# 11-711 HW#3: Project Proposal State-of-the-art Reimplementation

Amanda Shen, Boyi Qian, Yiwen Zhang

## 1 Literature Review

Our group mainly focused on the field of Emotion Recognition of all kinds, and tried to find the most interesting topic to work on. We conducted literature survey for the following 4 topics: Fined-grained emotion recognition with transformers, semi-supervised graph-based models for emotion recognition, unsupervised pretraining method with knowledge graph, and commonsense-based supervised model for conversational emotion detection.

### 1.1 Choice of Topic

We attempted re-implementation for all 4 methods within our team (see contributions in contributions.txt), re-evaluated the scope and feasibility of all methods, and decided to work on the task of Commonsense Emotion Recognition for Conversation as our final topic for the project.

1. **GoEmotions**:
    - GoEmotions [1] is a large-scale dataset for fine-grained multi-label emotion recognition task. It contains 58k manually annotated English Reddit comments, with a set of 27 categories + Neutral as the prediction labels. The official code used a pretrained bert-based model for finetuning, and applied transfer learning to test the model on different label sets (grouped and Ekman).
    - This was the first choice of our project, and we've also tested out the official code for re-implementation(check Table 5.1, 5.2 and 5.3 for results). We revised the original tensorflow framework to a pytorch version. A majority of the results are close to the official paper release(0~5% boost in precision and f1-score, as we've trained for more epochs than recommended). However, we could not achieve the same results for recall rates(some classes among the 27 are 5% below official results), probably due to framework change and implementation difference.
    - This task focuses on emotion recognition of multi-labels on a large dataset + diverse taxonomy. Compared to our final choice of Conversational Emotion Recognition with Commonsense, it focuses on comments rather than dialogs, which are independent of each other and do not have the interactions between speakers and listeners.

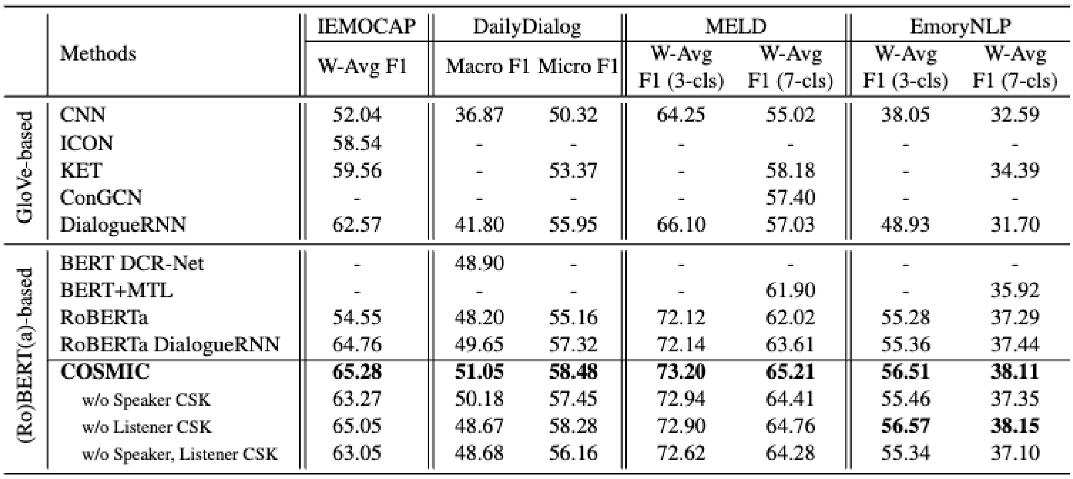
2. **CARER**:
    - This paper [2] proposed a semi-supervised method using graph algorithms for emotion recognition, which is aimed at capturing complex relationships between different linguistic elements. It also utilizes word embeddings to enrich pattern-based features.
    - The paper also proposed a new dataset, which was collected through noisy labels and annotated via distant supervision. Data was labeled into 6 emotions: anger, fear, joy, love, sadness, and surprise.
    - Authors of this paper did not provide official implementation for their algorithm, but we believe the methods mentioned in the paper and its derivative dataset can be very helpful to our project.
3. **SKEP**:
    - SKEP [3] proposed a pre-training algorithm enhanced with emotion-related knowledge. This algorithm employs an unsupervised method to automatically mine emotional knowledge, and then utilizes this emotional knowledge to construct pre-training objectives, which enables machines to understand emotional semantics.
    - The model primarily consists of two parts: 1. Emotional Masking, which automatically mines emotional knowledge in an unsupervised manner to identify emotional words in the input sentence, along with the polarity of these words and aspect-emotion pairs. These emotional words and aspect-emotion pairs are then masked to create a corrupted version of the sentence. 2. Recovery of these masked emotional words, their polarities, and the aspect-emotion pairs from the corrupted sequence.
    - We encountered issues while trying to replicate the code, specifically when installing the project dependency PaddlePaddle 1.6.3, which we were unable to resolve. However, the approach provided in the paper is greatly beneficial for our subsequent work.
4. **COSMIC**:
    - The COSMIC (Commonsense Ontology for Situations, Motivations, Interactions, and Characters) [4] framework is a novel approach which implements a common-sense guided framework for emotion identification in conversations. It focuses on extracting rich commonsense knowledge from knowledge graphs, including personality, events, mental states, intents and emotions, and eventually incorporates those features with context-independent sentence representations extracted from RoBerta-large models.
    - The paper delves into developing sophisticated models capable of processing and interpreting complex conversational elements. A key focus is on integrating commonsense reasoning to context independent representations. The causal relations would not only boost performance on all datasets, but also add interpretability to the results generated.
    - The paper incorporated representations learnt by COMET on ATOMIC knowledge graph, and tested out results for this add-on on the speaker side, the listener side, and both sides.

## 1.2 Literature Review in ERC

More specifically, for Emotion Recognition in Conversations, we did a literature survey as well.

- [5] provides a comprehensive review over the field. It discusses challenges in emotion recognition in conversations, reviews various datasets, and summarizes recent advances in the field. Methods were tested on IEMOCAP and DailyDialog.

- [6]proposes DialogXL, an extension of the XLNet model, to leverage **cross-utterance information** for better emotion recognition in dialogues. Methods were tested on IEMOCAP and DailyDialog.



| | Methods | IEMOCAP | DailyDialog | | MELD | | EmoryNLP | |
|---|---|---|---|---|---|---|---|---|
| | | W-Avg F1 | Macro F1 | Micro F1 | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) | W-Avg F1 (3-cls) | W-Avg F1 (7-cls) |
| GloVe-based | CNN | 52.04 | 36.87 | 50.32 | 64.25 | 55.02 | 38.05 | 32.59 |
| | ICON | 58.54 | - | - | - | - | - | - |
| | KET | 59.56 | - | 53.37 | - | 58.18 | - | 34.39 |
| | ConGCN | - | - | - | - | 57.40 | - | - |
| | DialogueRNN | 62.57 | 41.80 | 55.95 | 66.10 | 57.03 | 48.93 | 31.70 |
| (Ro)BERT(a)-based | BERT DCR-Net | - | 48.90 | - | - | - | - | - |
| | BERT+MTL | - | - | - | - | 61.90 | - | 35.92 |
| | RoBERTa | 54.55 | 48.20 | 55.16 | 72.12 | 62.02 | 55.28 | 37.29 |
| | RoBERTa DialogueRNN | 64.76 | 49.65 | 57.32 | 72.14 | 63.61 | 55.36 | 37.44 |
| | **COSMIC** | **65.28** | **51.05** | **58.48** | **73.20** | **65.21** | **56.51** | **38.11** |
| | w/o Speaker CSK | 63.27 | 50.18 | 57.45 | 72.94 | 64.41 | 55.46 | 37.35 |
| | w/o Listener CSK | 65.05 | 48.67 | 58.28 | 72.90 | 64.76 | **56.57** | **38.15** |
| | w/o Speaker, Listener CSK | 63.05 | 48.68 | 56.16 | 72.62 | 64.28 | 55.34 | 37.10 |

Figure 2.1: Baseline Results

- [7] introduces EmoBERTa, an adaptation of the RoBERTa model, tailored for emotion recognition in conversations with a focus on **speaker-awareness.** The method was evaluated on the EmpatheticDialogues dataset.
- [8] focuses on multimodal emotion recognition, integrating data from textual, auditory, and visual channels using advanced deep learning architectures. Methods were tested on the IEMOCAP and CMU-MOSEI datasets.
- [9]investigates **cross-cultural** aspects of emotion recognition in conversational AI, aiming to enhance the understanding of emotional expressions across different cultures. It employs culturally diverse datasets, including some language-specific emotional conversation datasets.

For the review we've done so far, we observed a trend for **utilizing transformer-based models** like RoBERTa and BERT, multimodal approaches, and the exploration of cross-cultural and **speaker-aware aspects**. Thus, we believe our baseline choice of COSMIC should be well representing the trend, which not only utilizes RoBerta as the word representation encoder, but also merged other knowledge base, and performed speaker-listner aware designs as well.

## 2 COSMIC Reimplementation

In our study, we retrained the COSMIC model from the referenced paper using the IEMOCAP dataset, and achieved a comparable result of **65.30%**, closely matching the original paper's 65.28% as shown in Figure 2.1.

### 2.1 Dataset: IEMOCAP

Among all the 4 datasets, we chose IEMOCAP as the main dataset to be trained and tested on, both for our baseline implementation and future research.

- **Dataset introduction**: IEMOCAP is a rich dataset featuring two-person conversations from ten unique speakers. The training dialogues are derived from the first eight speakers, while the testing dialogues involve the remaining two. Each utterance within this dataset is meticulously annotated with one of six emotions: happy, sad, neutral, angry, excited, or frustrated. The daily conversational property of IEMOCAP fits well with the design of the model, which utlizes COMET trained ON ATOMIC(also an everyday action based conversational commonsense knowledge base) as the commonsense source.

- **Why we choose IEMOCAP over other datasets**: The original paper conducted an in-depth case study exclusively on IEMOCAP. This study highlighted a test conversation where the emotional tone shifts rapidly, presenting a challenge for state-of-the-art models like DialogueRNN, which often struggle with sudden emotional transitions and subtle distinctions between similar emotions. In contrast, the COSMIC model, with its emphasis on commonsense knowledge propagation, demonstrates a superior ability to navigate these complexities.

  For instance, in a scenario where the conversation shifts from neutral to frustrated and back, COSMIC effectively identifies and differentiates between nuanced emotional states like frustration and anger. The model leverages commonsense insights about the speakers' reactions and the listeners' responses, enhancing its accuracy in classifying emotions. This nuanced understanding, as illustrated in our case study, underscores the value of re-implementing the IEMOCAP dataset in our research.

### 2.2 Reimplementation

In our research, we undertook a detailed re-implementation of the COSMIC model, applying it to the IEMOCAP dataset with a high degree of precision. This effort was in line with the methodologies detailed in the original paper, which emphasized the importance of averaging outcomes over multiple runs for robust result computation. Specifically, the original study's results were derived from an average of **five distinct runs**, with test scores being calculated at the point of optimal validation performance.

Adhering to this rigorous approach, we conducted a total of five runs for our experiment. During each run, we meticulously recorded the test accuracy, aligning it with the highest validation F1-score achieved. This methodical process was crucial in ensuring the reliability and consistency of our experimental outcomes.

The culmination of our extensive testing yielded a test accuracy of **65.30%**. This result is notably consistent with the 65.28% reported in the original study, demonstrating a remarkable alignment with the established findings. The close correlation of our results with those of the original paper not only validates our re-implementation process but also underscores the efficacy of the COSMIC model in narrative understanding tasks. This outcome significantly contributes to the existing literature, reinforcing the model's applicability and robustness in similar research contexts.

## 3 Error Case Analysis

In this section, we would explain the test results of the COSMIC model, and perform several analyses on error cases. For the six labels in the data, we used the abbreviations of the first three letters (Table 3.1 for detail with the corresponding **F1-score** of our test dataset). Table 3.2 shows three of **error prediction samples** for each label. We also analyze and visualize **label distribution** under different datasets and those are predicted wrong.

### 3.1 Analysis

We would analyze our test results in terms of per-class F1-score comparison, error case studies and distribution of the dataset.

|  | label | F1-score |
|---|---|---|
| Happiness | hap | 0.47385 |
| Sadness | sad | 0.79245 |
| Neutral | neu | 0.62315 |
| Anger | and | 0.53285 |
| Excitement | exc | 0.62879 |
| Frustration | fru | 0.66024 |

Table 3.1: The emotional meanings for each label and corresponding F1-score of test dataset

| Sentence | True label | Prediction |
|---|---|---|
| "Yeah, she-yeah, that's my type." | hap | exc |
| "Oh cool." | hap | exc |
| "six four nine four. That would be so great, thank you so much." | hap | neu |
| "No. I figured it was best to leave him alone." | sad | neu |
| "This what? What is this? This isn't even anything." | sad | fru |
| "Did you talk to him?" | sad | neu |
| "Hi, I need an ID." | neu | fru |
| "How did you know?" | neu | hap |
| "Right, but this is the wrong form. Somebody gave you the wrong form." | neu | fru |
| "What the hell is this?" | ang | fru |
| "There aren't any people upstairs; it's a photographer's studio." | ang | neu |
| "Well, you help me stay here." | ang | neu |
| "oh yeah." | exc | hap |
| "Oh my god, it was just last weekend." | exc | hap |
| "Yeah. Yeah, of course." | exc | hap |
| "I'm getting very bored with this conversation." | fru | ang |
| "And try to do something. I mean-" | fru | neu |
| "Yeah, and I look nice and I'm you know–" | fru | neu |

Table 3.2: Error prediction samples

### 3.1.1 Metrics (Per-class F1-score)

From the F1-scores obtained from the test dataset (Table 3.1), we can see that the model has the highest accuracy for predicting the '**sad**' label (about 0.79), followed by '**fru**' (frustration) and '**exc**' (excitement) with scores of approximately 0.66 and 0.63 respectively. The F1-scores for '**hap**' (happiness) and '**ang**' (anger) are lower (about 0.47 and 0.53), indicating that the model faces some difficulties in recognizing these two emotions.

### 3.1.2 Error Case Studies

If we randomly print out three incorrect predictions for each label (Table 3.2), we can observe that many sentences indeed have ambiguous emotions, and a significant portion requires context-based judgement. It appears that the pairs **'hap' and 'exc'**, as well as **'sad' and 'fru'**, are easily confused, both in terms of classification results and human interpretation of the sentence samples.

### 3.1.3 Visualization of mis-classified cases

To better understand the reasons behind the model's incorrect predictions, we calculated and visualized the distribution of incorrect predictions for each label (Figure 3.1). The results confirm our earlier assumptions: the labels with the most incorrect predictions were '**exc**' and '**neu**'.

A significant portion of the ’ang’ label mispredictions were predicted as ’fru’, while many in ’exc’ were incorrectly predicted as ’hap’, and vice versa. Notably, a large part of ’fru’ was wrongly predicted as ’neu’, which does not align with our assumption of ’ang’ and ’fru’ being a pair. Moreover, a substantial portion within ’neu’ was predicted as ’fru’. These findings will be crucial for the further improvement of our model. We would explain our possible improvements on this in the next section of Project Proposal.

### 3.1.4 Data Distribution

Observed that number of errors does not match our guesses for labels under-performing in F1-score, we also analyzed and visualized the distribution of different labels in each dataset (Figure 3.2). We found that ’fru’ (frustration) and ’neu’ (neutral) take up a large proportion. This might be the reason for the mutual misjudgment between these two labels observed earlier. Additionally, although ’exc’ (excitement) is not very numerous in the dataset, it has the highest number of errors, indicating that we need to improve its learning and prediction.

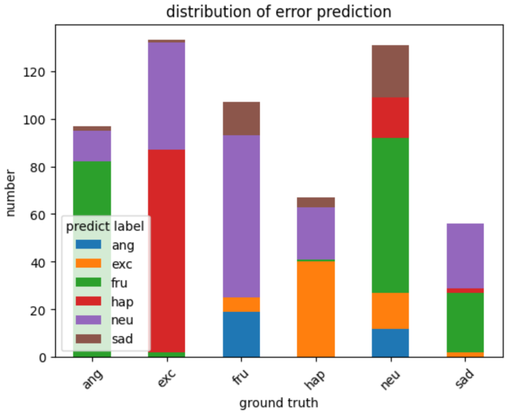


Figure 3.1: Distribution of error prediction under each label

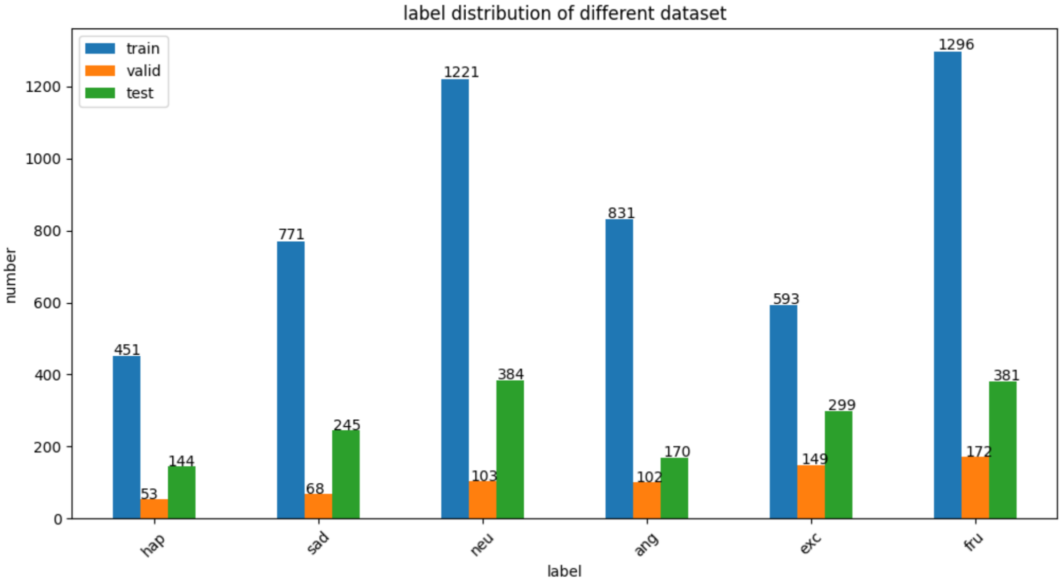


Figure 3.2: Label distribution under train/valid/test dataset

# 4 PROJECT PROPOSAL

Based on the previous literature review, re-implementation results and error analysis, we would like to make further improvements on our baseline for several aspects. Therefore, our project would be formed in the following way:

**Project Proposal:**

1. **Project Topic**: Supervised emotion recognition for conversational text input with a framework utilizing commonsense knowledge.

2. **Methodology**: Our baseline model can break down into the following 2 parts: Context-independent RoBerta representation, and the incorporation of commonsense vector from COMET into context. We would try to make improvements for these 2 parts separately.

   - **Context-independent representation**: For the backbone model that we use for context-independent word representation, we found that DialogRNN only got 2% of F1-score boost with Roberta learnt embeddings, compared to GLoVe-based embeddings. Thus, we plan to apply LoRA or other forms of Adapters mentioned in class to the transformer layers, so that it can largely save the parameters and compress the model consequently.
   - **Commonsense injection**: Instead of using the pretrained COMET representation of the utterance input, we plan to inject knowledge into language models by extending the original input (knowledge injection in class). To be more specific, given the utterance, we plan to use COMET to generate relations and append the result to the original sentence input, and then feed that into our backbone RoBerta model. This might not outperform the method mentioned in our baseline model paper, but it should simplify the vector representation to a large extent.

   Also, in terms of the error cases we analyzed in the previous section, we would like to make the following improvements:

   - **Data Augmentation**: We observed that 'happy' and 'angry' does not have the greatest number of error predictions on the test set, yet f1-score for both classes are way below average among all 6. We check the data distribution, and think that this might be due to class imbalance from the dataset. We found that most of the error cases of angry were misclassified as frustrated, which takes up a majority of the training and testing samples in the dataset. Thus, we're considering data augmentation or generating synthetic data for certain classes as the approach.
   - **Backbone Model**: According to the paper of our baseline model, the context-independent representation of utterance was trained solely on each dataset with the according labels. To enrich the representation of underrepresented emotion classes, we could choose models pretrained on multiple dataset, or finetune roberta over DailyDialog or CARER, and use that as our backbone model for encoding utterances.

3. **Dataset**: The dataset would be **IEMOCAP**. (We would like to add results on DailyDialog if time allows us to do so.)

   As previously explained, it's a dataset for dyadic daily conversation, which suits the choice of COMET as the commonsense generator (trained on everyday actions, ATOMIC). We aim at improving detecting subtle changes of emotions between utterances as hoped in the paper of our baseline model.

4. **Metrics**: In addition to the metrics of weighted f1-score performed by the original paper, we would run f1-score for each of the 6 classes of IEMOCAP, and do relevant error case analysis and ablation studies as well.

# 5 APPENDIX

| Emotion | Precision | Recall | F1-score |
|---|---|---|---|
| admiration | 0.5942 | 0.7321 | 0.6559 |
| amusement | 0.7253 | 0.9204 | 0.8113 |
| anger | 0.4718 | 0.5505 | 0.5081 |
| annoyance | 0.2885 | 0.4156 | 0.3405 |
| approval | 0.3225 | 0.3988 | 0.3566 |
| caring | 0.3452 | 0.4296 | 0.3828 |
| confusion | 0.4030 | 0.5163 | 0.4527 |
| curiosity | 0.4395 | 0.6021 | 0.5081 |
| desire | 0.4302 | 0.4457 | 0.4378 |
| disappointment | 0.2727 | 0.2781 | 0.2754 |
| disapproval | 0.3684 | 0.3932 | 0.3804 |
| disgust | 0.4295 | 0.4959 | 0.4603 |
| embarrassment | 0.4390 | 0.4864 | 0.4615 |
| excitement | 0.3720 | 0.4660 | 0.4137 |
| fear | 0.5656 | 0.7179 | 0.6327 |
| gratitude | 0.9059 | 0.9034 | 0.9046 |
| grief | 0.2 | 0.1666 | 0.1818 |
| joy | 0.6084 | 0.6273 | 0.6177 |
| love | 0.7188 | 0.8487 | 0.7784 |
| nervousness | 0.2857 | 0.2608 | 0.2727 |
| neutral | 0.6343 | 0.6485 | 0.6413 |
| optimism | 0.4912 | 0.6021 | 0.5410 |
| pride | 0.4615 | 0.375 | 0.4137 |
| realization | 0.2 | 0.2482 | 0.2215 |
| relief | 0.2666 | 0.3636 | 0.3076 |
| remorse | 0.5507 | 0.6785 | 0.608 |
| sadness | 0.5151 | 0.5448 | 0.5295 |
| surprise | 0.4748 | 0.6028 | 0.5312 |
| **macro-average** | **0.4694** | **0.5186** | **0.4886** |
| std | 0.1638 | 0.1893 | 0.1732 |

Table 5.1: Re-implementation of GoEmotion Taxonomy.

| **Sentiment** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|
| ambiguous | 0.5085 | 0.6174 | 0.5577 |
| negative | 0.6449 | 0.7052 | 0.6737 |
| neutral | 0.6267 | 0.6793 | 0.6519 |
| positive | 0.7685 | 0.8445 | 0.8048 |
| **macro-average** | **0.6296** | **0.7028** | **0.6636** |
| std | 0.0921 | 0.0831 | 0.0881 |

Table 5.2: Re-implementation of Sentiment-Grouped Data.

| **Ekman Emotion** | **Precision** | **Recall** | **F1-score** |
|---|---|---|---|
| anger | 0.4910 | 0.6033 | 0.5414 |
| disgust | 0.4796 | 0.4796 | 0.4796 |
| fear | 0.5811 | 0.6938 | 0.6325 |
| joy | 0.7667 | 0.8483 | 0.8055 |
| neutral | 0.6353 | 0.6843 | 0.6589 |
| sadness | 0.5395 | 0.6121 | 0.5735 |
| surprise | 0.5141 | 0.6159 | 0.5604 |
| **macro-average** | **0.5604** | **0.6318** | **0.5935** |

Table 5.3: Re-implementation of Ekman's Taxonomy.

# References

[1] Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, Sujith Ravi (2020). *GoEmotions: A Dataset of Fine-Grained Emotions* https://arxiv.org/pdf/2005.00547.pdf

[2] Saravia, E., Liu, H. C. T., Huang, Y. H., Wu, J., & Chen, Y. S.(2018). *CARER: Contextualized Affect Representations for Emotion Recognition*,Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 3687–3697). Brussels, Belgium: Association for Computational, https://doi.org/10.18653/v1/D18-1404

[3] Tian, H., Wu, H., Wang, H., Wu, F. (2020). *SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis*, https://ar5iv.org/abs/2005.05635

[4] Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, Soujanya Poria (2020). *COSMIC: COmmonSense knowledge for eMotion Identification in Conversations*, https://arxiv.org/pdf/2010.02795.pdf

[5] Soujanya Poria, Navonil Majumder, Rada Mihalcea, Eduard Hovy (2019). *Emotion Recognition in Conversation: Research Challenges, Datasets, and Recent Advances*, https://arxiv.org/pdf/1905.02947.pdf

[6] Shen, W., Chen, J., Quan, X., & Xie, Z. (2021). *DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition*, Proceedings of the AAAI Conference on Artificial Intelligence, 35(15), 13789-13797. https://doi.org/10.1609/aaai.v35i15.17625

[7] Taewoon Kim, Piek Vossen (2021). *EmoBERTa: Speaker-Aware Emotion Recognition in Conversation with RoBERTa*, https://arxiv.org/pdf/2108.12009.pdf

[8] H. Ranganathan, S. Chakraborty and S. Panchanathan,(2021). *Multimodal emotion recognition using deep learning architectures*, 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 2016, pp. 1-9, doi: 10.1109/WACV.2016.7477679.

[9] Laukka, P., & Elfenbein, H. A. (2021). *Cross-Cultural Emotion Recognition and Evaluation in Conversational AI Systems*, Emotion Review, 13(1), 3-11. https://doi.org/10.1177/1754073919897295