Team Members: Zhuoyuan Jiang, Yuan Chang

Multimodal Deep Learning for Slang and Informal Expression Detection

Problem Statement and Motivation

Non-native English speakers frequently struggle with understanding slang, phrasal verbs, and cultural references when learning through multimedia content. Traditional text-only analysis fails to capture the nuances conveyed through vocal tone, pauses, and visual context. This project investigates how multimodal deep learning can effectively detect and interpret informal expressions from textual, audio, and video sources, with the ultimate goal of developing a Chrome extension or software app that automatically highlights and contextualizes these expressions during multimedia consumption, seamlessly blending language acquisition with entertainment.

Related Work and Position

Prior research has focused on text-based informal language detection using transformer models like BERT and RoBERTa (Devlin et al., 2019; Liu et al., 2019), while audio-based emotion recognition has leveraged CNN and Wav2Vec2.0 architectures (Baevski et al., 2020). Multimodal approaches combining textual and audio inputs have shown promise in emotion detection through datasets like MELD (Poria et al., 2019) and IEMOCAP (Busso et al., 2008). However, there remains a significant gap in multimodal approaches specifically targeting slang and informal expression detection using deep fusion techniques that incorporate video alongside text and audio.

Approach

The project will focus on fusing textual, audio, and video modalities:

1. Textual Processing: Fine-tune a pre-trained RoBERTa model on labeled datasets containing slang and informal expressions
2. Audio Processing: Utilize Wav2Vec2.0 to extract audio features representing vocal tone and speaker intent
3. Video Processing: Extract visual cues and gestures using pre-trained vision models to capture contextual information
4. Multimodal Fusion: Implement a transformer-based fusion model with cross-attention mechanisms to effectively capture interactions between all three modalities

5. Extension Development: Create a Chrome extension that applies the model to highlight and save informal expressions encountered during multimedia consumption

Experimental Plan

Datasets

- Text-Only: Twitter Sentiment140 and Urban Dictionary datasets for baseline slang classification
- Multimodal: MELD and IEMOCAP datasets providing textual transcripts, audio recordings, and video content

Evaluation Metrics

- Quantitative: Accuracy, precision, recall, F1-score, BLEU, ROUGE scores, and etc
- Qualitative: Manual inspection of extracted expressions, attention heatmaps, and error analyses

Validation Approach

1. Establish baseline performance using single-modality models (text-only, audio-only, video-only)
2. Compare against bimodal (text-audio, text-video, audio-video) and full trimodal fusion approaches
3. Conduct ablation studies to determine the contribution of each modality
4. Perform user studies with language learners to assess the practical utility of the Chrome extension

This project will address the existing gap in multimodal slang detection literature by incorporating all three modalities—text, audio, and video—while developing a practical tool that enhances language learning through everyday multimedia consumption.