Exploiting BERT For Multimodal Target Sentiment Classification Through Input Space Translation

Zaid Khan, Yun Fu
{khan.za,y.fu}@northeastern.edu
Northeastern University
Boston, Massachusetts, USA

ABSTRACT
Multimodal target/aspect sentiment classification combines multimodal sentiment analysis and aspect/target sentiment classification. The goal of the task is to combine vision and language to understand the sentiment towards a target entity in a sentence. Twitter is an ideal setting for the task because it is inherently multimodal, highly emotional, and affects real world events. However, multimodal tweets are short and accompanied by complex, possibly irrelevant images. We introduce a two-stream model that translates images in input space using an object-aware transformer followed by a single-pass non-autoregressive text generation approach. We then leverage the translation to construct an auxiliary sentence that provides multimodal information to a language model. Our approach increases the amount of text available to the language model and distills the object-level information in complex images. We achieve state-of-the-art performance on two multimodal Twitter datasets without modifying the internals of the language model to accept multimodal data, demonstrating the effectiveness of our translation. In addition, we explain a failure mode of a popular approach for aspect sentiment analysis when applied to tweets. Our code is available at https://github.com/codezakh/exploiting-BERT-thru-translation.

CCS CONCEPTS
• Computing methodologies → Neural networks; • Information systems → Sentiment analysis.

KEYWORDS
sentiment analysis, BERT, deep learning, vision-language, twitter

1 INTRODUCTION
The spread of emotional narratives is conjectured to drive large-scale events [32] and Twitter is a critical venue for spreading emotional narratives [10] and misinformation [11, 38]. Understanding the spread of sentiment on the scale of social media requires computational approaches to emotion understanding. A holistic computational understanding of emotion requires comprehension of aspect/target - towards what is the sentiment directed? In addition, the inherent multimodality of tweets requires models capable of multimodal sentiment understanding, specifically vision/language, to leverage the emotional information and additional context carried by the images accompanying tweets.
While aspect sentiment analysis and multimodal sentiment analysis are established fields, multimodal aspect sentiment analysis - a combination of the two - is relatively new. Directly applying methods developed for aspect sentiment analysis or multimodal sentiment analysis to multimodal tweets is challenging for the following reasons: (1) tweets are shorter and less informative than the review texts commonly used for aspect sentiment analysis, (2) the visual modality is much more likely to be irrelevant/noisy for tweets than for videos of human speakers commonly used for multimodal sentiment analysis, (3) the visual scenes accompanying tweets are significantly more complex than videos of single human speakers.

We propose a model to address these challenges. Our approach adapts a transformer architecture for object detection to generate text instead, then uses the resulting model to translate images into the input space of a pretrained language model. We then feed the translated image into a BERT language model by constructing an auxiliary sentence, and use the resulting encoding for multimodal aspect sentiment analysis. This solves two problems: first, we increase the amount of text available for the language model, and second, the translation preserves complex scene information appropriate for understanding social media images. Finally, the translation renders the fusion relatively interpretable, because the image is expressed in natural language.

Our work is similar in spirit to [12], who proposed using image captions for the visual question answering task. However, [12] requires external information, does not leverage recent large language models, and is for the VQA task rather than sentiment classification. We draw inspiration from BERT-Pair-QA [33], which constructs an auxiliary sentence to aid aspect sentiment classification. Unlike their auxiliary sentence, we use the auxiliary sentence to pass multimodal information into BERT. In principle, our work has similarities to MCTN [28], which cyclically translates modalities for fusion, but we perform translation in the input space rather than the feature space, and do not need a GAN. Our approach shares characteristics with TomBERT [42], but unlike TomBERT, we perform multimodal fusion without modifying BERT, allowing us to achieve higher performance. We rely heavily on the architecture introduced by DETR [4] for end-to-end object detection with transformers, but we use the DETR layers to implement a non-autoregressive text generator rather than an object detector. Our contributions are the following:

1. A novel architecture for multimodal aspect sentiment analysis that translates in input space and utilizes an auxiliary sentence for fusion through a large pretrained language model.
2. We investigate and explain the performance degradation of state-of-the-art aspect sentiment analysis models in the short-text twitter setting.
3. We adapt the DETR [4] architecture for non-autoregressive text generation.

## 2 RELATED WORK

### 2.1 Multimodal Sentiment Classification

The goal of multimodal sentiment analysis is to regress or classify the overall sentiment of an utterance using acoustic, visual, and language cues. Because multimodal sentiment analysis is a large and well-established field, we direct the reader to [2, 21, 29] for an overview of the field, and MISA [8], MAG [31], and M3ER [20] as representative of recent state of the art works. We restrict our scope to describing differences and similarities between our setting and the classical multimodal sentiment analysis setting. The primary difference is that multimodal sentiment analysis datsets such as [1, 3, 27, 43] traditionally focus on videos of a speaker with their face in focus. In this scenario, it is overwhelmingly likely the visual modality is relevant, in contrast to the social media setting, where

| Label  | Tweet                                                                 | With Tweet Only | With Tweet + Image Context |
|--------|------------------------------------------------------------------------|-----------------|----------------------------|
| NEGATIVE | Harriette moved back to the # Chicago with a baby 6 months after her mom died. And it’s been terrible. # NextDayChi 17 | [X] NEUTRAL     | [X] NEGATIVE               |
|         | A woman is sitting on a sidewalk looking at her cell phone.            |                 | A woman is sitting on a sidewalk looking at her cell phone. |
| NEUTRAL | Starlin Castro with better glove work in the 6th, tagging out # Chicago | [X] POSITIVE    | [X] NEUTRAL                |
|         | A baseball player sliding into a base while another player tries to tag him out. |                 | A baseball player sliding into a base while another player tries to tag him out. |
| POSITIVE | RT @ AJPenn # Barry Allen Bruce Wayne. Oliver Queen. # Flash # Arrow # Gotham # Diehappy | [X] NEUTRAL     | [X] POSITIVE               |
|         | A group of young men standing next to each other.                     |                 | A group of young men standing next to each other. |

Figure 2: Examples the text-only BERT predicts incorrectly, but our proposed method gets correct. In Row 2, notice how the proposed method uses the image context to recognize the target is the player being tagged out and downgrades the sentiment.
images accompanying tweets can be completely irrelevant [37] or lack emotional content. In addition, multimodal sentiment analysis models operate on sequences of multimodal data, whereas our setting is more similar to the vision-language setting, where only the language input is a sequence, and the accompanying modality is a single image.

### 2.2 Target/Aspect Sentiment Classification Without Architectural Modifications

The goal of target/aspect sentiment classification is to classify the sentiment towards a target mentioned in text. The power and widespread availability of large pretrained language models such as BERT[17], roBERTa [15], GPT-2 [30], and XLNet[41] has resulted in pretrained language models dominating the field. Although architectural modifications to large pretrained language models have been successful [44], studies have shown that architectural modifications to large language models are [24] brittle and often do not transfer across implementations and applications. In addition, [22] finds that reproducing the results of modifications to BERT architectures is difficult. However, the remarkable ability of transformers to adapt without architectural modifications [18], as well as the practical desirability of deploying [26] popular baseline architectures without modifications, has motivated research into exploiting large language models for target/aspect sentiment classification without architectural modifications.

These methods either rely on data augmentation, novel inference procedures, enhanced training procedures [39], or a combination of the three. Our work fits most closely into this mold. The most related work to ours in the BERT-Pair model introduced by Sun [33]. The BERT-Pair model converts the multi-class aspect sentiment classification problem into a series of binary classification problems by using BERT in sentence-pair mode. To form the second sentence of the pair, they construct an auxiliary sentence that encodes label information as a query. In contrast to BERT-Pair [33], we encode multimodal image context through translation into the auxiliary sentence. Furthermore, we do not require 3 passes for every training instance during training and inference, as BERT-Pair [33] does. In addition, BERT-Pair [33] cannot use multimodal data.

### 2.3 Transformers for Vision-Language Tasks

Vision-language transformers are natively multimodal, taking in both language and visual input. The most similar to our work are two-stream encoders, such as ViLBERT [17]. These use separate input streams for vision and language before fusing them at a later step. An ideal task for vision-language transformers is object detection and image captioning. Our caption transformer is directly inspired by DETR [4], but in contrast to DETR [4], our DETR-based caption transformer predicts a natural language image description instead of a set of object detections.

In this regard, our work makes use of a non-autoregressive text-generation approach - predicting all tokens in one pass - which makes it similar to works such as LaBERT [5] and Transform and Tell [34], although we do not use a language model during the image captioning.

### Table 1: Key Differences Between The Proposed Model and Existing Approaches

|                        | Proposed | TomBERT | ViLBERT |
|------------------------|----------|---------|---------|
| Doesn’t Modify Language Model | ✓        | X       | X       |
| Trains End-to-End      | X        | ✓       | ✓       |
| Translates In Input Space | ✓        | ✓       | X       |
| Uses Object Level Information | ✓        | X       | ✓       |
| Explicitly Interpretable Fusion | ✓        | X       | X       |

### 2.4 Multimodal Target Oriented Sentiment Classification

Multimodal target-oriented sentiment classification is a new field that combines multimodal sentiment analysis and target-oriented sentiment classification. The most common setting for this task is multimedia collected from online social media. Xu [40] introduced a Multi-Interactive Memory Network for multimodal aspect sentiment classification, and a private Chinese-language dataset called Multi-ZOL curated from review sites. The YouTubean dataset was introduced by Marrese-Taylor [19], and contains target annotations 500 YouTube review videos.

The work most closely related to ours is TomBERT [42], which introduces two multimodal tweet datasets with target annotations that we use, as well as a target-oriented multimodal BERT model that we compare to. TomBERT builds on top of the baseline BERT architecture by adding target-sensitive visual attention and adding more self-attention layers to capture inter-modality dynamics. In contrast to TomBERT, we do not modify the baseline BERT model (other than one linear layer), and we do not add additional self-attention layers.

### 3 Fusion and Sentiment Analysis Through Input Translation

#### 3.1 Problem Definition

Let \( M \) be a set of multimodal samples. Each sample \( m_i \in M \) consists of a sentence \( S_i = (w_1, w_2, \ldots, w_n) \), where \( n \) is the number of words. Alongside the sentence is an image \( I_i \) and an opinion target \( T_i \), which is a subsequence of \( S_i \). The opinion target is assigned a label \( y_i \in \{\text{negative, neutral, positive}\} \). Our goal is to learn a function \( f : (T_i, S_i, I_i) \rightarrow y_i \). In short, given a tweet like "Einstein is my favorite scientist, but I hate physics", and the posted image, the model should be able to predict positive for the target "Einstein" and negative for the target "physics".

#### 3.2 Overview

Our model (Fig 3) consists of two parts. Given a multimodal input sample \( m_i = (T_i, S_i, I_i) \) consisting of a target \( T_i \), input sentence \( S_i \), and image \( I_i \), we first pass the image through a captioning transformer. Let \( 3, H, W \) represent the number of channels, width, and height of an image respectively. The captioning transformer turns the image \( I_i \in \mathbb{R}^{3 \times H \times W} \) into a vector \( \hat{I}_i \in \mathbb{R}^{l} \), where \( l \) is the maximum output length of the captioning transformer. Thus, the image is converted from a 3-D tensor representing continuous data into a 1D vector of integers representing symbolic data. The tuple
\( \hat{l}, \hat{s}, \hat{t} \in \mathbb{N}_0^n \) representing image context, the sentence, and the target now reside in the same space, which is the vector space of \( l \)-dimensional vectors with elements in \( \mathbb{N}_0 \). By sharing tokenizers and the vocabulary mapping symbolic words to elements in \( \mathbb{N}_0 \) between the captioning transformer and the large language model, the language model can understand \( \hat{l} \) by constructing an auxiliary aux; sentence from the target \( \hat{t} \) and the natural language image description \( \hat{s} \), and using the large language model in sentence-pair classification mode, feeding the pair \((\hat{s}; \text{Aux})\) through the language model to get a sentiment prediction \( y_t \) for the target. The overall training procedure is described in Algorithm 1.

### 3.3 Input Translation with Caption Transformers

We propose to translate an image \( I_i \in \mathbb{R}^{3 \times H \times W} \) into an element in the input space \( \hat{l_i} \in \mathbb{N}_0^n \) representing symbolic natural language input for a large language model, specifically BERT [6].

#### 3.3.1 Transformer Encoder

Given the input image \( I_i \in \mathbb{R}^{3 \times H \times W} \), we apply a CNN backbone (ResNet 101 [9]) to generate an activation map \( f \in \mathbb{C} \times H \times W \), where \( C = 2048 \), and \( H = W = \frac{12}{7} \). We then apply a 1x1 convolution to the channel dimension \( C \) to reduce it to \( d << C \), where \( d = 256 \). The new feature map \( z_0 \in \mathbb{R}^{d \times H \times W} \) is then flattened along the spatial dimensions, resulting in \( z_0 \in \mathbb{R}^{d \times HW} \).
The DETR encoder architecture is a standard multi-head key, query, and value architecture. We refer the reader to [4] for details, and provide an overview here.

A DETR encoder layer consists of multiple attention heads, each of which is parameterized by a weight tensor \( T' \in \mathbb{R}^{d \times d \times d} \). The attention head first computes the key, query, and value embeddings:

\[
[Q;K;V] = [T'_1(X_q+P_q);T'_2(X_k+P_k);T'_3X_{kv}]
\]

(1)

where \( P_q \in \mathbb{R}^{d \times N_q} \) and \( P_k \in \mathbb{R}^{d \times N_k} \) are position embeddings for the key and query respectively. Because the transformer is permutation invariant, the positional embeddings are added to each transformer encoder layer to help the transformer learn dependencies. The weight tensor \( T' \) is the concatenation of \( T'_1, T'_2, T'_3 \). The attention then computes the attention weights, which represents the importance of each element in the key-value sequence to each element of the query sequence. Let \( i \) be an index of the query sequence and \( j \) be the index of the key-value sequence, then the attention weights \( a_{i,j} \) are given by

\[
a_{i,j} = \frac{1}{\sqrt{d_i}} Q'_i K_j Z_i
\]

where \( Z_i = \sum_{j=1}^{N_k} e^{\frac{1}{\sqrt{d_i}} Q'_i K_j} \).

The output of the attention head is the weighted sum of the value sequence \( V \), using the attention as weights. The \( i-th \) row is given by \( \text{att}_i (X_q, X_{kv}, T') = \sum_{j=1}^{N_k} a_{i,j} V_j \). We then combine the single attention heads of a single encoder layer into a multi-head attention by concatenating the outputs of single attention heads \( 1 \ldots M \)

\[
X'_q = [\text{att}_1(X_q,X_{kv},T_1); \ldots ; \text{att}_M(X_q,X_{kv},T_M)]
\]

\[
\tilde{X}_q = \text{layernorm}(X_q + \text{dropout}([X'_q]))
\]

followed by a linear projection and a residual connection. This is the standard formulation of key, query, value attention by [36].

3.3.2 Transformer Decoder. The decoder transforms \( l \) embeddings of size \( d \), where \( l \) is the maximum sequence length the transformer can attend to. In contrast to DETR [4], which inputs object queries and trains the transformer with a set-matching loss, we use the decoder for non-autoregressive text generation by predicting a description for the input image in one forward pass. We generate a description as follows. Let \( \sigma^0 = \text{concat}(\text{TokenID}[\text{CLS}], \text{zeros}(l-1)) \) be the concatenation of the BERT token ID for the [CLS] token with the \( l-1 \) dimensional zero vector. The resulting vector \( \sigma^0 \) becomes a prompt for the decoder, indicating the start of a sentence. We add a standard positional encoding to \( \sigma^0 \) from [4] and forward it through the encoder to get \( l \) embeddings of size \( d \). In the decoder, we apply a 3-layer feedforward network with ReLU activations [23] to predict a word from BERT’s vocabulary at each query position. The token [PAD] is used to represent the absence of words at the end of the sentence. At each position in the prompt sequence \( \sigma^* \) (except for the first position, which contains a control code), we predict a 30522-dimensional probability distribution over BERT’s vocabulary:

\[
p(t|\sigma^*, i) = \text{SoftMax}(W_3 \times R(W_2 \times R(W_1 \times D_{ENC}(\sigma^*, i))))
\]

where \( D_{ENC}(\sigma^*)_i \) is the embedding learned by the DETR encoder at the \( i-th \) position of the prompt \( \sigma^* \), and \( W_1, W_2, W_3 \) are learned weight matrices, and \( R \) is the ReLU activation function.

\[
L_1 = \min \sum_{i=1}^{l} -\log p(t_i = t^*_i)
\]

(3)

We train the model by minimizing the sum of the negative cross-entropy over all positions, as depicted in Eq. 3. Here, \( t^*_i \) is the probability distribution from Eq. 2, and \( t_i = [\text{PAD}] \) and 0 otherwise.

3.4 Fusion Through The Auxiliary Sentence

Once the caption transformer described in Section 3.3 has been trained, we can use it to translate input images into natural language descriptions of the image. We adapt the auxiliary question [6, 33] mechanism for this. In the auxiliary sentence method, BERT is used in sentence-pair classification mode. In sentence-pair classification mode, input to BERT takes the sentence-pair form

\[
[\text{CLS}] A_1 \ldots A_{A-len} [\text{SEP}] B_1 \ldots B_{B-len} [\text{PAD}]
\]

(4)

where \( A_i \) are the tokens of sentence \( A \), and \( B_i \) are the tokens of sentence \( B \). For aspect sentiment classification, Sentence \( A \) is typically the text to be classified, and Sentence \( B \) contains auxiliary information used to prime BERT, such as the name of the desired sentiment target. Instead of encoding label information in the auxiliary question as in BERT-Pair [33], we enrich it with multimodal...
We carry out experiments to answer the following research questions:

- **RQ1**: Does our approach using input-space translation improve over a purely text-based language model, and is it competitive with other multimodal and unimodal approaches?
- **RQ2**: Does the image description help BERT to understand the relationship of the image to the tweet, and the target to the image?
- **RQ3**: How does the amount of multimodal information affect accuracy of predictions in our model?
- **RQ4**: What are the failure modes of powerful state of the art unimodal models on short-text social media data?

To evaluate our research questions, we use two benchmark datasets for target oriented multimodal sentiment classification, Twitter-15 and Twitter-17, both introduced by [42] and building on earlier work [16, 45]. Twitter-15 and Twitter-17 consists of multimodal tweets, where each multimodal tweet consists of text, an image posted alongside the tweet, targets within the tweet, and the sentiment of each target. Each target is given a label from the set {negative, neutral, positive}, and the task is a standard multi-class classification problem. The third dataset we use is the Bloomberg Twitter-Image Relationship dataset [37]. This dataset also consists of multimodal tweets, but the task for this dataset is a binary classification task, where the tweet-image pair must be classified into two classes: TEXT REPRESENTED, indicating that one or more entities in the tweet appear in the image, and TEXT NOT REPRESENTED, indicating that no entities from the tweet appear in the image. To train the caption transformer, we use the MS-COCO [14] dataset, but as our goal is not image captioning, we do not conduct an experimental evaluation for image captioning. All hyperparameters are described in Table 6. Most hyperparameters are directly borrowed from DETR [4, 35] or TomBERT [42]. Our method introduces no new hyperparameters. We report the average of 5 independent training runs for all our models, and report standard deviations wherever we claim SOTA.

### 4.2 Multimodal Target Oriented Sentiment Analysis Performance (RQ1)

We compare three configurations of our approach: (1) **LF-CapTrBERT**, which performs late fusion by skipping the construction of the auxiliary sentence and encoding the caption and tweet separately. (2) **EF-CapTrBERT**, which is the model described in Fig. 3 and Section 3.4. (3) **EF-CapTrBERT-DE**, where we initialize the BERT layers with domain-specific weights from [25] but leave the caption transformer unchanged to test if the model can overcome domain mismatch. For text only models, we compare to the baseline BERT model, as well as BERT+BL, which is BERT with another BERT layer stacked on it, and MGAN [7], which uses a multi-grain attention network for aspect understanding. We also compare with BERT-Pair-QA[33], which uses the auxiliary question method to obtain SOTA on SemEval 2014 Task 4. For multimodal comparisons, we compare to TomBERT [42], which is the current state of the art for target-oriented multimodal sentiment classification, as well as mPBERT [42]. We also compare with a strong baseline: Res-BERT+BL, which directly applies crossmodal attention to ResNet input features and the language features without any extra modifications.

### Table 2: Dataset Stats

| Split   | Negative | Neutral | Positive | Total | # Targets | Length |
|---------|----------|---------|----------|-------|-----------|--------|
| Train   | 368      | 1883    | 928      | 3179  | 1.3       | 16.7   |
| Validation | 149  | 679     | 303      | 1122  | 1.3       | 16.7   |
| Test    | 113      | 607     | 317      | 1037  | 1.3       | 17     |

| Split   | Negative | Neutral | Positive | Total | # Targets | Length |
|---------|----------|---------|----------|-------|-----------|--------|
| Train   | 416      | 1638    | 1508     | 3562  | 1.4       | 16.2   |
| Validation | 144  | 517     | 515      | 1176  | 1.4       | 16.4   |
| Test    | 168      | 573     | 493      | 1234  | 1.4       | 16.4   |

### Table 3: Bloomberg Dataset Stats

| Split   | Text Represented | Text Unrepresented | Total | Length |
|---------|------------------|-------------------|-------|--------|
| Train   | 927              | 1324              | 2251  | 11.7   |
| Validation | 232  | 332     | 564    | 11.7  |
| Test    | 290              | 414               | 704   | 11.4   |

4 EXPERIMENTS

We carry out experiments to answer the following research questions:

- **RQ1**: How does the amount of multimodal information affect accuracy of predictions in our model?
- **RQ2**: What are the failure modes of powerful state of the art unimodal models on short-text social media data?
Table 4: Multimodal Aspect Sentiment Classification. ± are standard deviations across five runs.

| Model           | TWITTER-15 | TWITTER-17 |
|-----------------|-------------|-------------|
|                 | Accuracy    | Mac-F1      | Accuracy    | Mac-F1      |
|                 |             |             |             |             |
| Image Only      |             |             |             |             |
| RES-Target      | 59.88       | 46.48       | 58.59       | 53.98       |
|                 |             |             |             |             |
| Text Only       |             |             |             |             |
| MGAN            | 71.17       | 64.21       | 64.75       | 61.46       |
| BERT            | 74.25       | 70.04       | 68.88       | 66.12       |
| BERT+BL         | 74.25       | 70.04       | 68.88       | 66.12       |
| BERT-Pair-QA    | 74.35       | 67.7       | 63.12       | 59.66       |
|                 |             |             |             |             |
| Text and Images |             |             |             |             |
| Res-MGAN        | 71.65       | 63.88       | 66.37       | 63.04       |
| Res-BERT+BL     | 75.02       | 69.21       | 69.2        | 66.48       |
| mPBERT (CLS)    | 75.79       | 71.07       | 68.8        | 67.06       |
| TomBERT (FIRST) | 77.15       | 71.75       | 70.34       | 68.03       |
| Proposed Model Configurations |       |             |             |             |
| LF-CapTrBERT    | 76.89       | 72.14       | 68.83       | 66.54       |
| EF-CapTrBERT    | **78.01**   | **73.25**   | 69.77       | 68.42       |
|                 | ±0.34       | ±0.36       | ±0.16       | ±0.48       |
| EF-CapTrBERT-DE | **77.92**   | **73.9**    | **72.3**    | **70.2**    |
|                 | ±0.83       | ±0.82       | ±0.27       | ±0.15       |

Table 5: The proposed model’s performance improvement over the text-only approach on the Bloomberg Twitter-Image relationship dataset. Standard deviations across 5 runs are reported.

|           | Accuracy  | Weighted-F1 | Modalities |
|-----------|-----------|-------------|------------|
| Vempala ’19 [37] | -         | 0.58        | Text+Image |
| BERT [6]   | 0.611     | 0.602       | Text       |
|            | ±0.021    | ±0.014      |            |
| Proposed Method | **0.648 (+6%)** | **0.640 (+6.3%)** | Text + Image |
|            | ±0.011    | ±0.013      |            |

4.2.1 Results & Discussion. The results are reported in Table 4. Two configurations of our model, EF-CapTrBERT and EF-CapTrBERT-DE, outperform all text-only models by a substantial margin. Against multimodal models, both EF-configurations outperform the field. However, TomBERT is quite competitive, and only EF-CapTrBERT-DE substantially outperforms TomBERT. **The comparison between LF-CapTrBERT and the EF-configurations is an ablation.** LF-CapTrBERT does not use the auxiliary sentence mechanism and instead encodes the tweet and caption separately, followed by a concatenation. This configuration performs much worse than the EF-configurations which use the auxiliary sentence. Finally, we note that the stability of EF-CapTrBERT-DE is lower than that of EF-CapTrBERT, which we conjecture is due to the domain mismatch between the caption transformer and the language model in EF-CapTrBERT-DE.

4.3 Image-Relationship Understanding (RQ2)

In this experiment, we use the Bloomberg Twitter Image Relationship Dataset (Table 3) to answer whether the language model is able to learn a correspondence between the image description, the target, and the tweet. We compare against text-only BERT, and use EF-CapTrBERT as our experimental configuration. We report accuracy and weighted-F1 for consistency with [37].

4.3.1 Results & Discussion. The results can be seen in Table 5. The text-only BERT model only outperforms the memory-based model in [37] by a small amount. However, adding multimodal image context with EF-CapTrBERT results in a substantial performance improvement. This shows that EF-CapTrBERT is able to fuse image and text well enough to improve over a purely-text model to understand when it needs to seek additional context from the image.

4.4 Further Analysis RQ3 + RQ4

4.4.1 RQ3: Effect of image description length on prediction accuracy? We compute the length of each caption, which can be considered a weak proxy for the amount of information in a caption. Then, we bin the captions by length and count the number of correct
Figure 5: We investigate why state of the art aspect sentiment analysis models, specifically BERT-Pair-QA, experience degraded performance in the social media setting, and find that their confidences are badly calibrated in the social media setting.

and incorrect predictions for each bin. The results can be seen in Fig 4. Surprisingly, we find that there is no strong relationship between the descriptiveness of the image caption and performance. Intuitively, we might expect longer captions to be more informative, and lead to higher accuracy. However, in practice we find that the longer captions also include more noise and more spurious objects, thus counteracting their potential for higher information.

4.4.2 RQ4: Why does BERT-Pair-QA’s performance degrade badly on tweets? BERT-Pair-QA is a state of the art and highly effective method for unimodal aspect sentiment analysis. Surprisingly, it performs badly on Twitter data, even though it exhibits strong performance in other domains. BERT-Pair-QA utilizes an auxiliary sentence like our EF-CaTrBERT and EF-CaTrBERT-DE models. However, unlike our method, it converts the 3-class sentiment classification problem into a 3-way binary classification problem and uses the confidence levels of each binary classifier to make a choice. We examine the confidence levels of a BERT-Pair-QA model, specifically

\[
\text{Pool}_\text{OUT} = \text{softmax}(W^T \text{Dropout}(H^{[\text{CLS}]}))
\]

when \(H^{[\text{CLS}]}\) contains different label queries in the auxiliary sentence. We find that the model consistently overestimates probabilities, and this is the reason for performance degradation. The results of our analysis can be seen in Fig 5. One explanation for why this happens is because BERT-Pair-QA was developed for review texts, which tend to be longer than tweets and more informative, so confidence levels may be more trustworthy for reviews than tweets.

5 CONCLUSION

We identify a number of challenges (high variance in usefulness of the visual modality, scene complexity, and short texts) that set the new field of multimodal aspect/target sentiment analysis apart from multimodal sentiment analysis and aspect/target sentiment analysis. We develop an approach, EF-CaTrBERT that uses translation in the input space to translate images into text, followed by multimodal fusion using an auxiliary sentence input to the encoder of a language model. We justify the auxiliary sentence fusion by showing our LF-CaTrBERT configuration, which lacks the auxiliary sentence, cannot match the performance of the EF-CaTrBERT configuration. We show that our input translation significantly improves performance on the task of detecting if entities in tweets are represented in images. We achieve state of the art performance on target/aspect sentiment analysis on the Twitter-15/17 datasets without modifying the architecture of our pretrained BERT language model. We investigate and explain the performance degradation of the otherwise powerful BERT-Pair-QA model on tweets.

ACKNOWLEDGMENTS

We are grateful to Daniele Magazzeni and Robert Tillman at JP Morgan AI Research for many suggestions and thought-provoking conversations which helped shape this paper. This work was supported by a JP Morgan Faculty Research Award.

REFERENCES

[1] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, Melbourne, Australia, 2236–2246. https://doi.org/10.18653/v1/P18-1208
