues removed
  - Manually corrected annotations
  - Around 24,600 records

Future Improvements

- Capture detailed aspects of sarcasm
- Ensure even representation of text, image, and audio data
- Continuously update with real-time data from various sources.
- Expand to include more diverse cultural and linguistic data.


Text: What a successful toast, it looks so delicious! # terrible_food emoji_39
Label: Not Sarcasm
(a)
Spurious Cues Removal
Text: What a successful toast, it looks so delicious! ~~# terrible_food~~ ~~emoji_39~~
Label: Not Sarcasm
(b)
Human re-annotation
Text: What a successful toast, it looks so delicious!
Label: Sarcasm
(c)

Future Work

- Cross-Cultural Sarcasm Detection
  - Investigate how sarcasm is expressed differently across cultures and languages
  - Models that can adapt to these variations can enhance accuracy
- Real-Time Sarcasm Detection
  - Implementation in social media and communication platforms
  - Particularly useful for moderating content and improving user interactions
- Integration with Other NLP Tasks
  - Study the integration of sarcasm detection with other natural language processing tasks
    - Sentiment analysis
    - Emotion detection
    - Humor recognition
  - Provide a more comprehensive understanding of the text

Thank you!