Multimodal Sentiment Analysis for Memes: BERT/CLIP & MMBT

Kritin Nandish*
Thrishna Bhandari†

University of California, Berkeley

Online communication increasingly relies on multimodal content—short texts paired with images, such as memes— to convey sentiment and opinion. Traditional sentiment analysis models, which tend to focus solely on text, often fail to capture the full spectrum of meaning embedded in these cues. In this paper, we explore two approaches to multimodal sentiment classification: (1) combining BERT for textual encoding with CLIP for visual features, and (2) using a Multimodal BiTransformer (MMBT) that jointly models text and image inputs with a transformer layer. We evaluate both models on a labeled meme dataset; our goal is to identify not only overt negativity, but also subtler forms of harmful sentiment—such as sarcasm or visual irony— that are often missed by text-only models. Although we implemented multiple architectures and tuning strategies, neither model significantly outperformed a simple baseline. Manual review of dataset samples revealed a high rate of labeling inconsistencies, suggesting that data quality likely constrained model performance. These findings underscore both the potential and limitations of multimodal sentiment analysis, particularly in noisy, crowdsourced and loosely moderated meme data in social media.

0 INTRODUCTION

As digital communication becomes more visual, language has evolved into a multimodal form in which text and images jointly construct meaning. This evolution is especially visible in the rise of memes: compact, image-text combinations that dominate platforms like Instagram, Reddit, and X (formerly Twitter). Memes often deliver layered or ambiguous messages where emotional tone is shaped not just by words, but by the accompanying visuals. These complexities pose a challenge for traditional sentiment analysis systems that operate only on textual input, potentially leading to misclassification or missed signals.

Beyond entertainment, memes can subtly convey harmful sentiment— negativity, hate speech, or misinformation— often masked through sarcasm, irony, or symbolic imagery. In such cases, understanding sentiment requires models that interpret both the language and the visuals in context.

In this study, we address this challenge by evaluating two approaches to multimodal sentiment classification. The first combines BERT for textual feature extraction with CLIP for visual encoding, feeding both into a classification layer. The second employs a Multimodal Bitransformer (MMBT), which jointly processes text and image embeddings within a unified architecture with transformers. Using a labeled meme dataset, we compare the ability of the two methods to identify both overt and subtle negative sentiment. This work contributes to building more context-aware and reliable sentiment analysis tools for use in social media environments, where multimodal content is the norm.

1 BACKGROUND

Multimodal sentiment analysis has gained increasing attention, with several recent studies emphasizing the importance of effectively fusing multiple modalities, such as text, images, and emoticons, to capture nuanced sentiment expressions. For example, Kiela et al. (2019)[1] introduced the Multimodal Bitransformer (MMBT), a supervised model designed to jointly process and classify text and image inputs. Unlike previous approaches that relied on late fusion techniques, MMBT employs early fusion by concatenating text and image embeddings and passing them through a shared transformer architecture. The authors demonstrated that MMBT outperforms existing multimodal models on several benchmark tasks, including the Hateful Memes dataset, highlighting its effectiveness in handling multimodal classification challenges. Inspired by

*kritin_nandish@berkeley.edu
†thrishnabhandari@berkeley.edu

these results, we adopted MMBT as one of our primary modeling approaches.

In a related domain, Javaid et al. (2023) proposed an image-text multimodal emotion analysis model that incorporates emoji features alongside text and image data, using multi-head attention to mitigate semantic mismatches. Their approach achieves improved sentiment classification accuracy [2], highlighting the value of incorporating cues like emojis, which are prevalent in online communication but often overlooked in traditional models.

Beyond simple fusion, advanced modeling techniques have been deployed to tackle challenges related to redundant or irrelevant information across modalities. Chen et al. (2023) introduced a multimodal sentiment analysis framework based on supervised contrastive learning combined with CNNs and Transformers, which enhances feature representation and outperforms previous state-of-the-art models on multiple datasets [3]. Such approaches show the critical role of feature engineering and fusion strategies in advancing multimodal sentiment analysis.

2 METHODS

2.1 Approach Overview

To address the task of tone detection and sentiment analysis, we adopted a structured three-step approach. Initially, our objective was to develop an effective binary classifier. Upon achieving satisfactory performance, we aimed to extend the model to handle multiclass classification, thereby enabling it to capture more nuanced tonal variations in meme content. After training and evaluating two distinct models for binary classification, we selected the one demonstrating superior performance—based on our predefined evaluation metrics—to serve as the foundation for the multiclass extension.

Following an initial review of relevant methodologies, we selected two model architectures for experimentation:

1) BERT + CLIP (Binary and Multiclass): This multimodal classification approach integrates both visual and textual features using pre-trained encoders. For the textual component, a BERT model encodes the meme captions and the meme image is processed using a frozen CLIP vision encoder (CLIPModel with ViT backbone). Both embeddings are concatenated and fed into a feed-forward classification head.
2) Multimodal Bitransformer (MMBT)[1]: This approach takes paired text and image inputs from memes. Text embeddings are generated using a BERT encoder, while image representations are extracted via a ResNet model. After respective feature extraction, the embeddings are concatenated and projected into a shared embedding space. This combined representation is passed through a transformer layer, facilitating cross-modal interactions via attention mechanisms that enable the model to capture complex interactions between modalities. The output is then fed into a classification head to produce the final logits.

2.2 Data Processing

We utilized the Memotion Dataset 7k from Kaggle [4], which comprises 6,992 annotated memes designed for sentiment analysis. Given our objective of building a model applicable to content moderation—particularly for identifying underlying tones in social media content—we focused on the offensive feature of the dataset. This feature is categorized into four classes: very offensive, slight, hateful offensive, and not offensive. To establish a baseline for our classification task, we initially reframed the problem as a binary classification. Specifically, we consolidated the very offensive, slight, and hateful offensive categories into a single offensive class (label 1), while retaining the not offensive category as a separate class (label 0)

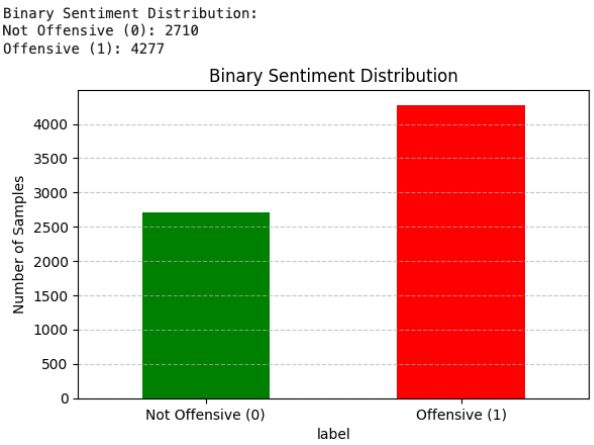

Fig. 1: Class distribution for our data post processing

2.3 BINARY CLASSIFIERS

Before diving into training our models, we wanted to establish a baseline predictor to moderate performance, that consistently predicted our majority class: offensive (label 1). Based on the 60-to-40% split of our training data, we noted the imbalance in class distribution, and included F1-score in our evaluation metrics to get a more cohesive picture. The baseline model results were:

Table 1: Baseline Model Performance & Metrics

Stage	Accuracy	F1-Score
Validation	0.6120	0.7593
Testing	0.6123	0.7595

Note. This table shows baseline model performance for predicting majority class 'Offensive' (1).

2.3.1 Model 1: BERT + CLIP

Our first model was a multimodal binary classification architecture that leveraged complementary visual and

textual features. The model integrated two pre-trained encoders-BERT and CLIP-to extract contextual embeddings from paired image and text inputs. We chose the CLIP vision encoder for its stronger alignment between the image and text modalities.

For the text input, we used Bert-based-uncased, which encodes tokenized meme text into a 768-dimensional contextual embedding vector. In parallel, image inputs were processed using the clip-vit-base-patch16 vision transformer encoder from the CLIP model. The resulting 512-dimensional image embedding corresponds to the pooled output of the CLIP visual transformer.

Both the BERT and CLIP encoders operate independently on their respective modalities. The outputs from the BERT encoder (768-dimensional) and the CLIP image encoder (512-dimensional) are then concatenated to form a single feature vector of 1280 dimensions.

This combined feature vector is then passed to a MultimodalClassifier, which is a simple feedforward neural network, which consists of two linear layers with ReLU activation and a dropout layer with a dropout rate of 0.3 between them. The first layer projected the 1280-dimensional input to a 256-dimensional space, and the second linear layer outputs 2 logits for the binary classification task.

The model was trained using the Cross-Entropy loss function and the Adam optimizer with a learning rate of 2e-5. Training was conducted for 5 epochs with a batch size of 32. Performance was monitored on a validation set, and the final evaluation was performed on a separate unseen test set. During training, the model exhibited consistent performance metrics across the epochs, suggesting the model quickly reached a performance plateau. The results observed indicate that the multimodal binary classification model did not significantly outperform the majority-class baseline, suggesting minimal learning for this specific classification task.

Table 2: BERT + CLIP Classifier Model Performance

Epoch	Training Accuracy	Validation Accuracy	Validation F1-Score
1	0.6119	0.6120	0.7593
2	0.6124	0.6120	0.7593
3	0.6112	0.6120	0.7593
4	0.6119	0.6120	0.7593
5	0.6116	0.6120	0.7593
	Test Accuracy: 0.6137	**Test F1-Score: 0.7602**	

Note. Training and validation metrics for the BERT + CLIP classifier. Test results align closely with a majority-class baseline, reflecting the challenge of surpassing simple heuristics on noisy data.

2.3.2 Model 2: MMBT

We implemented a Multimodal BiTransformer (MMBT) architecture to jointly process textual and visual information for binary sentiment classification. The model integrates pre-trained encoders for each modality, whose outputs are fused via a transformer encoder and passed to a classification head.

For textual input, we used the BERT-base-uncased model to produce contextualized token embeddings. In parallel, image inputs were processed using a pre-trained ResNet-50 model, truncated before its final classification layer, to yield 2048-dimensional feature vectors. Both text and image outputs were projected into a shared 256-dimensional embedding space via linear transformations. These embeddings were then concatenated, treating the image vector as an additional token appended to the text sequence.

The combined sequence was passed through a lightweight Transformer encoder composed of two layers with four attention heads per layer. These hyperparameters were selected to reduce model capacity and mitigate overfitting on our training data. The output corresponding to the first token position (analogous to the [CLS] token in BERT) was pooled and fed into a feedforward classification head consisting of two linear layers with ReLU activation and dropout regularization. The final output was a single logit used for binary classification.

Images were preprocessed using standard augmentation techniques, including resizing to 224×224 pixels, random horizontal flipping, +/- 10 degree rotation, brightness/contrast jittering, and normalization using ImageNet statistics. Text was lowercased and tokenized using the BERT tokenizer with a maximum sequence length of 128 tokens, applying padding and truncation as needed. Initial training was conducted using the Cross-Entropy loss function. To address class imbalance, we later adopted the BCEWithLogitsLoss function with class weights (detailed in the Challenges section). Experiments were conducted with batch sizes of 16 and 32 to evaluate the trade-offs between convergence stability and generalization. Models trained with batch size 32 exhibited more consistent and stable performance.

Training was performed using the Adam optimizer for up to 10 epochs, with early stopping based on validation F1 score and a patience of three epochs. The model checkpoint with the highest validation F1 score was saved and later used for final evaluation with the test set.

Despite these efforts, the MMBT model did not consistently outperform our majority-class baseline. Throughout optimization and fine-tuning, validation performance remained unstable and exhibited fluctuations. The limitations of model performance and their connection to data quality are discussed further in the following sections.

Table 3: MMBT Classifier Model Performance

Epoch	Training Accuracy	Validation Accuracy	Validation F1-Score
1	0.4933	0.6120	0.7593
2	0.5088	0.5088	0.4717
3	0.5491	0.5197	0.5729
4	0.5720	0.5297	0.5962
	Test Accuracy: 0.6123	**Test F1-Score: 0.7595**	

Note: This model run exactly matched the test performance of the baseline. **Hyperparameters**: Batch size: 32, Patience: 3 *Early stopping triggered*

Table 4: MMBT Classifier Model Performance

Epoch	Training Accuracy	Validation Accuracy	Validation F1-Score
1	0.5976	0.5354	0.6035
2	0.6474	0.4681	0.4620
3	0.7541	0.5719	0.7045
4	0.8426	0.5247	0.6188
5	0.9249	0.5412	0.6698
6	0.9404	0.5154	0.6134
	Test Accuracy: 0.5658	**Test F1-Score: 0.6973**	

Hyperparameters: Batch size: 16, Patience: 3 *Early stopping triggered*

Table 5: MMBT Classifier Model Performance

Epoch	Training Accuracy	Validation Accuracy	Validation F1-Score
1	0.4969	0.6034	0.7452
2	0.5391	0.5812	0.7014
3	0.5880	0.5297	0.5907
4	0.6815	0.4710	0.4538
5	0.8001	0.5089	0.5457
	Test Accuracy: 0.6023	**Test F1-Score: 0.7447**	

Hyperparameters: Batch size: 32, Patience: 4 *Early stopping triggered*

2.4 MULTICLASS CLASSIFIER

A multimodal classification model was developed to categorize data into three distinct classes: not-offensive, very-offensive, and slight. This model integrates pre-trained encoders for text and images, concatenates their outputs, and processes them through a feedforward neural network for multi-class prediction. Similar to the binary classifier, the BERT-base-uncased model was used to generate 768-dimensional token embeddings. Concurrently, image inputs were processed using the clip-vit-base-patch32 CLIP model, which yielded 512-dimensional image features. The embeddings and CLIP image features formed a single 1280-dimensional feature vector which was fed into a multi-modal classifier. The classifier is a feedforward neural network comprising three linear layers with ReLU activation functions and dropout layers (dropout rate of 0.3) between the first two linear layers. The first linear layer transforms the 1280-dimensional input to 512 dimensions, the second maps it to 128 dimensions, and the final layer outputs 3 logits, corresponding to the three target classes.The data was then split into training, validation, and test sets (60,20,20) respectively, with stratification to preserve the original class distribution. Text entries with missing values were removed.

Training was performed using the Cross-Entropy loss function, with class weights applied to address potential class imbalance, computed using a balanced weight class. The Adam optimizer with a learning rate of 2×10^{-5} was used. The model was trained for 5 epochs with a batch size of 32. Performance was monitored using accuracy and weighted F1-score on the validation set, and a final evaluation was conducted on the test set, providing overall accuracy, weighted F1-score, and a confusion matrix to visualize per-class performance.

The results, as recorded in the table below, suggested multi-class model struggled to achieve high accuracy, particularly given the challenges of multi-class classification and potential data complexities, further confirming observations from the binary classification task regarding the model's inability to learn due to inconsistent labeling.

Table 6: Multiclass Classifier Performance

Epoch	Training Accuracy	Validation Accuracy	Validation F1-Score
1	0.3695	0.3319	0.2907
2	0.3649	0.3326	0.3369
3	0.3666	0.3533	0.3122
4	0.3954	0.3415	0.3409
5	0.4095	0.3474	0.3298
	Test Accuracy: 0.3737	**Test F1-Score: 0.3544**	

Hyperparameters: Batch size: 32, Optimizer: optim.Adam, Loss Function: nn.CrossEntropyLoss

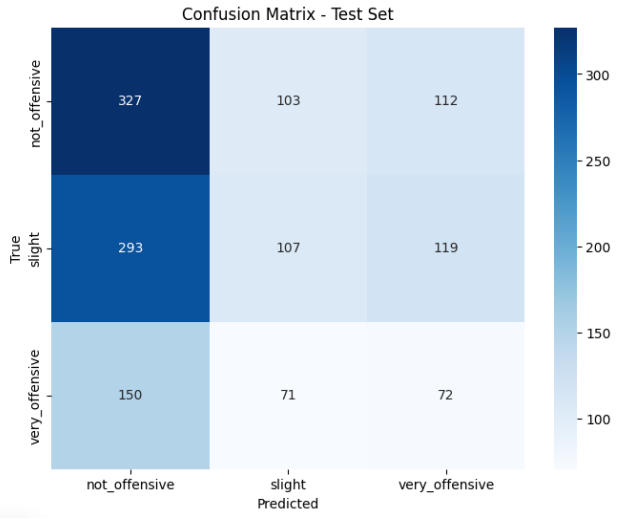

Fig. 2: Confusion matrix generated with BERT+CLIP Multiclass classifier results: visual summary of the model performance by showing the counts of actual versus predicted classes, highlighting where the model correctly predicted (top-left to bottom-right diagonal) and where it misclassified instances into other categories.

3 RESULTS & DISCUSSION

Over the course of our work, we started to notice our model's were rarely performing better than the baseline benchmark. The results were unstable, ranging on average 1-5 percent higher or lower than 0.61 validation/testing accuracy and 0.76 F1-scores. We tried a combination of hyperparameter tuning for both model architectures in attempts to increase performance while also taking a closer look at our core dataset, to try to isolate our root issue.

3.1 CHALLENGES

3.1.1 Training & Evaluation

During the training and evaluation phases of our classification tasks, we encountered several challenges that affected model performance.

The Multimodal Bitransformer (MMBT) model and the multi-class BERT/CLIP model both exhibited early signs of overfitting. While training accuracy increased steadily across epochs, validation accuracy remained low, suggesting that the model was memorizing the training data rather than learning patterns that generalize to unseen data. To address this, we experimented with various architectural and training adjustments discussed below.

Given the imbalanced nature of our dataset, we determined that accuracy alone was an insufficient evaluation metric. Although we continued to monitor accuracy, we shifted our primary focus to the F1-score, which better captured model performance under class imbalance. When training was guided by validation accuracy alone, both models underperformed relative to our binary classification baseline, with test accuracies averaging around 55% for MMBT and a lowly 34% for our multi-class.

To mitigate overfitting for the MMBT model, we implemented early stopping with a patience parameter set to 3 (with additional experiments at 2 and 4). This strategy halted training when the validation F1-score failed to improve for a specified number of epochs, thereby preventing continued memorization of the training set. While early stopping on its own did not yield substantial performance gains, shifting our optimization and evaluation focus to the F1-score led to improvements, with the MMBT model reaching closer parity with our baseline classifier. We also introduced a dropout layer with a rate of 0.4 in the classifier, aiming to promote generalization by randomly deactivating portions of the neural network during training. However, the results were inconsistent and did not lead to a clear performance improvement. Meanwhile, for our multi-class and binary class CLIP/BERT models we chose a different approach. We experimented with the actual classifications. For the initial binary CLIP/BERT model we chose to have the not-offensive label by itself and to group the rest. This was primarily due to the imbalance mentioned earlier. In the case of the multi-class, we found that the hateful-offensive class was being misclassified the most, and our model showed significant progress, increasing both in the F-1 metric and overall accuracy, when we chose to exlude this classification.

An adjustment that yielded more tangible results was tuning the batch size for both MMBT and multiclass. We experimented with batch sizes of 16, 24, and 32, and observed that a batch size of 32 consistently led to more predictable training behavior, with modest but consistent improvements in both accuracy and validation F1-scores. Specifically, F1-scores were, on average, approximately 0.025 higher using this configuration.

Many of our architectural and training adjustments were targeted at addressing class imbalance in the dataset. To better reflect the skewed distribution of offensive versus non-offensive labels, we replaced the standard cross-entropy loss with a weighted binary cross-entropy (BCE) loss function. The class weights were computed from the label distribution, tuning the model to penalize misclassifications of minority-class samples more heavily, encouraging more balanced learning during training. This change, especially when combined with F1-based evaluation, resulted in one of the most significant performance improvements. By this stage, overfitting in both the MMBT and multi-class models were substantially reduced. However, while validation and test performance had stabilized, the scores remained comparable to those of the baseline model, without showing significant improvement beyond it.

3.1.2 Dataset Quality

After extensive hyperparameter tuning across multiple model configurations, we observed that performance had plateaued. This prompted us to re-examine the dataset itself as a potential source of limitation. Given that sentiment analysis inherently involves a degree of subjectivity, we initially conducted a brief review of the dataset and found the labels to be generally acceptable.

However, considering the various modifications we had already made to address class imbalance during training, we hypothesized that the label distribution might still be affecting performance. To test this, we adjusted the training data to create an approximately even split between offensive and non-offensive labels. Despite this rebalancing, model performance deteriorated for the MMBT and the binary BERT/CLIP model. Our accuracy plummeted from 60 percent to the mid 40 percent on average. However, we noticed a slight uptick in our multi-class as it increased from 30 percent accuracy and a F-1 score of .288 to 37 percent accuracy and an F-1 score of 0.35. Despite this modest improvement, it was clear to us that no adjustments or rebalancing would provide substantial improvements.

This result led us to suspect that label quality—rather than label distribution—might be the primary issue. We conducted a manual review of over 200 meme samples and found that while the 'hateful offensive' category was labeled accurately in nearly all cases, approximately 35% of the subset appeared to be misclassified. These findings suggest that label noise, likely due to the inherent subjectivity of sentiment labeling, may be a key factor limiting model performance.

4 NEXT STEPS

Potential next steps we are interested in exploring involve applying our current model architectures to a more robust and reliably annotated dataset. Access to high-quality ground truth labels would allow for a more accurate assessment of model performance and provide clearer signals during hyperparameter tuning and evaluation. This would not only offer improved feedback loops for model optimization, but would also help isolate model limitations from those stemming from data quality.

With our current dataset, we aim to explore the performance of more powerful multimodal models, such as Gemma 3, or other recent transformer-based architectures with significantly larger capacities. This direction will require access to more advanced hardware with higher GPU resources. Evaluating a model of this scale on our existing data could serve a dual purpose: first, to test whether more sophisticated architectures can overcome the noise and subjectivity inherent in the dataset; and second, to further validate our hypothesis regarding label inconsistencies if even these models fail to perform satisfactorily under the same conditions as ours.

Future experimentation may explore human-guided annotation techniques, including interactive labeling workflows and active learning, to resolve ambiguous predictions during training and incrementally enhance the quality of the dataset. This would be aided by an overarching standardization of sentiment to increase the consistency in our classification labels.

5 CONCLUSION

While our project did not fully achieve the level of performance initially proposed, the process yielded valuable insights into the challenges of multimodal sentiment analysis, particularly in the context of social media content and memes. Despite implementing a range of model architectures and optimization techniques—including multimodal fusion via MMBT, BERT/CLIP layers with classifier and extensive hyperparameter tuning—our models were ultimately unable to significantly outperform a baseline classifier.

Upon closer investigation, we identified that a key limiting factor was the quality of the dataset itself. A substantial portion of the meme annotations were found to be inconsistently labeled, undermining the model's ability to learn reliable patterns, especially in a task as inherently subjective as tone and offensiveness detection.

Nevertheless, the project provided a rich hands-on experience in training and evaluating multimodal models under real-world constraints. We gained practical expertise in model debugging, working with imbalanced datasets, designing evaluation strategies (such as F1-based early stopping), and analyzing data quality issues. These experiences deepened our understanding of the complexities involved in deploying machine learning systems for nuanced language tasks and highlighted the critical role that high-quality labeled data plays in achieving robust performance.

6 AUTHORS' CONTRIBUTIONS

- Project Proposal + Research: Thrishna Bhandari & Kritin Nandish
- Data Preprocessing: Kritin Nandish
- Baseline Model: Thrishna Bhandari
- BERT + CLIP Binary Classifier: Kritin Nandish
- MMBT Binary Classifier: Thrishna Bhandari
- BERT + CLIP Multiclass Classifier: Kritin Nandish
- Report: Thrishna Bhandari & Kritin Nandish

7 REFERENCES

[1] D. Kiela, S. Bhooshan, H. Firooz, E. Perez, D. Testuggine, "Supervised Multimodal Bitransformers for Classifying Images and Text," (2020), URL https://arxiv.org/abs/1909.02950.

[2] G. Bao, L. Sun, R. Zhang, B. Zhang, Z. Shen, S. Chen, "Research on Image-text Multimodal Emotions Analysis with Fused Emoji," presented at the *2024 27th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, pp. 18–23 (2024), [Online]. Available: 10.1109/CSCWD61410.2024.10580738, URL https://ieeexplore.ieee.org/document/10580738.

[3] L. S. Dean, G. Devendra, J. Boonyanudh, N. Subia, M. D. Tallquist, V. R. Nerurkar, S. P. Chang, D. C. Chow, C. M. Shikuma, J. Park, "Phenotypic alteration of low-density granulocytes in people with pulmonary post-acute sequelae of SARS-CoV-2 infection," *Frontiers in Immunology*, vol. 13, p. 1076724 (2022), [Online]. Available: 10.3389/fimmu.2022.1076724, URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10007095/.

[4] W. Scott, "Memotion Dataset 7k," Kaggle dataset (2020), URL https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k/data.

[5] "Discussion on Memotion Dataset 7K: Data Quality and Labeling Issues," https://www.kaggle.com/datasets/williamscott701/memotion-dataset-7k/discussion/254590.

APPENDIX

We are dedicating the Appendix to demonstrate examples of mislabeled data in the 'Memotion Dataset 7k'. We captured evidence of ground truth samples whose overall messaging does not align with the expectation of their respective classification.

- **Category**: 'Not Offensive' **Expectation**: Acceptable. Does not contain disrespectful content, harmful messaging or including bigotry, racism, hate, etc. of any kind.

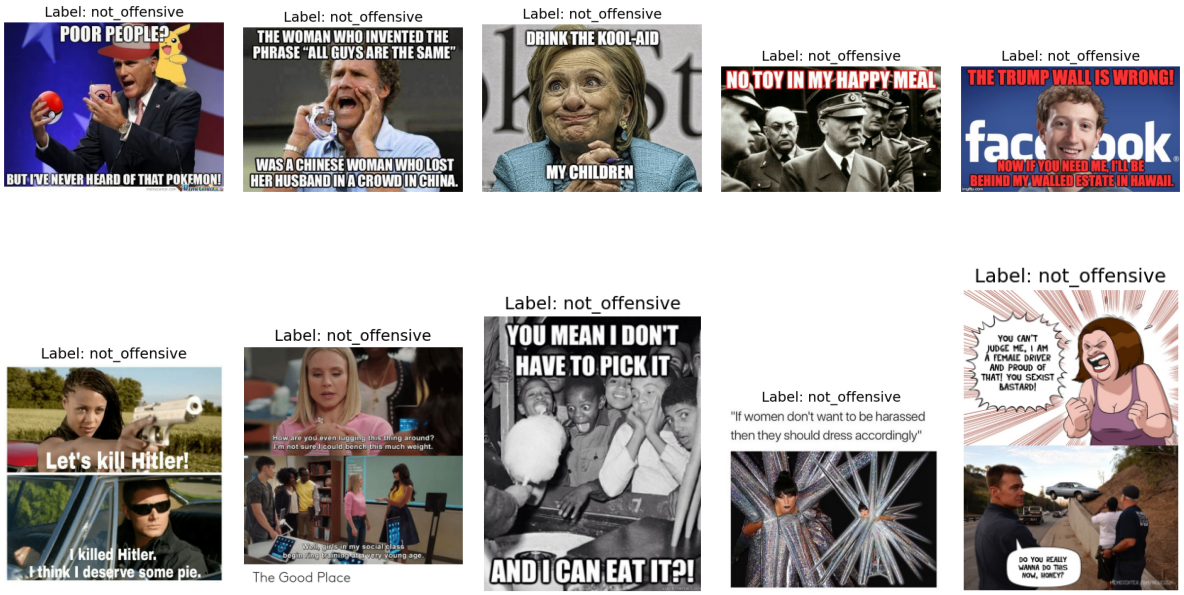

Fig. 4: Examples of inaccurate 'not offensive' ground truth samples from dataset.

- **Category**: 'Very Offensive' **Expectation**: Contains distinctive negativity, prominent and obvious disrespectful, extremely upsetting, or insulting content, harmful messaging.

Fig. 5: Examples of inaccurate 'very offensive' ground truth samples from dataset.

- **Category**: 'Hateful Offensive' **Expectation**: Contains severely negative and disrespectful content, targeted hostility, calls for violence, or including prejudice, bigotry, racism, hate, etc.

Fig. 6: Examples of inaccurate 'hateful offensive' ground truth samples from dataset.

- **Category**: 'Slight Offensive' **Expectation**: Contains content that invokes discomfort, mild annoyance but not specific targeting of an individual or group, general negative sentiment and absence of obvious disrespectful, extremely upsetting, or insulting content & harmful messaging.

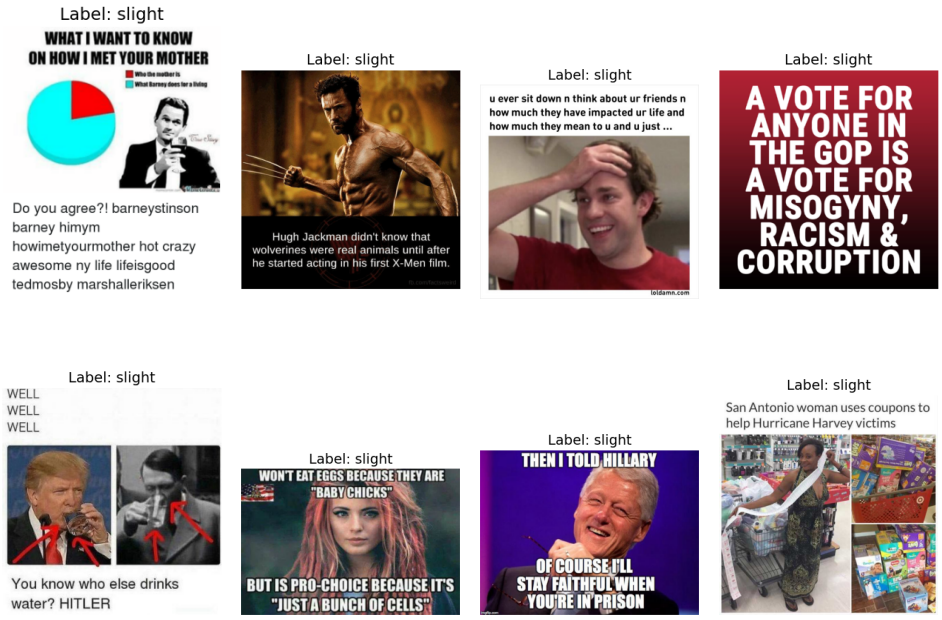

Fig. 7: Examples of inaccurate 'slight offensive' ground truth samples from dataset.

Additional Reported Concerns About Dataset Labelling: Discussion post on Kaggle[5]: