

Multimodal Sarcasm Detection

Aneesha Sampath^{*1} Neha Nishikant^{*1} Arthi Nagarajan^{*1}

Abstract

Sarcasm is a complex and subjective phenomenon that is oftentimes difficult for humans to understand, let alone computers. State of the art machine learning models fail to understand natural human language tendencies such as humor or sarcasm. Previous work has explored multimodal modeling of sarcasm through manipulation of modalities in self-attention and encoder initializations. However, these models consistently fail to classify sarcasm in samples where images contain words. To address this gap, our main contributions are twofold: we propose a method to (1) refine the cross-modal interaction by adding an optical character recognition modality to contextualize the image embeddings with the text contained within the images, and (2) a method to employ a symmetric attention block to ensure that images and text both contextualize each other equally (as opposed to using text as the primary modality and image only for contextualization). We conduct our experiments on Cai et al. (2019)’s public Twitter dataset. From our experiments, we found that our methods were able to address the error case of text in images, but our models failed to properly model cases where both the image and text modalities are necessary to understand the sarcasm, highlighting the need for inter-modal modeling in sarcasm detection. Our code is publicly available on GitHub¹.

1. Introduction

Sarcasm is defined as “the use of irony to mock or convey contempt.” Take the following sentence as an illustrative example: *“So happy about being snowed in... really makes it easy for me to get my tired fixed!”* In this example, the speaker conveys contempt about not being able to use their car by mockingly expressing displeasure about

being snowed in. A literal interpretation would not capture the speaker’s true intention, making sarcasm detection a challenging task.

1.1. Motivation

State of the art machine learning models are rapidly moving toward more human-like communication, but fail to understand natural human language tendencies such as humor and sarcasm. To create a human-like conversational bot and approach artificial general intelligence, the understanding of humor and natural human mannerisms is an integral next step for the machine learning community. Sarcasm is deeply integrated into daily human interaction – even in academic settings, such as textbooks or papers. Models that understand sarcasm will be able to better communicate with and be useful to humans.

Machine learning is essential to solving this problem, as rule-based approaches strongly rely on patterns identified in the data, making it infeasible to expand and apply to unseen patterns (Pan et al., 2020). Previous work has demonstrated the ability of machines to accurately detect sarcasm, and recent advances in deep learning enable us to move toward human-level performance in this field.

1.2. Sarcasm Detection

Incongruity or contradiction are often attributed as defining characteristics of sarcastic comments. In the previously mentioned example – *“So happy about being snowed in... really makes it easy for me to get my tired fixed!”* – the contradiction is between the audience’s expectation of what makes it easy for a person to drive and what the speaker expresses about their ease of driving in snow. Sarcasm is often difficult to detect due to its subtle nature and reliance on knowledge of the topic in question. Additionally, it is not often used in formal literature or conversation, and is instead used in informal conversation, such as social media. Thus, we focus on detecting sarcasm from Cai et al. (2019)’s publicly available multimodal sarcasm dataset², sourced from comments and images posted on Twitter.

¹Carnegie Mellon University, Pittsburgh, PA, USA. Correspondence to: Aneesha Sampath <aneeshas@andrew.cmu.edu>.

^{*} Denotes equal contribution.

¹<https://github.com/NehaNishikant/latr>

²<https://github.com/headacheboy/data-of-multimodal-sarcasm-detection>

1.3. Contributions

Our main contributions are twofold:

1. We propose a method to refine cross-modal interaction by **adding an optical character recognition modality** to contextualize the image embeddings with the text contained within the images adding a **2D positional embedding** for text within images to account for the spatial structure of images on social media.
2. We propose a **symmetric attention mechanism** to ensure that images and text both contextualize each other equally, as opposed to using text as the primary modality and image only for contextualization.

In the remainder of this paper, we discuss related works in the field of sarcasm, provide an overview of the current state of the art models on the Twitter dataset, perform extensive error analysis on those models, present a series of experiments designed to analyze the effects of each modality in enhancing model performance.

2. Related Work

2.1. Psychology of Sarcasm

Gibbs (1986) has studied the psycholinguistic effects of sarcasm. Interestingly, his studies show that humans do not take longer to process sarcastic language when compared to literal language. Additionally, he suggests that both literal and sarcastic interpretations are equally dependent on context, contrary to most beliefs that literal language can generally be interpreted without context. Thus, we look at sarcasm detection in a context-independent setting, and perform inference without knowledge of the user’s previous behavior.

2.2. Sarcasm Detection in Twitter

As mentioned in the motivation, sarcasm is often present in social media, leading researchers to use Twitter as a source of sarcastic comments. Chia et al. (2021) found that Twitter hashtags are often critical indicators of sarcastic comments, and also found that heavy pre-processing of text can lead to an oversimplification of the comment, can eliminate sarcastic indicators in the text. Pan et al. (2020) also looked at sarcasm in Twitter, but incorporated multimodal information via the images and text associated with each tweet. They proposed modeling incongruity in both inter-modal (across multiple modalities) and intra-modal (within the same modality) settings. They found that sarcasm was often detected in the presence of a contradiction between the content of an image and the content of the comment, which lead to the modeling of inter-modal sarcasm. However, they

also found that images were sometimes unrelated to the comment, and additionally modeled intra-modal sarcasm by treating hashtags as a separate modality, for comments such as *“What beautiful weather today! #not”*, in which the text and the hashtag contradict each other.

2.3. Multimodal Sarcasm Detection

Wu et al. (2021) explicitly model incongruity between modalities through an Incongruity-Aware Attention Network (IWAN). They propose modeling word-level incongruity between modalities as opposed to utterance-level or sentence-level modeling. They found that word-level modeling worked better than utterance-level modeling, and that multimodal information yielded better results than unimodal information.

As mentioned previously, Pan et al. (2020) modeled across modalities by accounting for incongruity between images and text, and modeled within the text modality by accounting for incongruity between text and hashtags. They manipulated the key, query, value mechanism in a self-attention block to use vision features as the key and value and text features as the query. This allowed them to capture incongruity between vision and text, as the textual features would bring importance to the incongruous features in the images.

While previous work has addressed the concern of multimodal sarcasm by accounting for incongruity, we observe that, to our knowledge, prior work has not accounted for the spatial structure of images that can be clear indicators of sarcasm. We found that this is especially prevalent in images posted on social media, so we focus our analysis on better image representations for sarcasm detection.

3. Dataset

We leveraged the Twitter dataset from Cai et al. (2019) for our experiments. In the following subsections, we detail our error analysis on the baseline models proposed by Cai et al. (2019) and Pan et al. (2020).

3.1. Dataset-Based Error Analysis

Many of the images in the misclassified samples included text within the images themselves. Upon sampling the test dataset, we found that about 50% of the images contain text, however, about 80-90% of the misclassified samples had images which contain text. As seen in Table 1, the misclassified images also have an underlying spatial structure to them. The first example is divided into four sections and is meant to be read from left to right. The second example has several numbers in the text and is intended to be read from top to bottom. Although the images have spatial structure, both Pan et al. (2020) and Cai et al. (2019) use ResNet (He et al., 2016) to embed the images. We hypothesize that this


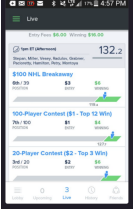
Image	Prediction	True
	Not Sarcastic	Sarcastic
	Not Sarcastic	Sarcastic

Table 1. Selected misclassified image examples when both baseline models predicted the wrong class. In both cases, the images contain large amounts of text, which are crucial to understanding the sarcastic nature of the tweets.

approach is not satisfactory, as the images used in tweets are a very specific subset of images and often have text that require spatial and sequential reasoning, as opposed to the ImageNet data (Deng et al., 2009) that ResNet was pre-trained on, which is often sourced from natural images or photographs.

3.2. Quantitative Analysis

Performance of baseline models shown in Table 2 and Table 3.2. While metrics of precision and recall seem fairly the same between and within models in 2, when we compare the models to each other and start taking intersections in Table 3.2, we notice that for the samples that only (Cai et al., 2019) misclassifies, the number of false negatives is much higher than the false positives. On the other hand, for the samples that only MsdBERT misclassifies, the number of false positives is much higher than the number of false negatives. The models seem to perform similarly, since most of the samples that each model classifies correctly is also classified correctly by the other. This suggests that there are some samples that are overall more complex to reason about than others.

4. Problem Statement

Our goal is to propose state of the art methods for multimodal sarcasm detection on the Twitter dataset. Each sample contains text (the comment of the tweet, including hashtags) and an image, which will both serve as the inputs to our model. The model prediction, \hat{y}_i is a binary class – 1 to represent that the tweet is sarcastic, and 0 to represent

that the tweet is not sarcastic. As in the baseline models, we use Cross Entropy Loss as our training objective, defined as:

$$J = - \sum_{i=1}^N [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)] + \lambda R$$

where y_i is the ground truth class, R is the standard L2 regularization term, and λ is the weight of R .

5. Multimodal Baseline Models

We compare our results to two baseline models – Cai et al. (2019)’s hierarchical fusion approach to sarcasm detection, and Pan et al. (2020)’s self and cross-attention approach to modeling both inter and intra-modal interactions.

5.1. Hierarchical Fusion

Cai et al. (2019) proposes a multi-modal hierarchical fusion model with modalities of text features, image features and image attributes.

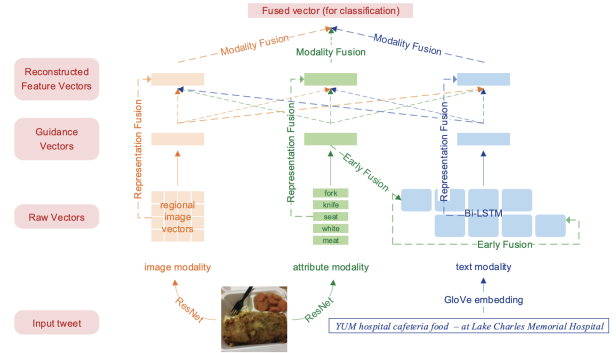


Figure 1. Cai et al. (2019)’s model for multimodal sarcasm detection.

Architecture Layout of Hierarchical Fusion:

- Early Fusion:** attribute features are used to initialize a Bi-LSTM network. This resulting network is used to extract the text features.
- Representation Fusion:** all three modality features are transformed into reconstructed representation vectors.
- Modality Fusion:** this layer performs a weighted average to the vectors and pumps them to a classification layer to yield the final result.

5.2. Self and Cross-Attention

Pan et al. (2020) utilizes the design key, query, value design of the self-attention mechanism to capture “inter-modality

Model	Accuracy	Precision	Recall	F1-Score
Hierarchical Fusion	0.8306	0.7958	0.799	0.7974
MsdBERT	0.8406	0.8417	0.8226	0.8294

Table 2. Performance of our re-implementation of the baseline models from Cai et al. (2019) (Hierarchical Fusion) and Pan et al. (2020) (MsdBERT) on the hold-out test set for evaluation metrics of accuracy, precision, recall, and F1-score.

Category	Both	Hierarchical Fusion	MsdBERT
Wrong	245	163	139
Correct	1862	139	163
False Positive	81	125	21
False Negative	164	38	118

This table shows which samples are wrongly or correctly classified by which models out of a total of 2409 samples in the test set

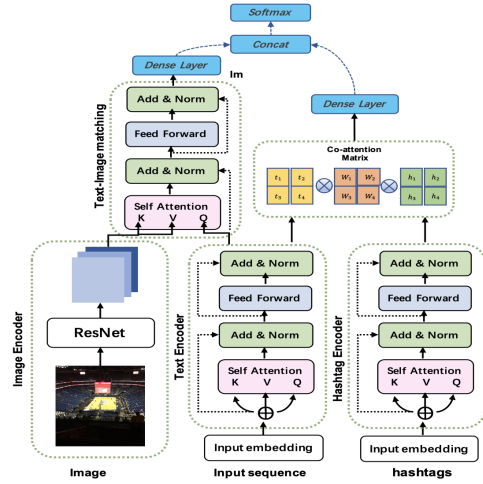


Figure 2. Pan et al. (2020)'s model for multimodal sarcasm detection.

attention.” The goal is to capture inter-modality incongruity (between Twitter images and their captions) in order to predict if an example is an instance of sarcasm or not.

The architecture is formed from three main components: the (1) Image and Text Processing Module, the (2) Inter-modality Attention Module, and the (3) Intra-modality Attention Module.

The inter-modality attention module is an adaptation of self-attention where the key and value are the image embeddings and the query is the text embedding. This is designed so that the text tokens are contextualized by the image regions that contradict the text.

The intra-modality attention module is a co-attention matrix between the caption embedding and the hashtag embedding.

This is to capture relationships between captions and the hashtags.

6. Experimental Methodology

6.1. Twitter Dataset

This Twitter was gathered by Cai et al. (2019) and consists of picture and caption pairs sourced from Twitter. The authors cite the importance of multimodality in detecting and understanding sarcasm. For example, the phrase “What a wonderful weather!” may not sound sarcastic until paired with dark clouds in the attached picture (Cai et al., 2019).

Positive samples are collected by searching for specific hashtags such as “sarcasm”. Negative samples are collected randomly by looking for tweets without hashtags. The authors collect about 24,000 total samples, and provide splits into train, evaluation, and test sets using a 80 : 10 : 10 ratio. The labels of the evaluation set and the test set are manually verified for quality.

6.2. Evaluation Metrics

We evaluate our model on the hold-out test set provided by Cai et al. (2019). We evaluate our models on accuracy to align with the evaluation metrics of Pan et al. (2020) and Cai et al. (2019). We use the default hyperparameters from both models in order to reproduce their results.

7. Method

7.1. Layout Aware Image Encoder

To our knowledge, previous work does not acknowledge that sarcasm detection on social media datasets should be treated differently from other interactions, such as TV show clips. Social media posts often contain images with lots of text that must be read in a particular order and contain structure.

For example, both Cai et al. (2019) and Pan et al. (2020) use a ResNet image encoder that was pre-trained on ImageNet for the images from the tweet. ResNet is trained on the ImageNet dataset (Deng et al., 2009), which is a collection on natural images and photographs. However, as seen in Table 1, memes are a very different genre of images since they rely heavily on spatial structure and optical character recognition (OCR) and should be treated accordingly.

Our error analysis specifically highlights images with text on them and images with spatial structure. We hypothesize that a spatially-aware image encoder would consider both of these. We wanted an image encoder that took in account the actual text and where it was on the image. Thus, we settled on the Layout Aware Transformer (LaTr) (Biten et al., 2022).

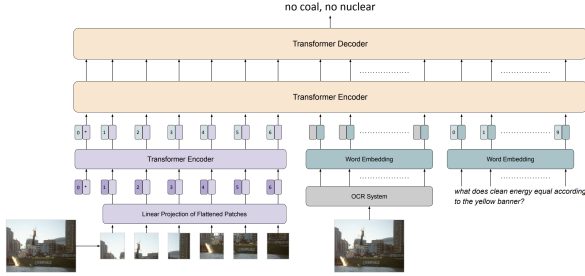


Figure 3. Layout-Aware Transformer architecture. The LaTr contains a 2D Spatial Embedding to account for the position of text within an image, and also had layout-aware pre-training, which treated documents as images to learn exact location of OCR tokens to be used.

LaTr was originally made for the task of Visual Question Answering. Given an image and a corresponding question, LaTr generates a text response for the question.

LaTr considers three modalities: image, text, and OCR. This third modality was the key for why we chose this model. The image is encoded using ViT, which divides the image into a fixed number of even patches. The text is embedded using the Text-to-Text Transfer Transformer encoder (Raffel et al., 2020). The OCR modality consists of the actual OCR tokens as well as a 2D spatial embedding representing a bounding box for where the token is on the image. Finally, the image patch embeddings, text word embeddings, and OCR token embeddings as well as their respective positional embeddings are passed into T5, which then generates an answer in natural language.

Furthermore, LaTr is optimized for text in images due to its Layout-Aware pre-training task. The model is pre-trained on scanned documents, which are treated as images, and then learns accurate OCR embeddings by predicting the masked tokens.

To adapt LaTr for the task of sarcasm detection, we fine-tuned LaTr on the Twitter Dataset by adapting the classification dataset for question-answering. Instead of asking a question, embed the Twitter caption and use that to contextualize the image (and vice-versa). We also modified the dataset in order to output the answer in natural language, so each sample had an answer of "Yes" or "No" to answer the question: "Is this sarcastic?"

For the OCR information, LaTr used Rosetta, a system for

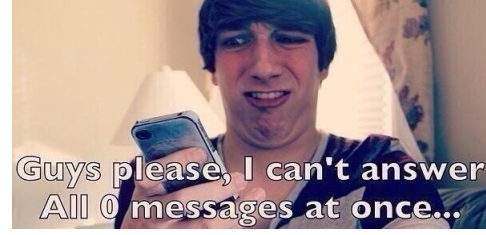


Figure 4. Tweet caption: "me and my mentions right now hahaha". In this example, the sarcasm is all contained within the image

understanding text for images (Borisyuk et al., 2018). However, this model is not publically available, so we instead used Google’s Tesseract-OCR engine (Smith, 2007).

7.2. Symmetric Attention

The MsdBERT model (Pan et al., 2020) computes attention between images and text as well as between texts and hashtags, as seen in Figure 2. The attention between the text and hashtag is computed via a (symmetric) co-attention matrix. The authors modify a typical transformer self-attention block by treating the text as the query and the image as the key and value. Since the dimensions of the query are retained after using attention weights, the image embeddings are used to contextualize the text embeddings, which means the current MsdBERT model treats text as the primary modality. However, that assumption is unfounded, and often the sarcasm is contained within the image as seen in Figure 4.

Thus, experimented with modifying the attention inputs for our model and analyzed the difference between unchaed MsdBERT (Pan et al., 2020) and where we switch key and value, i.e: Image is the query, Text is the key/value.

8. Results and Discussion

8.1. LaTr

Modality	Model	Accuracy
Visual	MsdBERT	0.7260
Visual	LaTr	0.7331
Text	MsdBERT	0.8385
Text	LaTr	0.6019
Visual+Text	HF	0.8344
Visual+Text	MsdBERT	0.8605
Visual+Text	LaTr	0.7090

Table 3. Ablation studies for LaTr and baseline models. The HF model is the hierarchical fusion model proposed by Cai et al. (2019), and MsdBERT is the model proposed by Pan et al. (2020).

We performed ablation studies on LaTr for the visual and text modalities. When given only the image, we use LaTr

as a VQA model as it was intended. We keep the image and thus the OCR information. For the text input, we ask a question, "Is this sarcastic?" When limiting LaTr to only text input, we keep the Twitter caption's text, and we mask out the image entirely, so the model is fed only a black square. This effectively removes the image and OCR modality information.

Pan et al. (2020) also performed ablation studies in their research which we included in Table 3. We included the results by Cai et al. (2019) in the "Visual+Text" row, since they did not release modality-specific results.

As shown in Table 3, LaTr with "Visual"-only performed better than "Visual+Text". This is surprising and shows that LaTr cannot make meaningful use of the textual information. This is corroborated by the fact that the performance for "Text"-only for LaTr is very low at about 60%.

On the other hand, MsdBERT does performs well for "Text"-only and only marginally improves performance when adding the visual modality, as shown in the "Visual+Text" row of Table 3. When MsdBERT masks out the image, they still maintain two out of their three modalities: text and hashtag. This is one possibility for why the performance of MsdBERT is largely unaffected when adding in the visual modality. Furthermore, these results indicate that the hashtag modality may be essential to detecting sarcasm in Twitter data, as MsdBERT found that often tweets were accompanied by short phrases such as "#not", which provided the necessary incongruity to detect sarcasm (Pan et al., 2020).

Although LaTr does not perform well on the text modality, the results indicate the LaTr improves performance on processing images with spatial structure. For "Visual"-only, MsdBERT's accuracy significantly decreases when compared to the "Visual+Text" setting as well as the "Text"-only setting. LaTr's image encoder appears to be better suited for Twitter data, likely because of the 2D spatial embedding and the layout-aware pre-training.

Motivated by these results, we created an ensemble model to leverage the strengths of both MsdBERT and LaTr. To accomplish this, we used the text embeddings of MsdBERT in the LaTr model instead of T5 embeddings. We saved the output after the hashtag and text cross-attention and ported them into the LaTr model for inference only. We used the LaTr visual encoder trained on "Visual+Text" in the "Visual+Text" setting, and used the ablated LaTr visual encoder trained on "Visual"-only for the "Visual" setting. These results can be seen in Table 4.

From Table 4, there are negligible improvements when switching the T5 encoder to MsdBERT's text encoder. While this seems to negate our hypothesis that leveraging the strengths from each model and ensembling them would



Figure 5. A mis-classified sample of LaTr caused by incorrect OCR information.

boost performance, we acknowledge that introducing these changes only during inference may inhibit the capabilities of the model, since the fine-tuned LaTr likely learned to ignore the text modality and more heavily weight the visual modality (due to the poor performance of text, shown in our ablation studies in Table 3. This is most evident in the "Visual+Text" setting, as the accuracy is identical to four decimal points of precision. This also further corroborates our hypothesis as to why there was no improvement when adding the text modality to our LaTr ablation studies (Table 3).

Motivated by these results, one idea for future work would be to have the MsdBERT text encodings at training time in addition to inference time. In this setup, would fine-tune LaTr from scratch with the MsdBERT text encoder instead of using T5 encoder that currently is used by LaTr.

8.1.1. ERROR ANALYSIS

While LaTr didn't do as well as the baselines as seen in Table 3, Table 5 shows that out of the samples that (Cai et al., 2019) and MsdBERT got incorrect, LaTr is able to correct about half. This shows that LaTr is good at a very different subset of images.

In terms of weaknesses, LaTr is dependent on the quality of the OCR. For example, consider Figure 5. This example was incorrectly classified by LaTr, but the OCR information was incorrect. The word "Sarcasm" was recognized as the word "race".

8.1.2. NOVELTY

The main novelty in our approach is adapting LaTr, a VQA model, for a sarcasm detection task. To our knowledge, no prior work on sarcasm detection accounts for spatial structure in images. We performed an extensive study on layout aware training for sarcasm detection via ablation studies, model ensembling, and error analysis.

Model	Training Modalities	Inference Modalities	Accuracy
LaTr	Visual (LaTr)	Visual (LaTr)	0.7331
LaTr	Visual+Text (LaTr)	Visual+Text (LaTr)	0.7090
LaTr+MsdBERT	Visual (LaTr)	Visual (LaTr) + Text (MsdBERT)	0.7335
LaTr+MsdBERT	Visual+Text (LaTr)	Visual (LaTr) + Text (MsdBERT)	0.7090

Table 4. Ablation studies for LaTr and baseline models. The HF model is the hierarchical fusion model proposed by Cai et al. (2019), and MsdBERT is the model proposed by Pan et al. (2020).

Model	% Corrected by LaTr
HF	0.4853
MsdBERT	0.5677

Table 5. Examples that LaTr corrects out of those that the baseline models get incorrect.

8.2. Symmetric Attention

Another motivation for this idea is present in Table 3 and our discussion in section 8.1 about MsdBERT. MsdBERT uses text very well, as it does well for “Text”-only, but experiences little improvement for “Visual+Text” and has worse results for “Visual”-only. One hypothesis for this is that since text is treated as the primary modality, the image information isn’t used to its full potential by MsdBERT.

Primary Modality	Accuracy
Text	0.8605
Visual	0.6021

Table 6. Results for experimental setups for comparing the different primary modalities in the approach proposed by Pan et al. (2020).

8.2.1. ERROR ANALYSIS

Our results are shown in Table 6. As seen in the Primary “Visual” row, the accuracy significantly drops, negating our hypothesis. We suspect that this occurs because the image encoder that MsdBERT uses – ResNet – is unsuitable for this dataset, as most of the Twitter images are not photographs or natural images, which is what ResNet uses for pre-training. This, using vision as the primary modality does not help improve performance. Future work could look to extend this idea by leveraging an image encoder that is more suited to the task.

To improve this in the future, we would visualize attention maps to understand what parts of the image and text are attending to each other. Further, we could analyze a combination of Primary-“Text” and Primary-“Visual” by performing both kinds of attention and then concatenating or fusing each aligned vector.



Figure 6. Tweet caption; “happens all the time ! lol”. This is an example of a tweet positively labeled as sarcastic, however we observe that this is a generally humorous comment, but not necessarily sarcastic.

8.2.2. NOVELTY

The main novelty in our approach is the ablation analysis-driven approach (as seen in 8.1) to perform a study of different attention inputs.

9. Future Work and limitations

9.1. The Task

Sarcasm is an inherently subjective task, one that is often confused with humor. With the dataset from Cai et al. (2019) specifically, we observed that positive labels for sarcasm would likely be more accurately represented by positive labels for *humor*, not sarcasm. For example, consider the sample in Figure 6. We find that the content of this sample is mostly humorous and not truly sarcastic. This is relevant because most of the prior work defines sarcasm to be an incongruity between or within modalities. However, samples such as the one in Figure 6 do not appear to contain an incongruity.

Furthermore, humor can be due to an unexpected incongruity, but often also comes from references, the way inside jokes work. Consider the example in Figure 7. This meme is funny because we collectively know that April 1st is April Fool’s day and this functions almost like an inside joke that we all know. However, the model doesn’t know this.

Thus, many of the assumptions made in (Cai et al., 2019), (Pan et al., 2020) and other prior work don’t always hold.



Figure 7. Tweet caption; "when it 's 2nd april and she is still pregnant .. juser emoji_19 emoji_19 emoji_19". This is an example of a positively labeled image that is funny but not because of incongruity.

9.2. Extensions

9.2.1. COUNTERFACTUAL DATA AUGMENTATION

We believe that our models could benefit from counterfactual data generation through the addition of negative samples. If an off-the-shelf image-captioning model is used on positively-labeled samples, negative samples would be generated as we would join the image from the positive sample with the neutral, generated caption. This could be used for any model, but could especially be helpful for LaTr. This would ideally help boost the model's text understanding, which currently is very poor. The model would be forced to understand the difference between the true sarcastic text and the counterfactual non-sarcastic text.

9.2.2. ENSEMBLE TRAINING

Another idea we have for future work is, as discussed in section 8.1, to include MsdBERT text embeddings at train time. Or, to extend this further, the MsdBERT text embeddings could be trained jointly with the LaTr image encoder, instead of sequentially.

9.2.3. CHARACTER-LEVEL EMBEDDINGS

Sample	Prediction	True
yeah it was lit and yeah i can handle it	Not Sarcastic	Sarcastic
damn! we lurveee our new logo. wait, not ours, sarawak fa's	Not Sarcastic	Sarcastic
nooo my name is just for show	Not Sarcastic	Sarcastic

Table 7. Selected misclassified textual examples when both baseline models predicted the wrong class. The character-level information – including the spelling – provide key insights into detecting the sarcastic tone of the tweets.

From analysis of our misclassified examples, as seen in Ta-

ble 7, we observed that captions often have domain-specific language – specifically, language that is unique to Twitter. Both the model from Cai et al. (2019) as well as (Pan et al., 2020) fail to account for the domain-specific language, as they use GloVe and BERT embeddings respectively. Twitter language is arguably its own dialect, and typical text encoders – such as BERT – are trained on more corpora with more formal language, such as Wikipedia, which have drastically different semantic and stylistic patterns. For example, the word "lit" in the first example in Table 7 is intended to mean "cool", but a BERT embedding would likely be unable to capture the informal nature of the word in the Twitter context. Furthermore, Twitter users often *intentionally* misspell words in order to communicate exaggeration or emphasis on certain parts of the sentence. As in the second and third examples in Table 7, "lurveee" is an intentional misspelling of "love" and "nooo" is an intentional misspelling of "no", respectively, but the choice of spelling adds a emphasis and changes the meanings of the words. Typical word embeddings – such as GloVe or BERT – are unable to capture this character-level word information since they are not pre-trained on corpora with more informal language.

References

- Biten, A. F., Litman, R., Xie, Y., Appalaraju, S., and Manmatha, R. Latr: Layout-aware transformer for scene-text vqa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16548–16558, 2022.
- Borisjuk, F., Gordo, A., and Sivakumar, V. Rosetta: Large scale system for text detection and recognition in images. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pp. 71–79, 2018.
- Cai, Y., Cai, H., and Wan, X. Multi-modal sarcasm detection in Twitter with hierarchical fusion model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 2506–2515, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1239. URL <https://aclanthology.org/P19-1239>.
- Chia, Z. L., Ptaszynski, M., Masui, F., Leliwa, G., and Wroczynski, M. Machine learning and feature engineering-based study into sarcasm and irony classification with application to cyberbullying detection. *Information Processing & Management*, 58(4):102600, 2021. ISSN 0306-4573.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.
- Gibbs, R. W. On the psycholinguistics of sarcasm. In *Journal of Experimental Psychology: General*, pp. 3–15, 1986.
- He, K., Zhang, X., Ren, S., and Sun, J. Identity mappings in deep residual networks. *CoRR*, abs/1603.05027, 2016. URL <http://arxiv.org/abs/1603.05027>.
- Pan, H., Lin, Z., Fu, P., Qi, Y., and Wang, W. Modeling intra and inter-modality incongruity for multi-modal sarcasm detection. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1383–1392, 2020.
- Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P. J., et al. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.
- Smith, R. An overview of the tesseract ocr engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, volume 2, pp. 629–633. IEEE, 2007.
- Wu, Y., Zhao, Y., Lu, X., Qin, B., Wu, Y., Sheng, J., and Li, J. Modeling incongruity between modalities for multimodal sarcasm detection. *IEEE MultiMedia*, 28(2):86–95, 2021. doi: 10.1109/MMUL.2021.3069097.