MORSE: MultimOdal sentiment analysis for Real-life SEttings

Yiqun Yao
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, United States
yaoyq@umich.edu

Verónica Pérez-Rosas
Computer Science and Engineering
University of Michigan
Ann Arbor, MI, United States
vrncapr@umich.edu

Mohamed Abouelenien
Computer and Information Science
University of Michigan
Dearborn, MI, United States
zmohamed@umich.edu

Mihai Burzo
Mechanical Engineering
University of Michigan
Flint, MI, United States
mburzo@umich.edu

ABSTRACT

Multimodal sentiment analysis aims to detect and classify sentiment expressed in multimodal data. Research to date has focused on datasets with a large number of training samples, manual transcriptions, and nearly-balanced sentiment labels. However, data collection in real settings often leads to small datasets with noisy transcriptions and imbalanced label distributions, which are therefore significantly more challenging than in controlled settings. In this work, we introduce MORSE, a domain-specific dataset for MultimOdal sentiment analysis in Real-life SEttings. The dataset consists of 2,787 video clips extracted from 49 interviews with panelists in a product usage study, with each clip annotated for positive, negative, or neutral sentiment. The characteristics of MORSE include noisy transcriptions from raw videos, naturally imbalanced label distribution, and scarcity of minority labels. To address the challenging real-life settings in MORSE, we propose a novel two-step fine-tuning method for multimodal sentiment classification using transfer learning and the Transformer model architecture; our method starts with a pre-trained language model and one step of fine-tuning on the language modality, followed by the second step of joint fine-tuning that incorporates the visual and audio modalities. Experimental results show that while MORSE is challenging for various baseline models such as SVM and Transformer, our two-step fine-tuning method is able to capture the dataset characteristics and effectively address the challenges. Our method outperforms related work that uses both single and multiple modalities in the same transfer learning settings.

KEYWORDS

Multimodal Sentiment Analysis; Dataset; Transfer Learning; Transformer; Imbalanced learning

ICMI '20, October 25–29, 2020, Virtual event, Netherlands

ACM ISBN 978-1-4503-7581-8/20/10…$15.00
https://doi.org/10.1145/3382507.3418821

ACM Reference Format:
Yiqun Yao, Verónica Pérez-Rosas, Mohamed Abouelenien, and Mihai Burzo. 2020. MORSE: MultimOdal sentiment analysis for Real-life SEttings. In *Proceedings of the 2020 International Conference on Multimodal Interaction (ICMI '20), October 25–29, 2020, Virtual event, Netherlands.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3382507.3418821

1 INTRODUCTION

In this paper, we focus on multimodal sentiment analysis, which is defined as the task of identifying the sentiment orientation (usually labeled positive, negative, or neutral) in multimodal data.

Despite the great improvement brought by existing datasets [33, 53, 54], which are mainly collected in controlled manners, fewer studies have been conducted on multimodal sentiment analysis tasks in real-life settings. The difference between real-life settings and controlled settings are (1) real-life videos usually need automatic transcription due to the high cost of manual transcription and issue of real-time response; (2) real-life tasks usually have small data scale and imbalanced label distribution.

To benefit the study on the above issues, we introduce MORSE, a domain-specific dataset for MultimOdal sentiment analysis in Real-life SEttings, built up with interview videos from a consumer research study on skin and health care products. In the videos, the panelists are asked to talk about their usage of their everyday products, but they do not necessarily express their sentiment towards it explicitly. The construction of our dataset highlights the "natural label distribution" scenario where we do not purposely seek for a relatively balanced number of positive or negative samples but rather keep the underlying sentiment distribution in all utterances. The ratio of samples annotated as negative, positive, and neutral in our dataset is around 1:10:30; having nearly 75% of neutral labels thus exemplifies a naturally imbalanced distribution. Besides, with only 63 of 2787 video clips annotated as negative, it is also a direct example of specific domain applications where the training samples for certain minority labels are extremely scarce. Another characteristic of our dataset is that it contains only real-life videos recorded under noisy environment conditions, and the voice of the interviewer frequently mixes with the panelist, making it more difficult to have clean automatic transcription.

Since the task setting of MORSE is a combination of noisy text, imbalanced labels and insufficient samples for minority classes, a widely-considered solution can be transfer learning. There has

Table 1: Comparison to related datasets.

Dataset	Task	Domain	#samples	#classes	Transcription	Distribution
MOUD [33]	Sentiment	General product review	412	2	Automatic	Balanced
MOSI [53]	Sentiment	Movie review	2199	7	Manual	Balanced
MOSEI [54]	Sentiment/Emotion	Open domain	23453	7/6	Manual	Balanced
IEMOCAP [7]	Emotion	Scripted action	10037	10	Manual	Balanced*
MELD [36]	Emotion	TV-series	13708	6	Manual	Imbalanced
MORSE	Sentiment	Domain-specific consumer interview	2787	3	Automatic	Imbalanced

* Although IEMOCAP has 10 emotion labels, previous work usually used the majority of 4 or 6 and discard the minorities.

been a surge of transfer learning research using the "pre-training and domain-adaptation" paradigm with Transformer model [47], both on natural language processing [14, 37, 50] and multimodal processing [15, 23, 42, 44, 55]. The majority of existing multimodal pre-training methods rely on the visual-groundings of language. However, in sentiment analysis tasks, finding visual-groundings is not as effective [38]. This is because most entities mentioned in the text do not even exist in the video. Besides, for the visual and audio modality, the key to correct sentiment prediction lays in the sequential change of facial expressions, gestures and speech tone, which can be automatically aligned with the textual word sequences [35, 53, 54].

There has not been much exploration on a joint pre-training method for this kind of tasks. On this side, we propose a novel two-step fine-tuning method based on BERT pre-training [14]. Our method enables the incorporation of sequential visual and audio features to a pre-trained textual model. In the task setting of our dataset, it outperforms basic models as well as both single-step textual BERT and related multimodal pre-training methods for video sentiment analysis, establishing a temporary state-of-the-art score and a strong baseline for future research.

The contributions of our paper are 3-fold:

- We propose a domain-specific dataset for MultimOdal sentiment analysis in Real-life SEttings (MORSE). It's especially suitable for applications regarding real-life videos and imbalanced label distribution.
- We provide rich baseline results using both non-sequential and sequential features, illustrating the characteristics and challenges covered by the MORSE dataset.
- We propose a novel two-step fine-tuning method that leverages multimodal information in the transfer learning for sentiment analysis [1]. It outperforms both single-modal and multimodal related work in our task settings.

2 REAL-LIFE SETTINGS

We define *real-life settings* as having the following two issues that are typical for multimodal sentiment analysis tasks regarding real-life videos, and not well-covered by existing work:

Firstly, *automatic and noisy transcriptions*. In most work [25, 35, 52–54], the models are built on manual transcriptions. However, for real-life applications, the sentiment analysis models usually deal with raw videos directly taken from recording devices. To make use of the textual information (which is important in sentiment analysis) in real-life videos, automatic transcription is commonly applied because of the high cost of manual transcription and the requirement of real-time response. This produces noise in the textual modality and potentially hurts the model performance.

Secondly, *small data scale and imbalanced label distribution*. For most existing datasets, the collection process ensures that the speech clips necessarily contain opinions, and the number of samples under each label (namely "label distribution") is usually sufficient and balanced [33, 53, 54]. However, real-life applications based on daily talks may not necessarily contain abundant opinion segments because people do not express their sentiments frequently when they do not intend to. Therefore, the "natural" label distribution in some real-life tasks can be very different from existing datasets built with only opinion segments; they are more likely to be imbalanced with "neutral/objective" label being the majority. Furthermore, in applications corresponding to specific domains, because of the small scale of the available data, the number of training samples under minority labels can even be scarce and insufficient for learning. This brings difficulty to sentiment analysis because affection labels (positive or negative) are what people mainly care about, however, these represent the minority classes in the over-all distribution in real-life settings. As demonstrated in our experimental results, minority labels can easily be overwhelmed by majorities, which affects the generalization of the models.

3 RELATED WORK

3.1 Multimodal Sentiment Analysis

To date, the majority of existing work in multimodal sentiment analysis uses three modalities: textual, visual, and audio; the classification task is usually formulated as assigning one of three sentiment labels: positive, negative, or neutral, or conducting regression on sentiment intensities. Commonly-used benchmarks for multimodal sentiment analysis include MOUD [33], MOSI [53] and MOSEI [54]. There are also related datasets in multimodal emotion recognition task, such as IEMOCAP [7] and MELD [36]. These datasets were either acquired under controlled lab settings (i.e., IEMOCAP) or using specific data collection guidelines that ensure that they have a balanced label distribution, have reasonable audiovisual quality, and also include high-quality manual transcriptions. This makes them valuable to explore the properties of multimodal human behavior for these tasks. However, an important drawback is that they do not represent the natural label distribution of the task and also do not fully portray the challenges of multimodal processing.

[1] Our code and data is available at https://github.com/FlamingHorizon/MORSE.

We address these issues while building MORSE. In particular, our data collection and annotation process preserves the natural label distribution present in consumer videos and portrays the challenges of multimodal processing, including ambient noises and different recording conditions, and the use of noisy automatic transcriptions.

Regarding baseline methods, Support Vector Machine (SVM) [46] is often the first choice of preliminary baselines [33, 53, 54]. For Neural Network-based models using sequential features, Gated Recurrent Units (GRU) [5] and Long Short-term Memory (LSTM) [21] are widely considered. [18, 19] proposed models based on Memory Network [41] and attention [5] mechanism for emotion recognition. Other models include multimodal factorization [24, 45]. If trained on balanced labels and sufficient data, existing models had solid performance, but limited studies have been conducted on their behaviors on training data with naturally imbalanced labels.

3.2 Multimodal Pre-training

There has been a series of work on joint pre-training of language and visual modalities [15, 23, 42, 44, 55] for tasks such as Visual Question Answering [3], Video Captioning [49] and Visual Dialog [11, 12]. The typical paradigm they use is to first encode different sub-parts of the image or video as a visual feature sequence following the textual word embeddings, and then pre-train a sequential model on top of the complete sequence using methods inspired by BERT[14]. The rationale behind this is that the Transformer self-attention heads learned during the pre-training phase can align the word semantics to specific sub-parts of the visual contents. For VQA and Visual Captioning tasks, these methods are highly effective because the visual entities are the key to questions. However, in sentiment analysis, the main goal is not linked to visual grounding but rather to connect sentiment with the sequential change of facial/body language and speech tone.

A common approach in multimodal sentiment analysis is to time-align words, speech, and visual content [35, 53, 54]. For instance, [38] used textual BERT for pre-training and added shifting gates to inject visual and audio features into the multimodal representation. Our proposed model is closely related to this method, but conduct two steps of fine-tuning instead of one, because an independent fine-tuning on textual modality is critical for our real-life settings, as we show in our experimental results. We also avoid the complexity brought in by shifting gate parameters to better address the extremely imbalanced label distribution.

3.3 Imbalanced Learning

Typical methods to deal with imbalanced labels include SMOTE [10, 30] and RUSTBoost [40]: the core idea is either over-sampling the minority labels or under-sampling the majority labels to build a balanced pseudo-dataset [1]. The main issue with these methods is that they work best with the ensemble of a relatively large number of simple classifiers, limiting their use in combination with neural network models that can generate better data representations.

Transfer learning is also widely considered as a solution to imbalanced label learning [6, 32]. Recent research has shown that domain adaptation with pre-trained language models [14, 34] usually alleviates the imbalance and scarcity issue to some extent [2, 26]. This is because the knowledge obtained from the large external corpus is encoded in the model weights, so samples of minority labels achieve meaningful representations before the fine-tuning process starts, thus mimicking the "over-sampling" process.

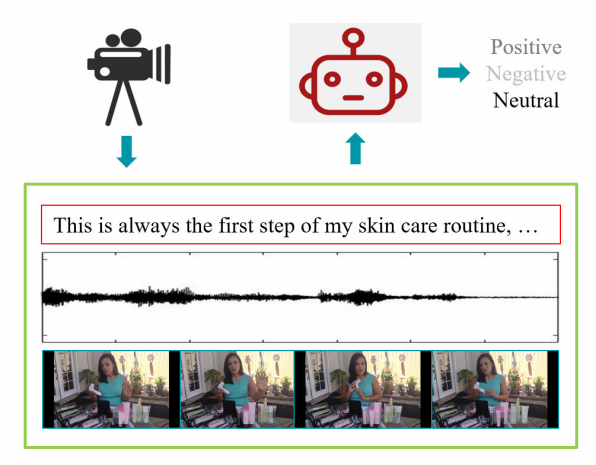

Figure 1: Demonstration of multimodal sentiment analysis task and typical video scenes of the MORSE dataset.

4 MORSE DATASET

4.1 Data Source

The source videos of MORSE are obtained from a consumer research study on skin and health care products conducted by Procter & Gamble during 2015. In these videos, panelists are interviewed about their skincare routine and product usage. During the interview, the panelist is asked to pick up the products they use for the daily skin care routine one-by-one, and talk about their product usage, including its function, frequency of use, effectiveness, and their experiences with the product. These conversations are conducted in a free manner and interviewers do not ask explicitly for users' likes or dislikes about the product, but rather focus on product usage and habits. This makes the interview similar to daily conversations.

Figure 1 shows a similar scenario to those in our videos [2]. The dataset is derived from individual recordings from 49 female panelists with ages ranging approximately from 20-40 years and located in two cities in the US. Each interview has an average duration of 40 minutes and portrays two speakers, the panelist and the interviewer. All participants expressed themselves in English.

4.2 Video Segmentation

As an initial step, we segment the videos to identify portions of the conversations where panelists talked about a specific product. Since the panelists are frequently interrupted by the interviewer to make clarifications or change topics, we devise a set of guidelines for video segmentation: 1) Select only the segments where the panelist is talking about a product; 2) Start segmenting when the speaker starts talking about the product (Not when the interviewer asks

[2]Due to privacy issues, we'll release the face features only instead of the actual videos. The pictures shown in figure 1 are from YouTube.

Table 2: Comparison between MORSE and MOSI [53].

Statistics	MOSI	MORSE
Total number of segments	2199	2787
Total number of videos	93	49
Total number of distinct speakers	89	49
Average segments in video	23.2	56.9
Average segment length	4.2 sec	19.4 sec
Average word count per segment	12	48
Total of unique words in segments	3107	5547
Total number of words in segments appearing at least 10 times in the dataset	557	834

the question); 3) Stop segmenting when the speaker is done talking about the product, i.e., changes the topic, is interrupted by the interviewer for more than 30 seconds or makes a long pause.

The motivation of these guidelines is to identify all the segments where panelists potentially express sentiment towards the product they are describing, although not exclusively, thus reflecting the *natural* distribution of sentiment expressed during the conversation.

All 49 interviews are segmented independently by three annotators using the ELAN annotation tool [48]. Before the segmentation phase, annotators participated in a calibration step where they discussed the criteria for segmenting each video, segmented two videos independently, compared the obtained segmentation, and discussed disagreements. To further verify that annotators were consistently following the segmentation guidelines, we measured the pair-wise overlap ratio for segmentation in a sample of 10 videos. This is done by pairing two segments from different annotators that are close to each other in timeline, and computing the Intersection over Union (IoU) score (ranging between 0 and 1), averaged over all pairs and all videos. The final overlap score is 0.3839, which suggests reasonable consistency among the three annotators.

We then proceeded to independently segment the remaining videos. Following this process, we obtained a total of 2,787 video segments with an average duration of 19.4 seconds.

4.3 Sentiment Annotations

After segmenting the videos, we move to an annotation step to assign a sentiment label that best summarizes the sentiment expressed by the panelist while talking about the product. For each video, we seek to assign one of three categories: positive, negative, or neutral. The annotation is also conducted using ELAN. As before, annotators started by double coding two conversations, then compared their annotations and resolved disagreements. This phase involved several iterations where annotators reconciled differences while assigning the labels. Then they proceed to annotate independently.

To measure inter-annotator consistency, we use the Cohen's Kappa score [28]. We reach a Kappa score of 0.6495 in 10 double-coded videos, indicating a satisfactory agreement. The final distribution of the annotated clips is 2,056 neutral, 668 positive, and 63 negatives. As observed, our dataset distribution is very skewed towards the neutral label. The ratio of labels is 32.6:10.6:1, reflecting the challenge of imbalanced labels in real-life data.

To further illustrate the challenges of our dataset, Table 2 shows a comparison of general data statistics with MOSI, a dataset widely used for multimodal sentiment analysis [53]. Note that in this table, we only compare with the "opinion segments" of MOSI as it contains a neutral label. As observed, our dataset has significantly longer video duration and word sequences, which in combination with the noisy transcriptions and imbalanced label distribution, makes our task much more challenging.

4.4 Transcription

To obtain the automatic transcription, we first extract the audio of the corresponding video and split it based on the segmentation. We obtain both the transcripts and word timestamps using the Google Speech-to-text API [3]. Manual inspection showed that the transcription is of reasonable quality; however, there is non-negligible noise introduced by speech recognition errors and interruptions between speakers. The transcripts consist of approximately 133,000 words, with each transcript ranging between 1-398 words and having an average of 48 words. Table 3 shows sample transcript excerpts of positive, negative, and neutral videos.

5 METHOD

We propose a novel two-step multimodal fine-tuning method based on the Transformer [47] architecture and BERT [14] pre-training. In this section, we first describe the textual, visual, and audio features used by our proposed models and baselines; then we describe the Transformer model used for the three-way sentiment classification of MORSE, followed by our proposed fine-tuning method.

5.1 Feature Extraction

We extract both sequential features and non-sequential features for each modality. Sequential features are used by Recurrent Neural Network (RNN) and Transformer baselines, while non-sequential features are single vectors used by basic models such as Support Vector Machines (SVM), Logistic Regression and Boosting.

During the feature extraction process, we follow the common practice of existing work [35, 38, 53, 54] of time-aligning visual and audio features using the timestamp for each transcription word.

5.1.1 Linguistic Features. We obtain two sets of linguistic features. The first consists of the tf-idf vector representations of the word distribution in each transcript. The minimum document frequency is set to five, resulting in a vocabulary of size 1387. The second is sequential features obtained by first tokenizing the transcript with the WordPiece tokenizer tool [51] and then adding a special token $[CLS]$ at the start.

5.1.2 Visual Features. We explore two sets of facial features: action units [17] (AUs) and face embeddings. We extract visual features at the word level by selecting the video frame at the word median duration. We start by reshaping the frame to a resolution of 640×360, and process it using the Face++ API [16] to recognize the speaker's face. The success rate of this process is 96.8%. [4] AUs are directly available from Face++, along with the recognized face box, we thus use the resulting 166-dimensional vector containing the x and y coordinates of 83 facial landmarks representing the

[3] https://cloud.google.com/speech-to-text

[4] For cases where face recognition fails, the corresponding visual features are obtained by smoothing from its adjacent neighbors.

Table 3: Examples of automatic transcription with positive, negative, and neutral labels.

Transcription	Label
I'm trying to think of the name of it melon. Melon Ashley. It did not work at all. And I guess I just decided I can't give somebody that kind of money and not get satisfaction and I sent her pictures afterwards and said	Negative
The reason I'm not too fond of this right now is because it doesn't have so screen looking so I actually have to give my new one with my sunscreen as you can tell I have any seats and stubborn.	Negative
I use Advanced Night Repair Estee Lauder. It's a like a recovery serum. It just helps with like age spots dark spots. If I had acne. It helps it. I can see it like healing faster and I don't know it might be a wrinkle fighter, but I'm not really sure it seems like it just helps.	Positive
Of the eye cream just because I don't want wrinkles around my eyes. So I want to prevent any damage any and all damage these have built in SPSS. So there are also some Block in a moisturizer.	Positive
I have Lancome me blush that I use	Neutral
And I spray this in my hair after I grade it and curl it just to keep Chris from developing.	Natural

different AUs. We further process these features by centering them to a (0,0) position and then standardize each feature to have zero mean and standard deviation of 1. To obtain face embeddings, we input the detected face box into FaceNet Inception [43], a Convolutional Neural Network (CNN) pre-trained on the VGGFace2 [9] face recognition dataset, and obtained 512-dimensional continuous features representing the speaker's face. We use two strategies to aggregate visual features at the video level. First, for non-sequential features, we take an average over the visual features (either AUs or face embeddings) corresponding to all words and use the resulting vector as a single representation of the visual modality. For the sequential features, we directly use the face embeddings obtained with FaceNet corresponding to each word in the transcript.

5.1.3 Audio Features. We use the Covarep [13] toolkit to extract raw audio features consisting of a 257-dimensional group delay spectrogram features obtained every 0.01 seconds. For sequential features, we average the audio features over the duration of each word. For non-sequential features, we average over the duration of all words in the transcript to obtain a single vector.

5.2 Basic Transformer Model

Our proposed method is based on a Transformer encoder [47] neural network that uses positional embedding to incorporate temporal information in the input sequences. In our work, the input of the Transformer encoder can be sequences of either word embeddings or joint-modal embeddings. Transformer uses multiple layers of self-attention operations to produce high-level representations at each position. The top-level representation, which captures long-distance connections inside the input sequence, is passed to a Multi-Layer Perceptron (MLP) classifier to predict the sentiment labels.

5.3 Two-step Fine-tuning

We propose a novel two-step fine-tuning method for multimodal sentiment analysis. The method uses two Transformer encoders: the first encodes the textual modality (TransEnc1) and the second (TransEnc2) encodes all modalities. Our pre-training scheme aims to address two main challenges in MORSE: Firstly, the noise introduced by automatic transcription, and secondly, the imbalanced label distribution. Figure 2 presents an overview of our method.

5.3.1 Pre-training. The inbalanced distribution of our dataset highlights the necessity of having large external corpus to learn the semantics of words and phrases expressing the sentiment. We use BERT [14], an unsupervised pre-training method for Transformer, to pre-train the language-only encoder (TransEnc1) using the BookCorpus and English Wikipedia datasets [14, 56]. BERT combines two different pre-training objectives: masked language model (MaskLM) and next sentence prediction (NSP).

5.3.2 Step 1: Language Fine-tuning. After the pre-training step, we obtain an open-domain encoder (TransEnc1). In order to make TransEnc1 more suitable for sentiment analysis, it must be further trained to encode sentiment. Therefore, in our step 1 of fine-tuning, we add an MLP classifier (namely Classifier1) to the TransEnc1 model, and fine-tune them using the Cross-Entropy loss function and our ground-truth sentiment labels. In this step, only textual transcripts are used as input as we seek to connect the learned language semantics to the sentiment labels. Thus, the process makes TransEnc1 capable of assigning distinctive representations for language inputs with different sentiment labels.

5.3.3 Step 2: Joint Fine-tuning. As mentioned in Section 1 and 3.2,most existing joint pre-training methods on visual-language tasks do not fit the problem setting of sentiment and emotion analysis where visual and audio features are naturally aligned with the language. To address this, we propose step 2: joint fine-tuning as a further step to incorporate aligned sequential features from visual and audio modalities. In step 2, we freeze the weights of TransEnc1, unplug it from Classifier1, and use it as a "language feature extractor". In particular, we use the top-level representations

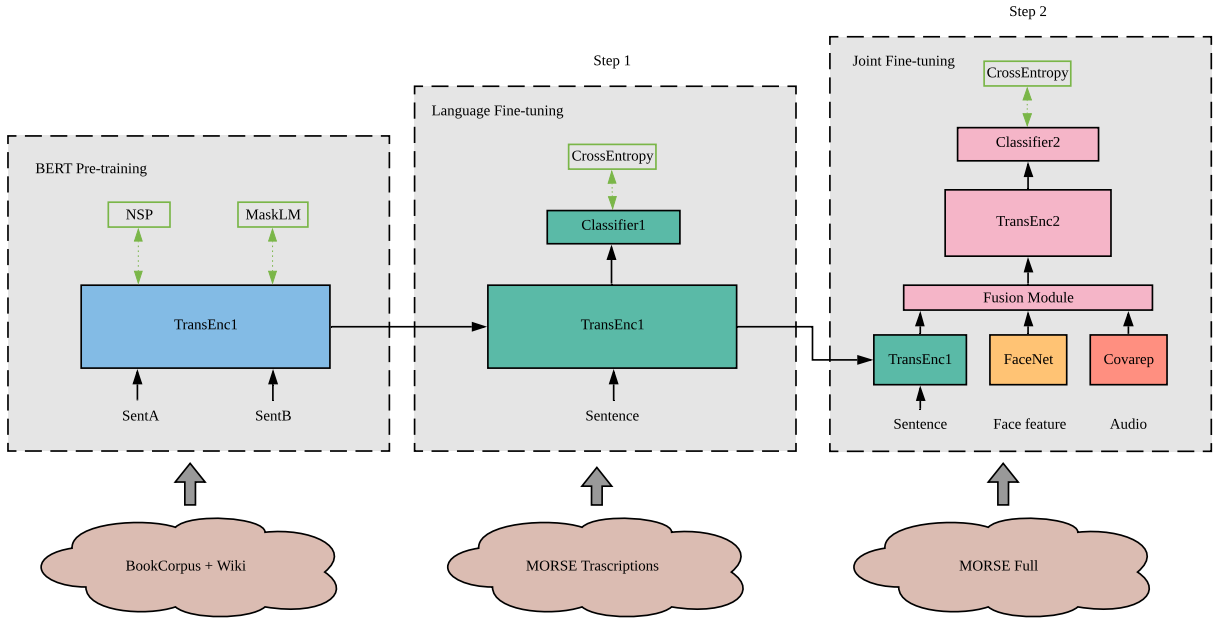

Figure 2: Our proposed two-step fine-tuning method for multimodal sentiment analysis.

of TransEnc1 (before the MLP classifier) as sequential linguistic features. We designed a fusion module to merge the three feature sequences from different modalities into one. Let L_i, V_i, A_i stand for the language, visual and audio feature corresponding to word token position i, respectively. The fusion module computes joint feature F_i via:

$$F_i = DP(LN(W_l * L_i + W_v * V_i + W_a * A_i + b)), \quad (1)$$

where W_l, W_v, W_a are projection weights and b is a bias vector; DP stands for a Dropout [20] layer with dropout rate 0.1; LN stands for Layer Normalization [4].

We learn to integrate multimodal representations using again a transformer model (TransEnc2). The F_i sequence is passed to TransEnc2 as input embeddings. A separate MLP network, namely Classifier2, is used in combination with TransEnc2 to predict the sentiment labels. In step 2, the weights of Classifier2, TransEnc2, and fusion module are randomly initialized and trained using Cross-Entropy loss. At test time, we run a forward pass with the modules in step 2 (Figure 2, right) and choose the output label with maximum softmax value from Classifier2.

It's worth noticing that we have two separate classifiers (Classifier1 and Classifier2) and conduct fine-tuning with Cross-Entropy loss in both Step 1 and Step2, which is different from related work that performs language fine-tuning using the same NSP and MaskLM objectives as BERT on domain-specific data [39]. Our motivation for choosing Cross-Entropy loss is that fine-tuning with NSP and MaskLM objectives relies on high-quality textual ground truth which is not our case as we use noisy transcripts obtained via automatic speech recognition.

6 EXPERIMENTS

6.1 Experimental Setup

Since MORSE has an extremely imbalanced label distribution in which the sentiment labels (positive and negative) are minorities, measuring only the overall accuracy would not be informative, i.e., a majority baseline predicting all "neutral" can have high overall accuracy but zero recall on minorities. Thus, our models are evaluated using precision, recall, and F-score metrics for each label. There are only 63 negative samples in MORSE. Thus, a single training-validation-testing split would produce results with high bias. To address this issue, we run 5-fold cross-validation and report the average results in our experiments.

We use two types of baselines. The first includes basic machine learning models that use non-sequential features, while the second consists of deep neural network models that use time-aligned sequential features.

6.2 Non-sequential Baselines

We experiment with four baseline methods that use non-sequential features for all the three modalities i.e., tf-idf vectors for the linguistic modalities, action units and face embeddings for the visual modality, and audio features obtained as described in section 5.1.

Support Vector Machines (SVM): SVMs are strong classifiers for small-sized datasets and at times outperform neural counterparts [8]. We use the SVM implementation from scikit-learn [31] with its default configurations.

Multi-layer Perceptron (MLP): The MLP feed-forward network has two hidden layers of 256 neurons and ReLU [29] activation. We

Table 4: Cross-validation results for basic models using non-sequential features. p, r, f stands for precision, recall and f-1 score, respectively. All visual stands for Action units + Face embeddings

Features	p-neg	r-neg	f-neg	p-pos	r-pos	f-pos	p-neu	r-neu	f-neu
SVM									
Ling. + Audio + Action units	0.070	0.127	0.089	0.359	0.500	0.417	0.808	0.687	0.742
Ling. + Audio + Face embeddings	0.060	0.113	0.077	0.373	0.489	0.423	0.812	0.706	0.755
Ling. + Audio + All visual	0.080	0.190	0.112	0.397	0.539	0.457	0.825	0.695	0.754
MLP									
Ling. + Audio + Action units	0	0	0	0.504	0.398	0.444	0.804	**0.881**	0.841
Ling. + Audio + Face embeddings	**0.200**	0.031	0.053	0.507	0.421	0.459	0.806	0.872	0.838
Ling. + Audio + All visual	0	0	0	0.493	0.403	0.442	0.803	0.873	0.836
SVMSMOTE+SVM									
Ling. + Audio + Face Embeddings	**0.200**	0.053	0.083	**0.524**	0.550	**0.537**	0.840	0.846	**0.843**
RUSBoost+DT									
Ling. + Audio + Face Embeddings	0.060	0.032	0.040	0.273	0.437	0.334	0.754	0.616	0.676
RUSBoost+Logistic									
Ling. + Audio + Face Embeddings	0.125	**0.318**	**0.179**	0.423	**0.631**	0.508	**0.873**	0.705	0.784

use an Adam [22] optimizer with an initial learning rate of 1e-3 to train for a maximum of 200 epochs.

SVMSMOTE: SVMSMOTE [30] addresses the imbalanced distribution by over-sampling the two minority classes with synthetic samples.

RUSBoost: RUSBoost [40] is an ensemble method that balances the classes by under-sampling the majority class during each training iteration. We use 1-depth Decision Tree (DT) and Logistic Regression as the weak classifiers for the ensemble, with the number of classifiers set to 1000.

Classification results using the non-sequential baseline models along with linguistic and audio features in combination with the two different sets of visual features (action units and face embeddings) are provided in Table 4. Note that for imbalanced learning methods we report results with face embeddings only as it worked better than in combination with action units and action units alone.

There are two main observations from the non-sequential model results: First, all models tend to neglect the negative label regardless of the features used, except when using imbalanced learning methods. This shows that in a small dataset with imbalanced distribution, it is necessary to use methods that can deal with imbalanced labels. Second, although SVMSMOTE over-samples both positive and negative classes, it has difficulties synthesizing new negative samples due to scarcity of existing samples and noisy transcriptions, but it works best for the positive label with relatively high sample numbers. RUSBoost under-samples the neutral label and hurts its performance, while benefiting the two minority classes.

6.3 Sequential Baselines

Most advanced models for sentiment analysis in videos use aligned feature sequences based on the temporal order. To compare sequential baselines models built with different modalities and architectures, we experiment with the following methods:

GRU-all: A 2-layer Recurrent Neural Network with GRU [5] cells. The dimension of the hidden layers is 256. The same fusion module as described in equation 1 is used to jointly project the features from different modalities. For language representation, we use a 200-dimensional word embedding which is updated as parameters during the training procedure. GRU-all is trained with Adam optimizer with learning rate 1e-3 and batch size 32.

Transformer: We use a transformer model with 12 layers, 12 attention heads and 768-dimensional hidden states. We experiment on single-modal Transformer with textual word embedding, action units, face embeddings, or audio features as input. We also experiment with a multimodal Transformer that jointly uses the three modalities with the fusion module as described in equation 1, namely *Transformer-all*. These models are trained on our MORSE dataset from scratch without pre-training, using the AdamW [27] optimizer for 200 epochs with initial learning rate 2e-5 and batch size 32. During training, the gradients are clipped to have a maximum norm of 1.0.

For all the sequential models, we limit the maximum sequence length to be 128. Table 5 shows the results of sequential baselines. We observe that Transformer using linguistic features performs better than other Transformers using single-modal visual and audio features, and comparable to the multimodal method Transformer-all. This indicates that for our task, the language features are more critical than visual and audio features. We also observe that Transformer-all outperforms GRU-all with the same features from all the 3 modalities. For completeness, we also trained Transformer-all on a manually-balanced dataset with 50 training samples for each class. The test F-score for negative, positive and neutral labels are (0.066, 0.375, 0.519), compared to (0.133, 0.379, 0.786) using the original imbalanced training data. Thus, supporting the use the imbalanced dataset for all of our experiments.

6.4 Two-step Fine-tuning for Multimodal Sentiment Analysis

We implement and apply our proposed two-step fine-tuning method on MORSE dataset. After the first step (language fine-tuning), we freeze the fine-tuned TransEnc1 and use it to extract language representations. In the second step, we fine-tune with multimodal features using the multimodal Transformer (TransEnc2), as described in Section 5.3. The TransEnc2 has the same structure and

Table 5: Results of sequential models using aligned features. p, r, f stands for precision, recall and f-1 score, respectively.

Model	p-neg	r-neg	f-neg	p-pos	r-pos	f-pos	p-neu	r-neu	f-neu
GRU-all									
Ling.+ Audio + Face emb.	0	0	0	0.323	0.321	0.322	0.764	0.781	0.773
Transformer									
Linguistic	0.073	**0.385**	0.123	0.402	0.552	0.465	0.836	0.623	0.714
Action units	0.333	0.077	0.125	0.238	0.149	0.183	0.737	0.844	0.786
Face embeddings	0.0	0.0	0	0.217	0.246	0.21	0.729	0.715	0.722
Audio	0.500	0.077	0.133	0.331	0.313	0.322	0.759	0.793	0.776
Transformer-all	0.500	0.077	0.133	0.371	0.388	0.379	0.781	0.791	0.786
Pre-training + Fine-tuning									
BERT-Linguistic	0.608	0.365	0.441	0.615	**0.557**	0.581	**0.850**	0.883	0.866
Shifting Gate [38]	0.333	0.154	0.211	0.378	0.485	0.425	0.800	0.740	0.769
Joint-two-step (ours)	**0.617**	0.365	**0.444**	**0.631**	0.548	**0.583**	0.848	**0.891**	**0.868**

hyper-parameters as TransEnc1, but we only run step-2 fine-tuning for 5 epochs because the model converges reasonably fast. In order to measure the effectiveness and efficiency of our proposed two-step fine-tuning method, we compare against the following two fine-tuning methods that are closely related to ours:

BERT-Linguistic: We first use BERT to pre-train the textual Transformer TransEnc1 using an external linguistic corpus as described in Section 5.3.1, and then perform our step-1: language fine-tuning using the transcriptions and sentiment labels of MORSE dataset. The fine-tuning takes 200 epochs; the learning rate and batch size are kept the same as all the Transformer baselines. In our two-step fine-tuning literature, this obtains the fine-tuned model after step-1 and before step-2. We name it BERT-Linguistic.

Shifting Gate: This is our implementation of the multimodal pre-training method proposed by [38], in which a trainable shifting gate is used to select among different modalities of inputs and inject them into the BERT-Linguistic fine-tuning process. The Transformer structure, hyper-parameters, and train settings are kept the same as BERT-Linguistic.

The last three rows in table 5 shows the results of the fine-tuning methods on MORSE. We point out the following observations:

- BERT-Linguistic significantly outperforms all the Transformers without pre-training, which shows that transfer learning helps solve the issues of noisy transcriptions and imbalanced labels to some extent.
- On our MORSE dataset with scarce training samples for minority labels, the shifting gate mechanism does not seem as effective as fine-tuning the language modality alone (BERT-Linguistic). This is potentially because the nature of the sentiment analysis task makes the language modality more critical than others, but the shifting gate itself brings complexity and needs sufficient data to learn which modality to choose, otherwise it has a negative impact on the language fine-tuning process. However, the performance is still significantly better than Transformers without pre-training.
- Our proposed two-step fine-tuning method outperforms the related Shifting Gate method [38] and slightly outperforms BERT-Linguistic. This improvement comes from our two-step paradigm: by first fine-tuning with BERT-Linguistic and freezing it, the language representation learned by language fine-tuning is better preserved than directly using a shifting gate to intervene in the fine-tuning procedure. In step 2, since the language representations are already distinguishable enough between different classes, the joint encoder learns to do minor corrections using the information from visual and audio modality. This cascaded manner fits the task intuition that language is more important than other modalities, resulting in fast convergence and improved performance. The performance gain is steady because we observed that Joint-two-step is better than BERT-Linguistic in all folds of the cross-validation.

7 CONCLUSION

In this work, we introduced a domain-specific dataset for MultimOdal sentiment analysis in Real-life SEttings (MORSE). It covers the properties of noisy transcriptions, imbalanced label distributions, and scarcity of minority labels, benefiting future research on related issues. We provided both sequential and non-sequential baseline methods using either single-modal or multimodal features. The performances of these baselines illustrate the challenge of real-life settings in our task and the necessity of transfer learning. Based on the Transformer architecture and BERT pre-training, we proposed a novel two-step fine-tuning method that first adapts language representations and then incorporates visual and audio features for multimodal fine-tuning. Experimental results show that our method captures the characteristics of MORSE dataset well, outperforming strong baseline models with advanced structures, as well as fine-tuning strategies proposed by related work.

ACKNOWLEDGMENTS

We are grateful to Matthew Barker and Rada Mihalcea for their insightful comments and helpful discussions. This material is based upon work supported by the Procter & Gamble Company through the Advanced Machine Learning Collaborative program at the University of Michigan (grant #007905) and by the National Science Foundation (grant #1815291). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the Procter & Gamble company or the National Science Foundation.

REFERENCES

[1] M. Abouelenien and X. Yuan. 2012. SampleBoost: Improving boosting performance by destabilizing weak learners based on weighted error analysis. In *Proceedings of the 21st International Conference on Pattern Recognition (ICPR2012)*. 585–588.

[2] Mohammed Al-Hashedi, Lay-Ki Soon, and Hui-Ngo Goh. 2019. Cyberbullying Detection Using Deep Learning and Word Embeddings: An Empirical Study. In *Proceedings of the 2019 2nd International Conference on Computational Intelligence and Intelligent Systems*. 17–21.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*. 2425–2433.

[4] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450* (2016).

[5] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[6] Siddhartha Banerjee, Cem Akkaya, Francisco Perez-Sorrosal, and Kostas Tsioutsiouliklis. 2019. Hierarchical Transfer Learning for Multi-label Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6295–6300.

[7] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation* 42, 4 (2008), 335.

[8] Evgeny Byvatov, Uli Fechner, Jens Sadowski, and Gisbert Schneider. 2003. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *Journal of chemical information and computer sciences* 43, 6 (2003), 1882–1889.

[9] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. 2018. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*. IEEE, 67–74.

[10] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[11] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2.

[12] Abhishek Das, Satwik Kottur, José MF Moura, Stefan Lee, and Dhruv Batra. 2017. Learning cooperative visual dialog agents with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*.

[13] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. 2014. COVAREP—A collaborative voice analysis repository for speech technologies. In *2014 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 960–964.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[15] Hamed Firooz Davide Testuggine Douwe Kiela, Suvrat Bhooshan. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. *arXiv preprint arXiv:1909.02950* (2019).

[16] Faceplusplus. [n.d.]. *Face Detection API.* https://www.faceplusplus.com/face-detection/

[17] E Friesen and Paul Ekman. 1978. Facial action coding system: a technique for the measurement of facial movement. *Palo Alto* 3 (1978).

[18] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2594–2604.

[19] Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 2122–2132.

[20] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. 2012. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580* (2012).

[21] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[22] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[23] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[24] Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019. Strong and Simple Baselines for Multimodal Utterance Embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. 2599–2609.

[25] Paul Pu Liang, Ziyin Liu, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Multimodal Language Analysis with Recurrent Multistage Fusion. In *EMNLP 2018: 2018 Conference on Empirical Methods in Natural Language Processing*. 150–161.

[26] Ping Liu, Wen Li, and Liang Zou. 2019. NULI at SemEval-2019 Task 6: transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*. 87–91.

[27] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101* (2017).

[28] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica: Biochemia medica* 22, 3 (2012), 276–282.

[29] Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*. 807–814.

[30] Hien M Nguyen, Eric W Cooper, and Katsuari Kamei. 2011. Borderline oversampling for imbalanced data classification. *International Journal of Knowledge Engineering and Soft Data Paradigms* 3, 1 (2011), 4–21.

[31] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[32] Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. *arXiv preprint arXiv:1906.05474* (2019).

[33] Verónica Pérez-Rosas, Rada Mihalcea, and Louis-Philippe Morency. 2013. Utterance-level multimodal sentiment analysis. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 973–982.

[34] Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of NAACL-HLT*. 2227–2237.

[35] Hai Pham, Thomas Manzini, Paul Pu Liang, and Barnabas Poczos. 2018. Seq2seq2sentiment: Multimodal sequence to sequence models for sentiment analysis. *arXiv preprint arXiv:1807.03915* (2018).

[36] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 527–536.

[37] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.

[38] Wasifur Rahman, Md Kamrul Hasan, Amir Zadeh, Louis-Philippe Morency, and Mohammed Ehsan Hoque. 2019. M-BERT: Injecting Multimodal Information in the BERT Structure. *arXiv preprint arXiv:1908.05787* (2019).

[39] Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2019. Adapt or get left behind: Domain adaptation through bert language model finetuning for aspect-target sentiment classification. *arXiv preprint arXiv:1908.11860* (2019).

[40] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. 2009. RUSBoost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans* 40, 1 (2009), 185–197.

[41] Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*. 2440–2448.

[42] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE International Conference on Computer Vision*. 7464–7473.

[43] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. 2017. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-first AAAI conference on artificial intelligence*.

[44] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5103–5114.

[45] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176* (2018).

[46] Vladimir Vapnik. 1998. The support vector method of function estimation. In *Nonlinear Modeling*. Springer, 55–85.

[47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*. 5998–6008.

[48] Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *5th International Conference on Language Resources and Evaluation (LREC 2006)*. 1556–1559.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*. 2048–2057.

[50] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in neural information processing systems*. 5754–5764.

[51] Zhifeng Chen Quoc V. Le Mohammad Norouzi Wolfgang Macherey Maxim Krikun Yuan Cao Qin Gao Klaus Macherey Jeff Klingner Apurva Shah Melvin Johnson Xiaobing Liu Łukasz Kaiser Stephan Gouws Yoshikiyo Kato Taku Kudo Hideto Kazawa Keith Stevens George Kurian Nishant Patil Wei Wang Cliff Young Jason Smith Jason Riesa Alex Rudnick Oriol Vinyals Greg Corrado Macduff Hughes Jeffrey Dean Yonghui Wu, Mike Schuster. 2019. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144* (2019).

[52] Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louisphilippe Morency. 2018. Memory Fusion Network for Multi-view Sequential Learning. In *AAAI-18 AAAI Conference on Artificial Intelligence*. 5634–5641.

[53] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[54] 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)*. 2236–2246.

[55] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. 2019. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059* (2019).

[56] Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*. 19–27.