Multimodal Emotion Fusion: Integrating Modalities for Enhanced Emotion Detection

AI6103 - DEEP LEARNING & APPLICATIONS - Project

Chithra Ramesh Asswin[†], Bendale Aneesh Santosh[†], Maheswaran Rohin Kumar[†], Shwetha Ravi[†], Asvini Selvaraj[†]

[†] School of Computer Science Engineering
Nanyang Technological University
Singapore

Abstract

Our study delves into the realm of emotion recognition, employing a multi-modal approach with the MELD dataset. Our objective is to create a more robust and sophisticated emotion identification system by utilizing the synergies and complementary information that different modalities offer in a seamless manner. We introduce a fusion strategy meticulously engineered to harmonize outputs from audio, text, and video models. Capitalizing on the pretrained capabilities of wav2vec for audio, CLIP-VIT for video, and the eminent BERT and RoBERTa for text, our study orchestrates a symphony of modalities. Significantly, our text model undergoes fine-tuning, elevating its predictive acumen. To achieve this, we introduce a fusion strategy that amalgamates the outputs from audio, text, and video models. This involves computing a late fusion weighted average of individual modalities, culminating in a comprehensive final prediction for the emotional state.

INTRODUCTION

Multimodal emotional analysis is a burgeoning field that combines various modalities, such as speech, facial expressions, and context, to provide a comprehensive understanding of human emotions. This approach is particularly beneficial in real-world scenarios where emotions are expressed through multiple channels simultaneously. It also aids in the development of personalized and adaptable systems, ensuring fair use of emotion recognition technologies

In recent years, multimodal emotional analysis has become more important in studies because it helps researchers understand human emotions more precisely and in more context. This way of thinking about analysis is very important in many areas, including human-computer interaction (HCI), healthcare, business, education, and more.

Combining different types of signals, like body language, facial emotions, and voice intonations, not only clears things up and makes them more robust, but it also makes sure that emotional states are shown more accurately. This is especially important in natural conversation situations where people show their feelings in more than one way at the same time.

Additionally, using multimodal emotional analysis helps create personalized and adaptable systems, which leads to technologies that change based on how people are feeling. A multimodal method is important for more than just its technical benefits. It is also important for ethical reasons, like reducing bias and making sure that emotion recognition technologies are used fairly. Basically, adding multimodal emotional analysis not only makes emotion detection systems more accurate, but it also makes sense from an ethical point of view and has real-world uses. This makes it an important new area of study and technology development right now.

RELATED WORK

Multimodal emotion recognition has undergone significant evolution, integrating text, audio, and video modalities. Early contributions by Busso et al. (2004) [1] explored audio-visual emotion recognition, paving the way for advancements. In 2017, Poria et al (2017) [2] conducted a comprehensive review, setting the stage for the integration of deep learning and notably enhancing emotion detection accuracy.

The introduction of datasets like MELD [3] facilitated sophisticated analyses, combining text, audio, and visual cues through fusion techniques. Fusion mechanisms, including feature-level and decision-level fusion, were explored by Morency et al. (2011) [4] and Baltrušaitis et al. (2018) [5], showcasing improved understanding of emotional context. Deep learning models have prominently led this integration, continually pushing boundaries and achieving state-of-the-art performance on benchmarks like MELD (Bagher Zadeh et al., 2018) [6].

Recent studies have delved into innovative fusion approaches [7]. demonstrated refined fusion mechanisms, showcasing the potential of modality combination in achieving superior performance. Furthermore, the significance of multi-task learning in multimodal emotion recognition has been highlighted by [8]. This approach leverages shared information and complementary features across tasks, contributing to a holistic understanding of emotions.

DATASET AND METHODS

MELD, an extension of the EmotionLines dataset, enriches emotion analysis by incorporating audio and visual modal-

	Train	Validation	Test
No of Samples	9989	1112	4807

Table 1: MELD data split

ities alongside text. Derived from the Friends TV series, MELD comprises around 16000 samples split as shown in **Table 1**, introducing an extra layer of complexity with multiple speakers. Each statement in the dataset is labeled with one of seven discrete emotions: Anger, Disgust, Sadness, Joy, Neutral, Surprise, or Fear. This diverse corpus serves as a valuable resource for studying emotions in a multimodal context.

Text Data Preprocessing

For text data extracted from MELD dialogues, the preprocessing pipeline begins by parsing file paths of associated MP4 files to extract dialogue and utterance IDs. Through meticulous dissection of file paths, numeric components following predefined patterns are isolated. These identifiers then filter a CSV file containing comprehensive dialogue and utterance information. The final step involves extracting specific utterance text based on identified dialogue and utterance IDs, resulting in a streamlined and organized dataset.

Audio Data Preprocessing

The raw audio data undergoes a multi-step transformation process facilitated by the Wav2Vec model during preprocessing. Custom emotion labels are defined to tailor the classification task. The audio is re-sampled to 16 kHz and features such as MFCC for specific audio file windows are extracted before being fed into the Wav2Vec model.

Video Data Preprocessing

The video clips within the MELD dataset exhibit varying lengths and resolutions. To preprocess these clips, the following steps are employed:

- Every 10th frame of the video is selected for emotion detection, reducing computational load while avoiding redundancy inherent in consecutive frames.
- Each image frame is resized to 224x224 px (ViT input size) and normalized.
- Extracted frames are converted into a NumPy array, allowing for emotion detection on a reduced yet representative subset of frames from each video clip.

EXPERIMENTAL METHODOLOGY

A. Text Emotion Recoginition

Emotion recognition in text is a critical task in natural language processing (NLP), enabling machines to understand and interpret human emotions. This section discusses the implementation of emotion recognition using two pretrained models, **BERT** (Bidirectional Encoder Representations from Transformers) [9] and **RoBERTa** (Robustly optimized BERT approach) [10] used, on the MELD (Multimodal EmotionLines Dataset).

BERT: BERT's core mechanism is a multi-layer bidirectional Transformer encoder, a type of attention mechanism that learns contextual relations between words in a text. Unlike prior language representation models, BERT is designed to pre-train deep bidirectional representations by joint conditioning on both the left and right context in all layers. This is achieved through two novel pre-training tasks: (a) Masked Language model (MLM), (b) Next Sentence Prediction (NSP).

Pretrained BERT Base Uncased in Hugging Face The **'bert-base-uncased'** model is a transformer model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pre-trained on the raw texts only, with no humans labeling them in any way with an automatic process to generate inputs and labels from those texts. The model has 12-layer, 768-hidden, 12-heads, and 110M parameters.

RoBERTa: RoBERTa refines the pioneering BERT's approach, leveraging more extensive training data and iterations, dynamic masking that updates the masked tokens throughout pre-training, and a byte-level BPE tokenizer for finer-grained word representations. It employs larger batch sizes and learning rates, enhancing convergence efficiency. Furthermore, RoBERTa discards the Next Sentence Prediction (NSP) task, deemed ineffective. These enhancements allow RoBERTa to outperform BERT, which itself revolutionized NLP by pre-training on vast text data using the Transformer architecture before task-specific fine-tuning, laying the groundwork for many advanced models and applications in the field.

Pretrained RoBERTa Base in Hugging Face Hugging Face provides a pre-trained model called 'roberta-base'. This model is trained on the same data and with the same hyperparameters as the original RoBERTa model described in the paper. It's called 'base' because it has the same model size as BERT base, with 12 layers, 768 hidden units, and 12 attention heads, totaling 125M parameters.

The 'roberta-base' model is a transformer model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pre-trained on the raw texts only, with no humans labeling them in any way with an automatic process to generate inputs and labels from those texts.

Methodology used for BERT and RoBERTa

Pre-Processing Tokenization was performed using BERT and RoBERTa's respective tokenizers. Each utterance was converted into a sequence of token IDs, with padding and truncation applied to maintain uniform sequence length.

Model Fine-Tuning Both BERT and RoBERTa were fine-tuned on the MELD dataset for emotion classification. The fine-tuning involved adapting the pre-trained models to the specific task of classifying text into emotion categories.

Hyperparameters Fine-tuning involved tweaking various hyperparameters:

- Learning Rate: Lower learning rates (e.g., 2e-5 for BERT, 1e-5 for RoBERTa) were chosen to make small adjustments to the model weights.
- Batch Size: Adjusted to manage computational load and training stability (8, 16, 32 for BERT and RoBERTa).
- Number of Epochs: Generally set to 3 to prevent overfitting while ensuring adequate learning.
- MAX_TOKEN length: 256, 512 for BERT and RoBERTa depending upon the batch size.

The corresponding experiments and results discussion are in the Results & Discussion Section.

B. Audio Emotion Recognition

The parsing of file paths from MP4 files within the MELD dataset is the first step in this process.Through meticulous dissection, unique identifiers—namely, dialogue and utterance IDs—are extracted from these paths, forming the foundational elements for subsequent data filtering. The extracted IDs are subsequently utilized to selectively access and filter a CSV file containing comprehensive dialogue and utterance information.Leveraging Wav2Vec, the model then conducts emotion classification on the extracted audio, generating predictions with associated confidence scores. The outcomes, encapsulating both true emotion labels and predicted emotions, are systematically organized into a structured DataFrame, providing a comprehensive and coherent record for subsequent analysis. This methodical approach ensures a systematic and reproducible integration of Wav2Vec for audio extraction and emotion classification within the context of the MELD dataset.

C. Video Emotion Recognition

In this study, we employ a zero-shot image classification model, specifically the `openai/clip-vit-large-patch14` model, to perform emotion recognition on video data. This model is a product of OpenAI and uses a Vision Transformer (ViT-L/14) architecture as an image encoder and a masked self-attention Transformer as a text encoder. These encoders are trained to maximize the similarity of (image, text) pairs via a contrastive loss, which allows the model to perform image classification without any fine-tuning or labeled data, by simply providing natural language descriptions of the classes.

The choice of this model is motivated by its ability to interpret images in the context of natural language descriptions, which makes it suitable for deciphering the complex visual elements associated with emotions. Furthermore, the integration of CLIP with ViT offers a powerful combination for interpreting visual data, leveraging the strengths of both models. ViT's effectiveness in image classification tasks and CLIP's proficiency in understanding images in a textual context provide a comprehensive approach to emotion recognition in video frames.

The experimental methodology involves the following steps:

1. Loading the model from the Hugging Face library.
2. Emotion Detection Function: We define a function `detect_emotion` that takes an image and the model as inputs, and returns the probabilities of each of the seven emotion labels for the image.
3. We load the test videos and extract every 10th frame for emotion detection. We then
4. Emotion Recognition: We then get the cumulative weighted scores of every frame to get the final prediction labels along with their scores.

RESULTS & DISCUSSION

A. Text Model Results

Statistical Significance: Tests were conducted to determine if the improvements or regressions in performance were statistically significant.

Observations

Based on the results observed from **Tables 2, 3, 4, and 5** the fine-tuned BERT and RoBERTa models were evaluated on emotion recognition tasks using the MELD dataset. Our analysis of the results considers several hyperparameters such as **maximum token length, batch size, learning rate (LR), and weight decay (WD)** as well as performance metrics like training and validation loss, accuracy, F1 score, precision, and recall.

BERT Model Inference:

- **Maximum Token Length:** The BERT model was fine-tuned with a maximum token length of 512, which was consistent across all experiments. One experiment attempted to use a token length of 256 but faced out-of-memory (OOM) issues, indicating resource constraints at lower token lengths.
- **Batch Size and Learning Rate**: A smaller batch size of 8 with a higher learning rate (1e-4) showed an increase in recall on the test set compared to a batch size of 16 with a lower learning rate (2e-5), indicating that the model may be better at identifying relevant instances of emotion at this configuration. However, this comes with a slightly higher validation loss, suggesting a potential overfitting risk.
- **Performance Over Epochs:** Model is been trained for 3 epochs(Since its already a pre-trained model) over three epochs, there is a trend showing that with a **lower learning rate (2e-5)**, the model achieves a higher weighted average accuracy and F1 score. This suggests that the model is becoming more effective at classifying emotions correctly, as indicated by the improvement in these metrics over time.

RoBERTa Model Inference:

- **Consistent Token Length and Epochs:** Similar to BERT, the RoBERTa model maintained a consistent token length and number of epochs across different experiments. The model achieved the highest recall on the test set with a token length of 512 and a batch size of 16, indicating a stronger ability to identify all relevant instances of emotions.

Batch Size	Learning Rate	Weight Decay	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
8	2e-4	0.01	0.7223	1.1372	0.6517	0.6355	0.6261	0.6517
16	1e-4	0.01	1.2752	1.1007	0.6411	0.6082	0.5937	0.6411
32	2e-5	0.01	1.1398	1.1378	0.6281	0.5978	0.5751	0.6281

Table 2: Training and Validation results based on different Hyperparameters for BERT

Weighted Accuracy	F1-Score
0.6325	0.6145
0.6344	0.5983
0.6475	**0.6204**

Table 3: Test results based on the different hyperparameters used in BERT

- **Learning Rate Effects:** RoBERTa appears more sensitive to changes in learning rate, with the higher learning rate (1e-5) yielding better recall but slightly lower precision, reflecting a trade-off between correctly identifying as many relevant cases as possible and the accuracy of the predictions made.
- **F1 Score and Weighted Average:** The F1 score, which balances precision and recall, is relatively stable across different configurations for RoBERTa. However, a batch size of 16 yields the best F1 score, suggesting it is the most balanced configuration for emotion recognition tasks with this model.

Comparative Inference:

- **Roberta vs. BERT:** RoBERTa tends to have a higher recall but slightly lower precision compared to BERT. This indicates that while RoBERTa may identify emotions more comprehensively, BERT could be more accurate in its predictions.
- **Optimal Hyperparameters:** Both models achieve better performance with a token length of 512. For BERT, a learning rate of 2e-5 appears optimal across different batch sizes. For RoBERTa, the learning rate of 1e-5 yields the best recall but should be balanced with precision considerations.
- **Emotion Recognition Performance:** In terms of emotion recognition, the choice between BERT and RoBERTa would depend on the specific requirements of the task. If identifying as many emotional instances as possible is paramount, RoBERTa with a higher recall would be preferred. If precision is more critical, BERT may be a better choice.

Baseline Comparison In evaluating the efficacy of our fine-tuned models for emotion recognition, a comparison with **state-of-the-art (SOTA)** models was conducted using the MELD dataset. As outlined in **Tables 6 and 7**, our fine-tuned BERT and RoBERTa models achieved weighted F1-Scores of 62.04% and 62.96%, respectively. Notably, our BERT model aligns closely with TRMSM-Att's performance, while our RoBERTa variant lags slightly behind the leading SAC-LSTM model, highlighting the potential for leveraging LSTM's sequential processing strengths. The benchmarking process was pivotal in contextualizing our models within the current landscape of emotion recognition research and set the stage for ongoing optimization

B. Audio Model Results

- **F1-Score Evaluation:** The Audio Wav2Vec2 Model exhibits a moderate F1-score of 0.34 as can be seen in **Table 8**, suggesting some challenges in achieving precision and recall balance.
- **Possible Reasons for Lower Score:**
  - *Limited Training Data:* Insufficient diverse training data might hinder the model's ability to generalize across various emotional expressions.
  - *Class Imbalance:* Uneven distribution of samples among emotion classes may affect the F1-score, especially in recognizing minority classes.
- **Possible Improvements:**
  - Fine-Tuning Parameters: Careful fine-tuning of hyperparameters, including learning rate and batch size, can optimize the model for better F1-score performance.
  - Addressing Class Imbalance: Implementing strategies to address class imbalance, like oversampling underrepresented classes, may improve the F1 score for all emotion categories.

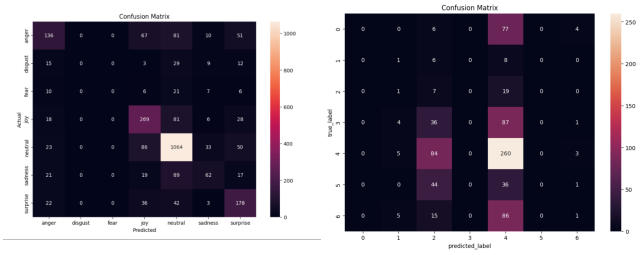

Figure 1: Confusion Matrix for RoBERTa (left) and Wav2Vec model (right)

C. Video Model Results

The classification report for the Video CLIP-Vit Model is presented in **Table 9**, summarizing precision, recall, and F1-Score metrics. The performance metrics are evaluated for each emotion category and are categorized into Accuracy, Macro Avg, and Weighted Avg. The Video CLIP-Vit Model results may deviate from expectations due to the following possible reasons:

Batch Size	Learning Rate	Weight Decay	Training Loss	Validation Loss	Accuracy	F1	Precision	Recall
8	1e-5	0.01	0.9198	1.1214	0.6411	0.6147	0.6178	0.6411
16	1e-5	0.01	0.9623	1.0991	0.6436	0.6153	0.5922	0.64
32	1e-5	0.01	1.2142	1.086	0.6406	0.6073	0.5873	0.6406

Table 4: Training and Validation results based on different Hyperparameters for Roberta

Weighted Accuracy	F1-Score
0.6464	0.6240
0.6548	0.6296
0.6517	0.6215

Table 5: Test results based on the different hyperparameters used in RoBERTa

Model	Weighted F1-Score
RGAT	60.91
BERT+MTL	61.90
TRMSM-Att	62.36
ERMC-DisGCN	64.22
Our fine-tuned BERT Model	**62.04**

Table 6: SOTA architecture vs Proposed model comparison (BERT)

Model	Weighted F1-Score
SACL-LSTM	66.86
CoMPM	66.52
COSMIC	65.21
DAG-ERC	63.65
Our fine-tuned RoBERTa Model	**62.96**

Table 7: SOTA architecture vs Proposed model comparison (RoBERTa)

	Precision	Recall	F1-Score
Accuracy			**0.34**
Macro Avg	0.52	0.15	**0.1**
Weighted Avg	0.59	0.34	**0.25**

Table 8: Classification Report for Audio Wav2Vec Model

- Lack of fine-tuning for emotion recognition.
- Limited diversity in training data may hinder the model's ability to generalize emotions.
- Mismatch in model architecture for emotion recognition can contribute to suboptimal performance.

Although the ViT model may be able to correctly classify the images in most of the video frames as shown in **Figure 2**, our implementation to get the overall emotion of the video may have further scope for improvement.

MULTI-MODAL FUSION MODEL

Our late fusion model (as seen in figure 3) synergistically combines textual, auditory, and visual inputs to enhance emotion recognition. Utilizing a weighted approach, we integrate the outputs of individual modality-specific models—leveraging the strengths of each to produce a unified prediction. Textual emotions are weighted most heavily, reflecting their significant contribution to context understanding. The audio and visual modalities provide complementary cues, essential for capturing the full spectrum of emotional expression.

	Precision	Recall	F1-Score
Accuracy			**0.24**
Macro Avg	0.31	0.2	**0.17**
Weighted Avg	0.34	0.24	**0.25**

Table 9: Classification Report for Video CLIP-Vit Model

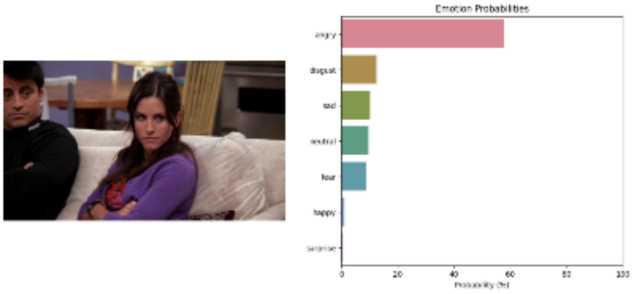

Figure 2: ViT Frame Sample

The results, as depicted in the provided **Table 10**, reveal the individual and collective efficacy of the modalities. Text (T) alone achieves an F1 Score of 0.6547, showcasing its prominence in conveying emotion. Audio (A) follows with a score of 0.3375, and Video (V) contributes a score of 0.2479. Surprisingly, the combined score (T+A+V) is 0.45, which enhances the performance of the individual audio and video metrics.

	T	A	V	T+A+V
F1 Score	0.6547	0.3375	0.2479	0.45

Table 10: Multimodel fusion Comparison

FUTURE WORK

Future research could focus on optimizing the multimodal fusion technique for efficiency.

Exploring methods to mitigate biases in training data and enhancing the models' ability to understand nuanced and complex emotional expressions would be beneficial. This

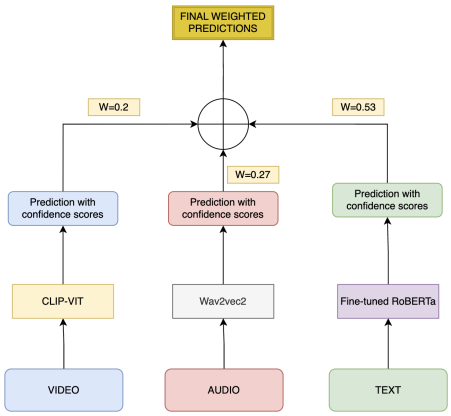

Figure 3: Our Proposed Weighed Average Multimodel Fusion Architecture

could be achieved through data augmentation techniques like Mix-Up Augmentation on video frames.

Additionally, extending the system to recognize a wider range of emotions, including subtle and culturally specific expressions, could improve its applicability in various contexts. Investigating lightweight models or transfer learning approaches might also offer pathways to more accessible and scalable emotion recognition systems.

CONCLUSION

Our findings reveal that each model brings unique strengths to the task. BERT and RoBERTa excelled in contextual understanding of text, Wav2Vec2 showed proficiency in capturing subtle auditory emotional cues, and CLIP-ViT effectively interpreted visual emotional expressions. The proposed multimodal fusion technique aimed to capitalize on these strengths, suggesting a promising direction for future emotion recognition research. The proposed multimodal fusion technique aimed to capitalize on these strengths for combined insight into the emotion being communicated in a scene.

However, the study also highlighted challenges, particularly in computational demands and the complexity of effectively integrating multimodal data. The potential for biases in the training data and the subjective nature of emotions pose additional hurdles.

This research contributes to the evolving field of emotion recognition by demonstrating the effectiveness of a multimodal approach. Advancements in accurately recognizing human emotions through AI have profound implications for various applications, from enhancing human-computer interaction to providing deeper insights into emotional dynamics in communication.

References

[1] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces*, ICMI '04, page 205–211, New York, NY, USA, 2004. Association for Computing Machinery.

[2] Soujanya Poria, Erik Cambria, Rajiv Bajpai, and Amir Hussain. A review of affective computing: From unimodal analysis to multimodal fusion. *Information Fusion*, 37:98–125, 2017.

[3] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy, July 2019. Association for Computational Linguistics.

[4] Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi. Towards multimodal sentiment analysis: Harvesting opinions from the web. In *Proceedings of the 13th International Conference on Multimodal Interfaces*, ICMI '11, page 169–176, New York, NY, USA, 2011. Association for Computing Machinery.

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy, 2017.

[6] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 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)*, pages 2236–2246, 2018.

[7] Tatiana Voloshina and Olesia Makhnytkina. Multimodal emotion recognition and sentiment analysis using masked attention and multimodal interaction. In *2023 33rd Conference of Open Innovations Association (FRUCT)*, pages 309–317, 2023.

[8] Md Shad Akhtar, Dushyant Chauhan, Deepanway Ghosal, Soujanya Poria, Asif Ekbal, and Pushpak Bhattacharyya. Multi-task learning for multi-modal emotion recognition and sentiment analysis. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *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)*, pages 370–379, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[10] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.