Final Project Evaluations

Predicting Valence and Arousal by Aggregating Acoustic Features for Acoustic-Linguistic Information Fusion

Published Conference:- IEEE REGION 10 CONFERENCE (TENCON)
Published Year:- 2020

Presented By:-

Himanshu Gupta (2022102002)

Dan Koshy George (2024814001)

Overview:

1. Objective
2. Work Done
   - Extracted acoustic and linguistic features at low level (frames and words)
   - Aggregated features to utterance level
   - Concatenated the acoustic and linguistic vectors to obtain input
   - Used EmoEvaluation data to train the SVR model and predict valence and arousal.
3. Result and analysis
4. Detailed acoustic Feature
   - MFCC, Prosody, Emobase feature sets (mean and maximum)
   - Their use in Valence and arousal
5. Detailed linguistic Feature
   - Direct sentence embeddings using BERT
   - Pretrained word embeddings that were aggregated for each sentence
     - Word2Vec
     - Fasttext
     - Glove
6. Future Scope

Objective:

The main objective of the project is to create an SVR model to predict valence and arousal (emotional attributes) based on speech and text data from the IEMOCAP dataset.

Work Done (Features Extraction):

We used several different feature sets for acoustic data **(MFCCs, prosody, Emobase, Compare)** as well as linguistic data **(BERT, Word2Vec, Fasttext and GloVe)**.

Two different aggregation methods were used for both, to obtain utterance level features (**mean and max** pooling for **audio**, and **mean and sum** for **text**).

All the different possible combinations of features were then used to train SVR models, all the models were then evaluated on a test set, and the R2 score, Explained Variance score, Root Mean Squared Error and Max Error metrics were used. The total number of models trained is 64.

Linguistic Features

BERT: We used the sentence_transformers library to extract sentence embeddings using both the **MiniLM L6 v2** and **MPnet base v2 models**. The BERT architecture is famous for its bidirectional approach, making encodings more robust and context-heavy.

Word embeddings:

Word2Vec: A popular method that learns word embeddings by predicting the surrounding words in a context window. Google’s pre trained embeddings were used.

GloVe (Global Vectors for Word Representation): Learns word embeddings by factorising the co-occurrence matrix of words. The 42B parameter Common Crawl embeddings were used.

FastText: Extends Word2Vec by representing each word as a bag of character n-grams and learning embeddings for these n-grams. The cc.en.bin.300 embeddings were used.

Acoustic Features

MFCCs for ASR: Mel-Frequency Cepstral Coefficients (MFCCs) are widely used in Automatic Speech Recognition (ASR) as they capture the spectral envelope of speech, closely mimicking human auditory perception. Their robustness to variations in pitch and speaker identity makes them foundational for speech decoding tasks.

Prosody: Prosodic features such as pitch, intensity, and rhythm convey emotional cues in speech. These features help in emotion detection as they reveal patterns in speech dynamics, like stress or intonation changes, which are critical for expressing emotions.

Emobase 2010: Emobase 2010 is an OpenSMILE feature set designed for emotion recognition, combining spectral, prosodic, and voice quality features. Its broad feature representation enables effective differentiation of emotional states across varied vocal expressions.

ComParE Feature Set: The Computational Paralinguistics Challenge (ComParE) feature set is comprehensive, covering spectral, prosodic, and functional descriptors. It is valuable for emotion detection as it captures subtle paralinguistic variations that correlate with emotional expressions.

Result and Analysis:

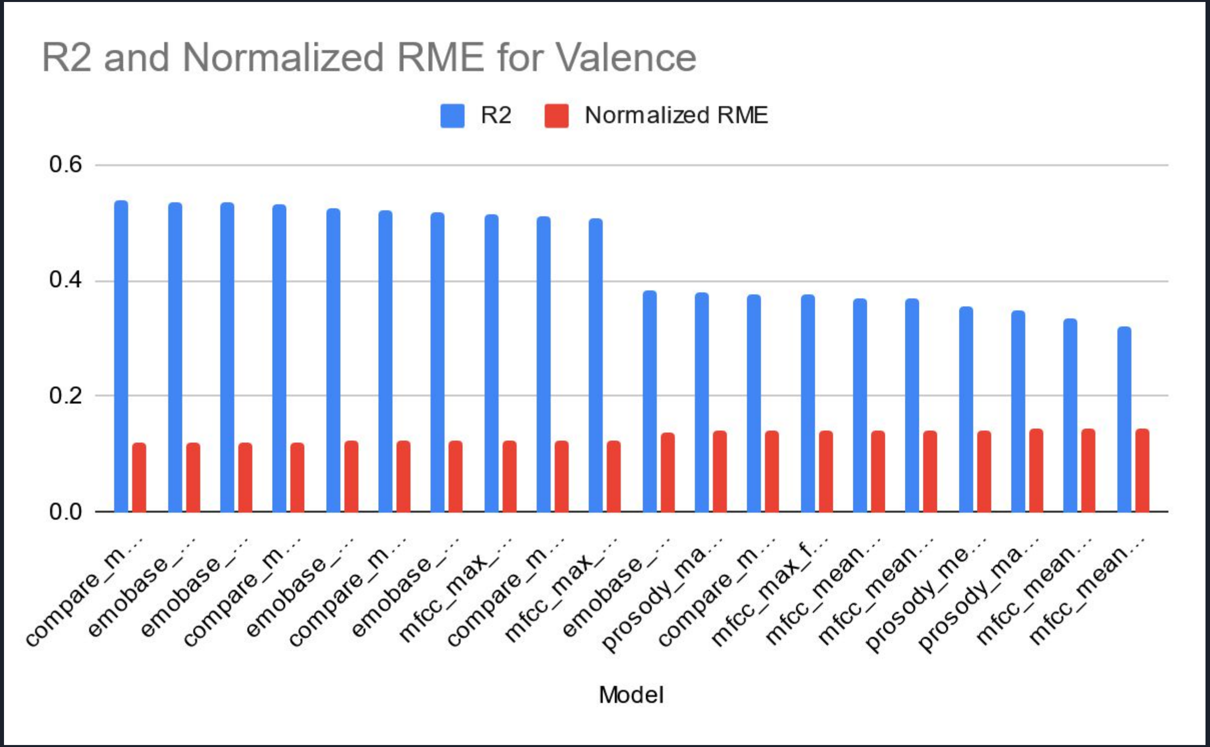
Metrics:

R2 Score: The coefficient of determination, also commonly denoted as R^2 is a common regression metric given by the proportion of the variation in the dependent variable that is predictable from the independent variable(s).

Root Mean Squared Error: The Root Mean Squared Error (RMSE) is one of the two main performance indicators for a regression model. It measures the average difference between values predicted by a model and the actual values. It provides an estimation of how well the model is able to predict the target value (accuracy).

Result and Analysis:

- Results for Valence:

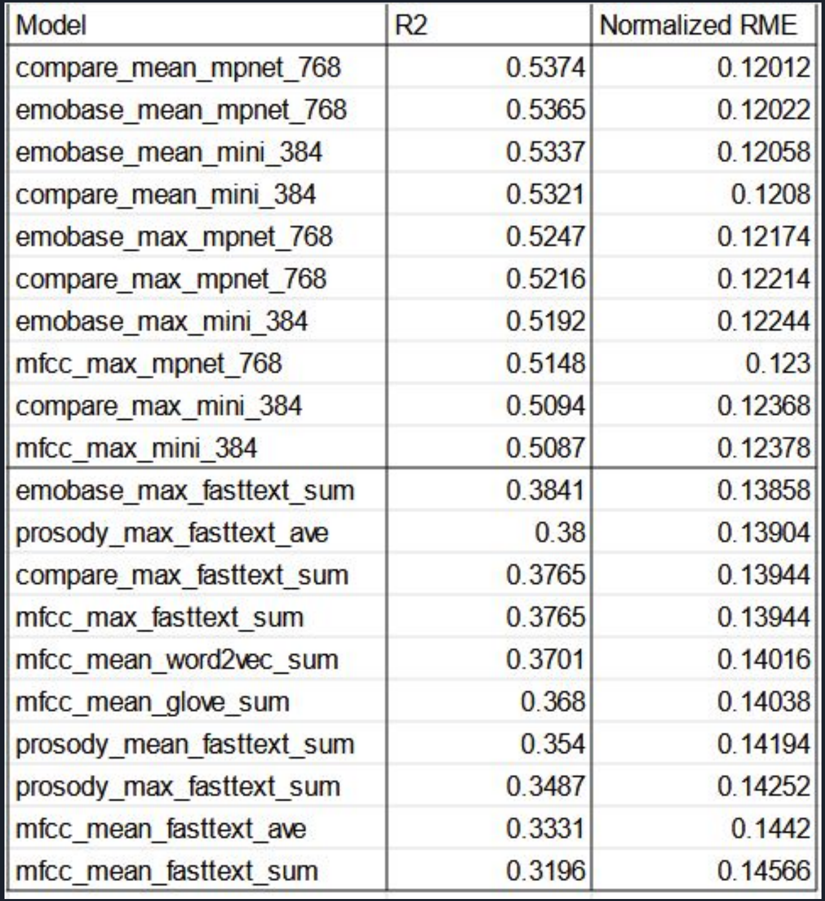


Result:

Dimensions:

Compare: 66
Emobase: 39
MFCC: 41
Prosody: 5

MPnet: 768
Mini: 384
Fasttext: 300
GloVe: 300
Word2Vec: 300



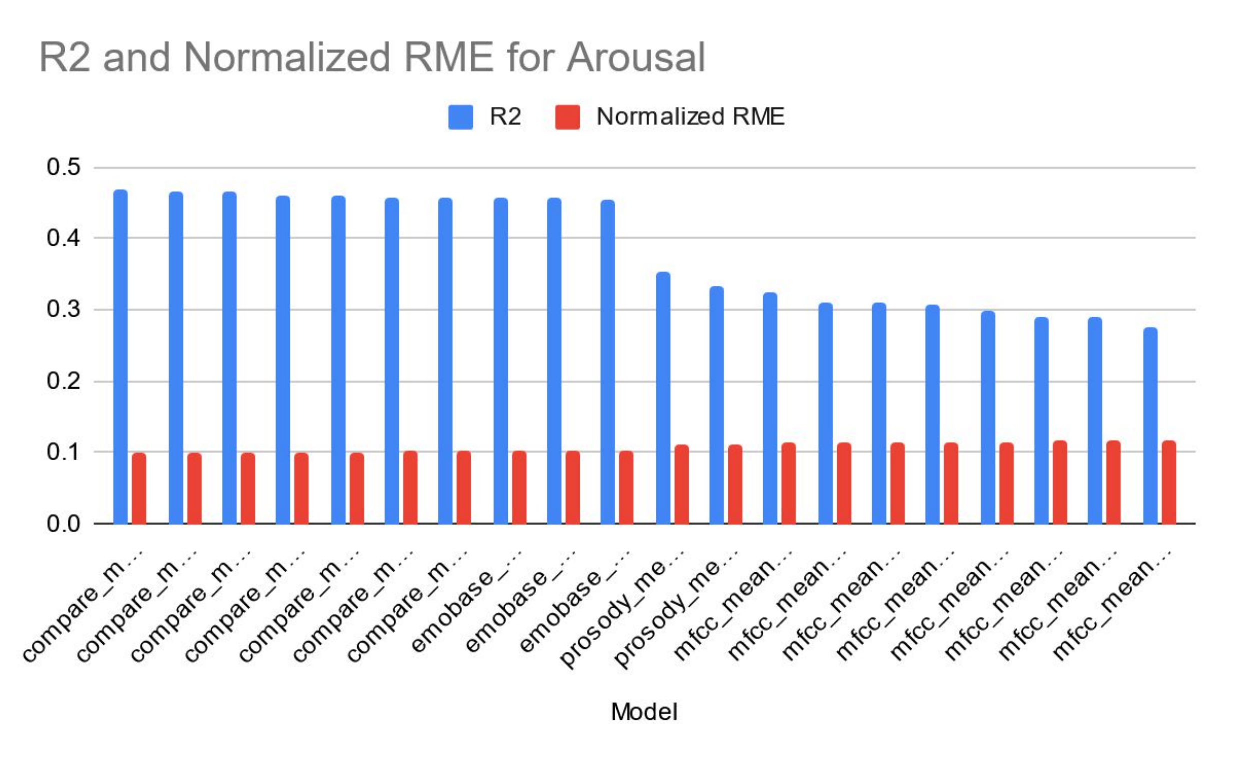
Model	R2	Normalized RME
compare_mean_mpnet_768	0.5374	0.12012
emobase_mean_mpnet_768	0.5365	0.12022
emobase_mean_mini_384	0.5337	0.12058
compare_mean_mini_384	0.5321	0.1208
emobase_max_mpnet_768	0.5247	0.12174
compare_max_mpnet_768	0.5216	0.12214
emobase_max_mini_384	0.5192	0.12244
mfcc_max_mpnet_768	0.5148	0.123
compare_max_mini_384	0.5094	0.12368
mfcc_max_mini_384	0.5087	0.12378
emobase_max_fasttext_sum	0.3841	0.13858
prosody_max_fasttext_ave	0.38	0.13904
compare_max_fasttext_sum	0.3765	0.13944
mfcc_max_fasttext_sum	0.3765	0.13944
mfcc_mean_word2vec_sum	0.3701	0.14016
mfcc_mean_glove_sum	0.368	0.14038
prosody_mean_fasttext_sum	0.354	0.14194
prosody_max_fasttext_sum	0.3487	0.14252
mfcc_mean_fasttext_ave	0.3331	0.1442
mfcc_mean_fasttext_sum	0.3196	0.14566

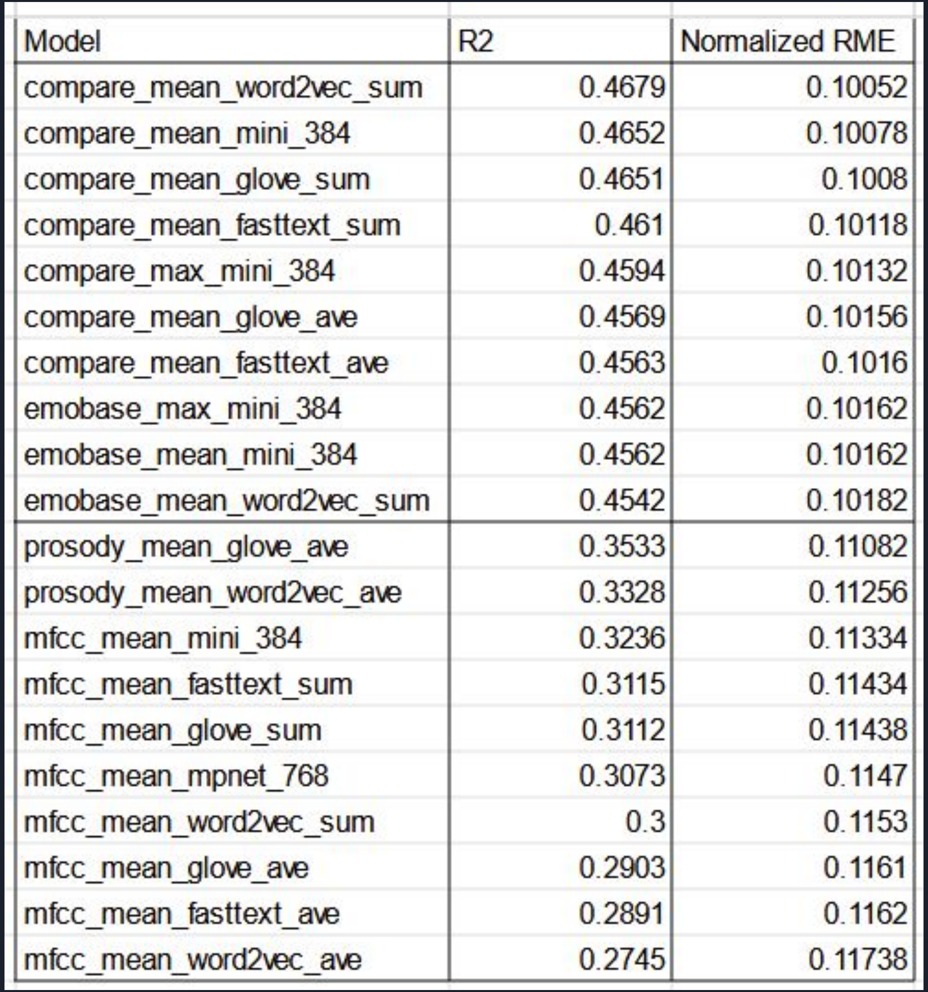
Observations

- The top 10 are completely dominated by the 768 and 384 sized BERT text embeddings.
- Acoustic feature sets are almost tied between Compare and Emobase, with two appearances by MFFC
- The bottom 10 are dominated by Fasttext, particularly with sum aggregation, which is the single worst performing text embedding
- Though Emobase, Compare and Prosody all appear in the bottom 10, MFFC, particularly with mean aggregation, got last place

Result and Analysis:

- Results for Arousal:





Model	R2	Normalized RME
compare_mean_word2vec_sum	0.4679	0.10052
compare_mean_mini_384	0.4652	0.10078
compare_mean_glove_sum	0.4651	0.1008
compare_mean_fasttext_sum	0.461	0.10118
compare_max_mini_384	0.4594	0.10132
compare_mean_glove_ave	0.4569	0.10156
compare_mean_fasttext_ave	0.4563	0.1016
emobase_max_mini_384	0.4562	0.10162
emobase_mean_mini_384	0.4562	0.10162
emobase_mean_word2vec_sum	0.4542	0.10182
prosody_mean_glove_ave	0.3533	0.11082
prosody_mean_word2vec_ave	0.3328	0.11256
mfcc_mean_mini_384	0.3236	0.11334
mfcc_mean_fasttext_sum	0.3115	0.11434
mfcc_mean_glove_sum	0.3112	0.11438
mfcc_mean_mpnet_768	0.3073	0.1147
mfcc_mean_word2vec_sum	0.3	0.1153
mfcc_mean_glove_ave	0.2903	0.1161
mfcc_mean_fasttext_ave	0.2891	0.1162
mfcc_mean_word2vec_ave	0.2745	0.11738

Observations

- In this case, the top 7 spots are held by Compare, with Emobase filling in the rest.
- In linguistic features, the 384 sized embeddings are the most common, but scattered, with the very top spot going to Word2Vec with sum aggregation.
- The 8 lowest models all used MFFC with mean aggregation, with the remaining 2 in the bottom 10 being Prosody.
- The linguistic features are mostly evenly represented in the bottom 10.

Link for all result:-

https://drive.google.com/file/d/1vbQ-KlyEiNk6XevqSsXSuO5pK4BbPZjR/view?usp=drive_link

Overall Observations and Conclusions

- For valence, linguistic features were the main contrast (BERT on top, Fasttext on the bottom)
- For arousal, acoustic features took the lead (Compare on top, MFCC on the bottom)
- This is because we can get the positive/negative part of emotion more easily from semantics, while prosody is uniquely suited to capture arousal.
- In linguistic features, average aggregation did better for valence, while sum aggregation did better for arousal
- In acoustic features, both forms of aggregation were close to even for both valence and arousal, with mean aggregation slightly ahead.
- MFCC with mean aggregation was the single worst acoustic feature set, while there is no consistently low performing linguistic feature set.
- In acoustic feature sets, Compare was the best performing, closely followed by Emobase, and the 368 sized BERT embeddings are the best choice if only one linguistic feature set has to be chosen.

Detailed Acoustic Features:-

1. MFCC:-

- We extract MFCCs along with their statistical features (like maximum and mean) for valence and arousal prediction. As it provides additional layers of information that are critical for understanding emotional content in speech.

MFCCs are a way to simplify and summarize sound signals in a way that makes sense to how humans hear. They focus on the most important parts of the sound that affect how we perceive it, making the data easier to work.

- It captures the spectral envelope, which reflects vocal tract characteristics.
- It represent the sound in a way that aligns with human auditory perception (via the Mel scale).
- It provides robust features for analyzing changes in tone, and spectral dynamics

These all are crucial for emotional expression.

Mean and Maximum of MFCC:-

While MFCCs provide frame-wise spectral information, speech is a dynamic signal, and emotions are expressed over time. So to account for this we calculate mean and maximum:-

1. Maximum MFCC Values:-

- Represent the highest energy concentrations in specific frequency bands.
- Indicate emotional intensity, as arousal often corresponds to high energy in certain bands.

2. Mean MFCC Values:-

- Provide an overall average representation of the spectral characteristics of the speech.
- Capture the general tone which is linked to emotional valence (e.g., positive vs. negative emotions).

By using these statistics feature, we look into the time-varying nature of speech into a fixed-length features, which makes them suitable for machine learning models to process.

Use in Valence and Arousal Prediction:-

1. MFCC Features to Valence

- MFCCs mean capture overall spectral brightness, which is higher in positive emotions (e.g., happiness) and lower in negative emotions (e.g., sadness).
- MFCCs maximum Reflect moments of peak spectral intensity, which might occur in high-valence emotional expressions.

2. MFCC Features to Arousal

- High arousal emotions (e.g., anger, excitement) often have higher energy in certain frequency bands, leading to distinct peaks in MFCC values.
- Low arousal emotions (e.g., calm, sadness) typically show more muted spectral energy, reflected in lower mean values.

2. Prosody

We extract the prosodic features as it captures the rhythm, intonation, and stress patterns in speech, which are critical for conveying emotions. Extracting the mean and maximum values of these features provides a way to compute speech characteristics over time and make them useful for predicting valence (positivity/negativity) and arousal (intensity of emotion).

1. Emotion Representation:-

- Prosody features, such as pitch (F0), energy (intensity), and speaking rate, are directly linked to how emotions are expressed in speech.
- For example:- Excitement or anger might involve higher pitch and faster speaking rates and sadness might involve slower speech with lower energy.

2. Complementary to MFCCs:-

- While MFCCs capture the timbral (tone-related) characteristics, prosody focuses on the temporal and dynamic patterns of speech, providing a broader view of emotional expression.

Mean and Maximum of prosody features:-

1. Mean Prosody Values:-

- It represent the average behavior of a feature over time.
- Example: A higher mean pitch might indicate a more excited or positive emotional state, while a lower mean pitch could signal calmness or sadness.

2. Maximum Prosody Values:-

- It captures the most extreme (peak) expressions of a feature.
- Example: A high maximum energy might indicate moments of intense emotion or high arousal.

Prosody features can vary significantly within a single utterance. Calculating statistical values like mean and max helps us to summarize the changing patterns in speech into a fixed-sized data, which make it easier to use in models.

Use in Valence and Arousal Prediction:-

1. Prosody Features to Valence Prediction

- Positive emotions (e.g., happiness) often have higher mean pitch and energy, with faster speech rates.
- Negative emotions (e.g., sadness) tend to have lower pitch and energy, with slower rates and longer pauses.

2. Prosody Features to Arousal Prediction

- High arousal emotions (e.g., anger, excitement) have higher maximum energy, larger pitch variations, and faster speech rates.
- Low arousal emotions (e.g., calmness) have lower energy, steadier pitch, and slower speech rates.

From prosody features into statistics like mean and max, we create consistent inputs for our models, which makes it easier to predict emotional states.

3. Emobase

EmoBase is a feature set designed for analyzing emotional expressions in speech. It includes a comprehensive range of acoustic features such as:

- MFCCs: Capture spectral properties aligned with human hearing.
- Pitch (F0): Reflects intonation patterns linked to emotions.
- Energy: Indicates intensity, useful for detecting arousal levels.
- Duration and Tempo: Capture speech rate and pauses, relevant for emotional context.

EmoBase is widely used in emotion recognition tasks because it provides a rich set of features for modeling emotional states like valence and arousal.

Detailed Linguistic Features:-

1. BERT (Bidirectional Encoder Representations from Transformers)

We use BERT to extract linguistic features because it captures the meaning of words and their context within sentences, making it highly effective for predicting valence (positivity/negativity) and arousal (emotional intensity) in text-based emotion analysis.

Why Use BERT for Extracting Linguistic Features?

- ❖ Contextual Understanding: BERT understands a word's meaning based on surrounding words, capturing subtle differences like "fine" in "I am fine" vs. "I am fine!".
- ❖ Bidirectional Encoding: It reads text both ways, capturing richer meanings than one-directional models.
- ❖ Pre-Trained Knowledge: Trained on large datasets, it recognizes common emotional patterns.
- ❖ Versatility: Extracts features at word, sentence, or document levels, adapting to various emotion analysis needs.

Use in Valence and Arousal Prediction:-

1. Valence Prediction:

- Captures emotional tone in context, e.g., "I loved the movie!" (high valence) vs. "The movie was terrible." (low valence).
- Detects nuances like sarcasm, e.g., "Oh great, another traffic jam." (negative sarcasm).

2. Arousal Prediction:

- Identifies emotional intensity, e.g., "I’m so excited!" (high arousal) vs. "I feel calm." (low arousal).
- Tracks changes in intensity, e.g., "It started slow but got exciting by the end."

Advantages of Using BERT:

❖ Handles word ambiguity based on context.

❖ Works well for both short and long texts.

❖ Can be fine-tuned for accurate emotion prediction.

MiniLM-L6-v2 and MPNet-base-v2:

1. Size and Speed

- MiniLM-L6-v2: Small (6 layers), faster, and memory-efficient.
- MPNet-base-v2: Larger (12 layers), slower but more powerful.

2. Training Approach

- MiniLM: Uses distillation for efficiency.
- MPNet: Combines masked and permuted language modeling for better context.

3. Performance

- MiniLM: Good for quick tasks with limited resources.
- MPNet: Better for high-accuracy tasks needing deep language understanding.

4. Use Cases:

- MiniLM: Real-time or mobile apps.
- MPNet: Advanced semantic tasks like document search.

2. Word2Vec

Word2Vec converts words into numerical vectors that capture their meanings and relationships, helping predict valence and arousal by clustering words with similar emotions.

Why Use Word2Vec for Extracting Linguistic Features?

- **Semantic Representation:** Word2Vec turns words into vectors that capture their meanings and relationships, e.g., happy and joyful have similar embeddings and revealing emotional patterns.
- **Efficient Training:** It uses CBOW or Skip-gram models to quickly learn word representations from large text datasets.
- **Pre-Trained Models:** Ready-to-use pre-trained models, like Google News, save time and provide high-quality features.
- **Low Computational Cost:** Word2Vec is lightweight and ideal for simpler tasks or limited resources compared to complex models like BERT.

Use in Valence and Arousal Prediction:-

1. Valence Prediction:

- Words with similar emotions (e.g., happy, joyful) have similar embeddings, while negative words (sad, angry) form separate clusters. **Example:** Sentences with success or love suggest high valence, while failure or hate indicate low valence.

2. Arousal Prediction:

- High-energy words (excited, shouting) cluster together, while low-energy words (calm, silent) are separate. **Example:** Sentences with excited suggest high arousal, while relaxed indicates low arousal.

Word2Vec embeddings can be averaged to summarize the emotional tone of longer texts like sentences or paragraphs.

3. Glove

We use GloVe to extract linguistic features because it captures word meanings and relationships by combining co-occurrence statistics and embeddings. This helps predict valence (positivity/negativity) and arousal (emotional intensity) with meaningful numerical representations.

Why Use Glove for Extracting Linguistic Features?

- **Global Context**: Captures broad semantic relationships by analyzing word co-occurrence across the entire corpus (e.g., happy and joy have similar embeddings).
- **Pre-Trained Models**: Ready-to-use embeddings trained on large datasets like Wikipedia save time and effort.
- **Compact Vectors:** Creates dense, numerical representations of word relationships.
- **Semantic Insights:** Identifies analogies and emotional relationships, like happy → joyful and sad → unhappy.

Use in Valence and Arousal Prediction:-

1. Valence Prediction:

- Groups positive words (happy, love) and negative words (sad, failure) into distinct clusters, capturing emotional tone.

2. Arousal Prediction:

- Differentiates high-energy words (excited, thrilling) from low-energy words (calm, relaxed), capturing emotional intensity and identify the contextual pattern.

The Text Representation combines word embeddings to represent the overall emotional tone of sentences or longer texts.

Advantages of Using GloVe:

- ❖ Semantic Coherence:- Captures word relationships and emotional nuances.
- ❖ Efficient:- Pre-trained embeddings are lightweight and faster than models like BERT.
- ❖ Adaptable:- Effective for both general and domain-specific datasets.

Future Scope

- Word/Sentence embeddings trained directly on a Sentiment Analysis corpus, for even better valence predictions.
- Extending the model to be more complete by predicting emotion directly instead of valence and arousal values.
- A special acoustic feature set tailored for the task.
- Multimodal Integration: Combine audio, text, and visual data for more accurate emotion recognition.
- Personalized Models: Develop systems that adapt to individual emotional patterns for better predictions.
- Real-Time Applications: Build efficient models for real-time emotion tracking in virtual assistants and mental health tools.

Github link:- https://github.com/DanTheMan314/SAL-Project

THANK YOU!