Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2023”

June 14–16, 2023

Bimodal sentiment and emotion classification with multi-head attention fusion of acoustic and linguistic information

Anastasia Dvoynikova
St. Petersburg Federal Research Center of the Russian Academy of Sciences / Saint-Petersburg, Russia
dvoynikova.a@iias.spb.su

Alexey Karpov
St. Petersburg Federal Research Center of the Russian Academy of Sciences / Saint-Petersburg, Russia
karpov@iias.spb.su

Abstract

This article describes solutions to couple of problems: CMU-MOSEI database preprocessing to improve data quality and bimodal multitask classification of emotions and sentiments. With the help of experimental studies, representative features for acoustic and linguistic information are identified among pretrained neural networks with Transformer architecture. The most representative features for the analysis of emotions and sentiments are EmotionHuBERT and RoBERTa for audio and text modalities respectively. The article establishes a baseline for bimodal multitask recognition of sentiments and emotions – 63.2% and 61.3%, respectively, measured with macro F-score. Experiments were conducted with different approaches to combining modalities – concatenation and multi-head attention. The most effective architecture of neural network with early concatenation of audio and text modality and late multi-head attention for emotions and sentiments recognition is proposed. The proposed neural network is combined with logistic regression, which achieves 63.5% and 61.4% macro F-score by bimodal (audio and text) multitasking recognition of 3 sentiment classes and 6 emotion binary classes.

Keywords: sentiments; emotions; CMU-MOSEI; attention mechanism; bimodal; multitask
DOI: 10.28995/2075-7182-2023-22-51-61

Бимодальная классификация сентимента и эмоций на основе объединения акустической и лингвистической информации с помощью механизма внимания

Двойникова А. А.
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук / Санкт-Петербург, Россия
dvoynikova.a@iias.spb.su

Карпов А. А.
Санкт-Петербургский Федеральный исследовательский центр Российской академии наук / Санкт-Петербург, Россия
karpov@iias.spb.su

Аннотация

Данная статья посвящена описанию решений нескольких задач: предобработка базы данных CMU-MOSEI для улучшения качества данных и bimodal multitask классификация эмоций и сентимента. С помощью экспериментальных исследований выявляются репрезентативные признаки для акустической и лингвистической информации среди предобученных нейронных сетей с архитектурой Transformer. Наиболее репрезентативными признаками для анализа эмоций и сентимента являются EmotionHuBERT и RoBERTa для аудио и текстовой модальности, соответственно. В статье устанавливается baseline для бимодального многозадачного распознавания сентимента и эмоций – 63,2 % и 61,3 % macro F-score, соответственно. Также проводятся эксперименты с различным подходами к объединению модальностей – конкатенация multi-head attention. Предлагается наиболее эффективная архитектура нейронной сети с ранней конкатенацией аудио и текстовой модальности и позднем multi-head attention для распознавания эмоций и сентимента. Предложенная нейронная сеть объединяется с логистической регрессией, с помощью чего достигается 63,5 % и 61,4 % macro F-score при бимодальном (аудио- и текстовый) многозадачном распознавании 3 классов сентимента и 6 бинарных классов эмоций.

Ключевые слова: сентимент; эмоции; CMU-MOSEI; механизм внимания; бимодальность; многозадачность

1 Introduction

Many existing studies are devoted to recognition of emotions and sentiments, because this area is in demand and there are still many unsolved problems [1-4]. People express emotions through visual, verbal and non-verbal manifestations. Based on this, developing a system for recognizing emotions and sentiments, it is necessary to analyse as many different sources of emotion manifestation as possible (video, audio, text modality) [5]. Based on the specifics of the data and the task, each modality can make a different contribution to the reliability of the system [6]. Therefore, it is important to conduct experimental studies to identify representative modalities for each task.

There are several approaches to emotions and sentiments recognition: emotions and sentiments analyses separately [2, 3, 5] and together (multitask) [4]. Multitasking systems have advantages in summarizing information better and finding correlations between different tasks.

In this article, we solve 2 important tasks: preprocessing the CMU-MOSEI [7] database, and bimodal multitask recognition of emotions and sentiments with different approaches to fusion modalities. A multimodal CMU-MOSEI database was used for experimental studies. This corpus has some problems, such as incorrect timings of speech utterances, extracted subtitles from videos instead of transcriptions of speakers' speech. Therefore, the CMU-MOSEI data corpus has been significantly modified by semi-automatic methods. This was done to improve the quality of the data. However, the experimental studies carried out on the modified data corpus become incomparable with existing studies with this database. Therefore, in this article we are setting a baseline for multitask recognition of multiclass sentiments (3 classes) and multilabel emotions (6 classes) by acoustic and linguistic information of speech utterances.

The article contains the following structure: Section 2 presents an analysis of existing solutions in the field of multimodal and multitask classification of emotions and sentiments on the CMU-MOSEI database. Section 3 describes the CMU-MOSEI data and the data processing algorithm. Section 4 contains experimental studies aimed at identifying relevant acoustic and linguistic features, establishing a baseline and bimodal multitask approach to emotions recognition and sentiments using various methods of merging modalities. Conclusions are given in Section 5.

2 Related Work

All the articles described in this section relate to research with the CMU-MOSEI database. Some researchers [1, 2] suggest later or early fusion of modalities using concatenation. This approach supposes the equal importance of information in each modality. In the real world, the relevance of information from each modality is unbalanced. This is due to presence of noise in the data, equipment malfunctions during recording, etc. More complex approaches to fusion modalities, such as hierarchical [8], attention mechanisms [1-4, 6, 9], allow to get a more reliable automatic system for recognizing emotions and sentiments. The authors of most recent studies [1-3, 10] use multi-head attention (MHA) to combine modalities in the tasks of emotion and sentiments classification. The advantages of MHA are that the algorithm uses several parallel streams of self-attention (head attention). This allows to find more relationships in the information. There are various areas of research: the application of MHA to each modality separately [1], to several modalities at an early stage [1-3], later combining several modalities [1, 6, 10]. [1] compares several ways of combining modalities (video, audio, text): with early concatenation and with early and late MHA. Studies show that various methods with early concatenation show emotion recognition accuracy on average 1-2% lower than methods with MHA. In [2], on the contrary, a simple concatenation for combining modalities (audio text) shows the accuracy of emotion recognition on the CMU-MOSEI corpus by 1% better than with concatenation after MHA for each modality. Experiments in the article [10] prove that later combining of modalities using MTA can improve the accuracy of emotion recognition by 4% than using late concatenation. Despite a large number of experiments in this field, researchers have not been able to come to a conclusion on the most effective ways of combining modalities for recognizing emotions or sentiments. This is due to the different nature of information, modalities and their combinations, neural network architecture, features, etc.

A large number of experimental studies in the field of emotions and sentiments recognition are conducted on the data of the CMU-MOSEI corpus [7]. The authors of many studies [4, 6, 9] used the features provided by the authors of the database – Emotive FACET, OpenFace, COVAREP, Glove for visual, acoustic and linguistic modalities. However, there are works by [1-3] in which the authors used other

features, for example BERT, spectrograms, pre-trained VGG16, ResNet50, etc., and achieved the highest accuracy. Some researchers used video, audio and text modalities simultaneously [1, 4, 6, 9], some bimodal recognition (audio and text) [2, 3]. Using the database, the tasks of emotions recognition are solved [1, 3, 6] and sentiments [9], both separately and simultaneously (multitask approach) [4]. In [4], using video, audio and text modalities and multitask approach, recognition 78.6% and 78.8% F-scores, 62.8% and 80.5% weighted accuracy is achieved for 6 classes of emotions and 2 classes of sentiments, respectively. Research by [11] shows that with an increase in the number of modalities, the accuracy of sentiments recognition on the CMU-MOSEI corpus increases. The article [4] uses a multimodal (video, audio, text) and multitask (using cross-modal attention) approach for recognizing sentiment and emotions on the CMU-MOSEI. It achieves recognition F-score of 78.6% and 78.8% for 6 classes of emotions (multilabel) and 2 classes of sentiments (multiclass), respectively. The presented results can be considered as the baseline of existing studies, since the work is as close as possible to the present research in this article. However, this study uses a modified CMU-MOSEI corpus, so the research results are not comparable to the baseline of existing studies.

3 Multimodal Database

The CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) database was used to conduct experimental studies [7]. The CMU-MOSEI includes video monologues from YouTube. These videos contain only one person in the frame who discusses one topic of interest. The videos were selected by 250 tags such as reviews, debates, business, products, speech, politics, etc. The volume of the corpus is 3,228 videos, the number of unique speakers is 1000. The authors of the corpus manually extracted transcriptions of speech from speakers' monologues (in fact, subtitles from videos). Then 23453 sentences were selected, timestamps of the beginning and end of the phrase were set. After that, each sentence was annotated by sentiment [-3; 3] and 6 basic emotions according to Ekman [12] (joy, sadness, anger, fear, disgust, surprise) on a scale of [0; 3] each. One utterance can contain several emotions. Each phrase was annotated by 3 annotators from a crowdsourcing platform. The annotators were not provided with instructions on how to annotate emotions so that they could interpret "how they feel".

3.1 Data preprocessing

In the original database, some phrases have incorrect timings: the first phrase in the monologue has a negative start time or the timings of phrases do not correspond to the pronunciation time. Such timings were adjusted manually. There were also situations when the timings of two adjacent phrases intersected. Such timings were corrected automatically by calculating the mean between the end of one phrase and the start of the next and subtracting (or adding) the resulting mean from these two timings.

The authors of the CMU-MOSEI provided their own data separation into training, validation and test sets. However, not all files from this distribution have an annotation. Therefore, 2 files from the test set were deleted.

The authors of the database extracted transcriptions of speech from videos. These transcriptions are subtitles from the video and sometimes these subtitles do not match the speaker's speech. In addition, the database contains videos of monologues of a speech-impaired person using gestures, so there is no acoustic and linguistic information. The analysis of speech transcriptions, rather than subtitles, is more correct for the analysis of emotional speech utterances. Therefore, speech transcriptions were extracted from the corrected audio timings using the automatic speech recognition System (ASR) - Vosk[1]. The modified data can be found in [2].

This study solves the problem of multilabel classification. Therefore, the labels of each emotion for each phrase were converted to binary categorization (0 – no emotion, >0 – there is an emotion). Continuous sentiment labels were transformed into 3 category classes: negative (<0), neutral (=0), positive (>0).

[1] https://alphacephei.com/vosk/

[2] https://github.com/Dvoynikova/CMU-MOSEI-modified.git

3.2 Data analysis

After the data preprocessing stage, the volume of the database has changed. Information about the total duration of the audio and the number of sentences in each of the training, validation and test sets is shown in Figure 1. Figures 2 (top) and 2 (bottom) show the distribution of sentiments and emotions in the modified CMU-MOSEI database.

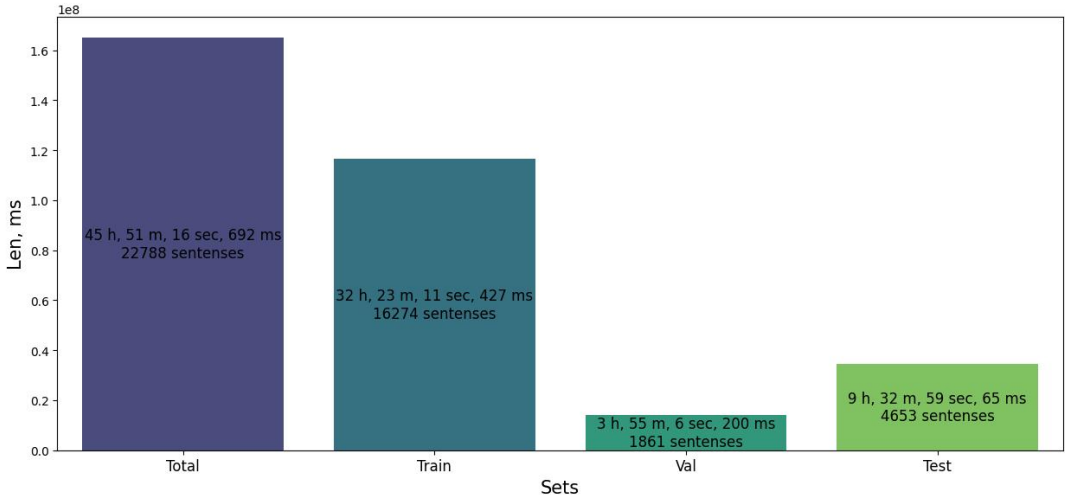


Figure 1: Modified CMU-MOSEI data distribution

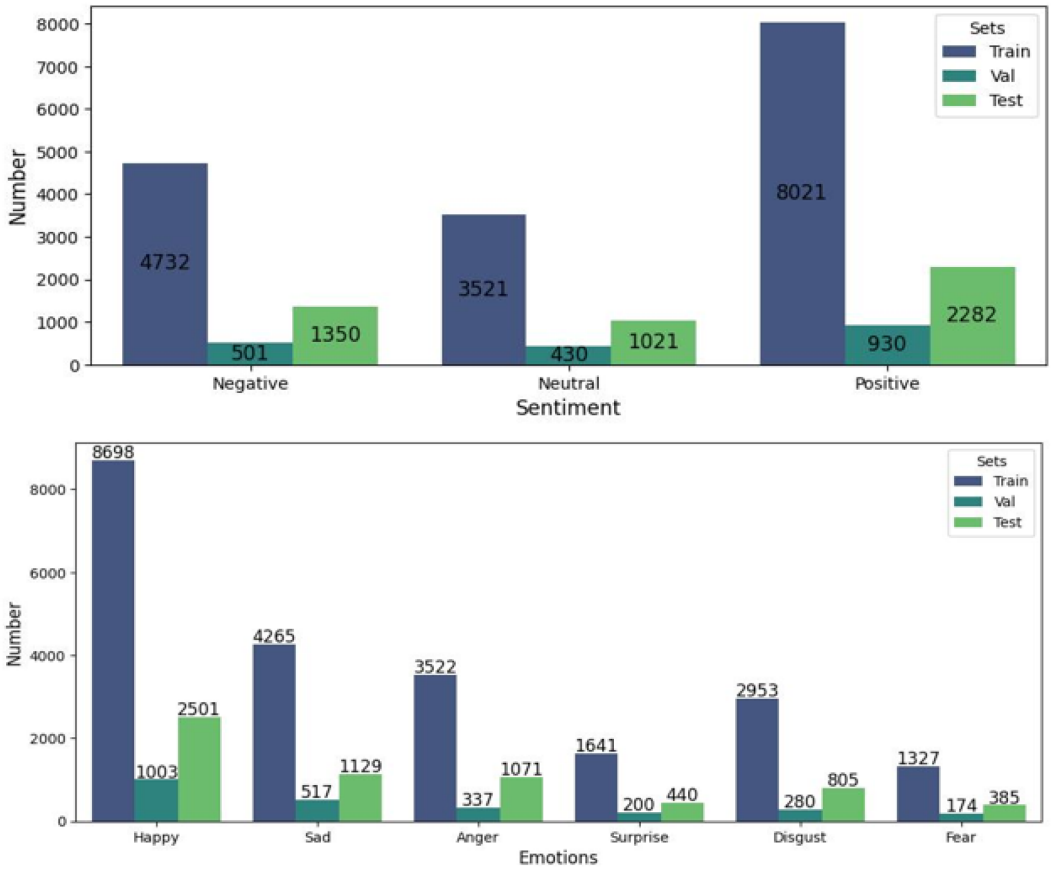


Figure 2: Distribution of sentiments in the modified CMU-MOSEI (top); Distribution of emotions in the modified CMU-MOSEI (bottom)

As mentioned above, each sentence might have several emotion labels. Figure 3 shows the number of pairwise intersections of emotions (on a binary scale) in the statements of the entire database. As can be seen from the heatmap matrix, the emotions of anger, sadness and disgust are most often found to overlap on the statements of the CMU-MOSEI corpus. Also, the labels of some statements may contain both joy and sadness. This raises doubts about the quality of the annotation, because the nature of these emotions has different valences, which makes it difficult to manifest these emotions simultaneously or with a short time interval.

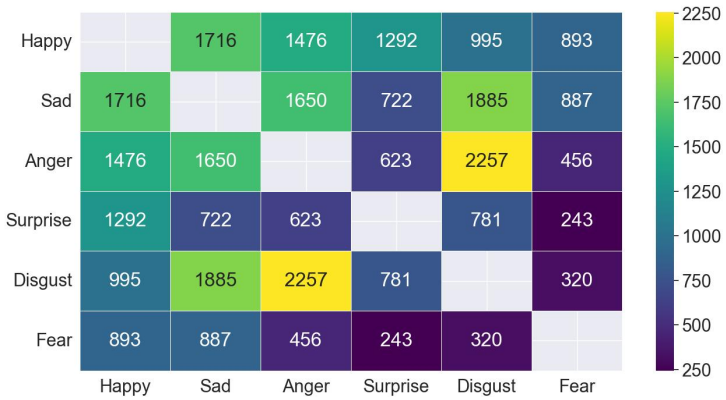


Figure 3: Numbers of intersections of emotions in sentences

Transcriptions of speech were extracted from speech utterances using ASR. Figure 4 shows the word cloud of the resulting linguistic information. It can be seen from the picture that the speakers' speech was free, and one of the most popular topics was the discussion of films. At the same time, it can be noted that words with emotional content are rarely found in statements.

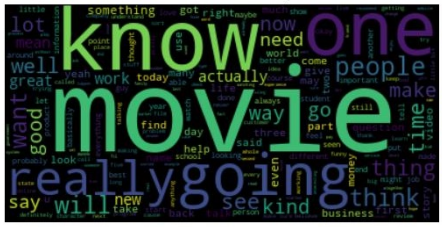


Figure 4: WordCloud ASR transcriptions of the modified CMU-MOSEI

4 Experimental studies

In this study, the bimodal multitask classification problem is solved. The acoustic and linguistic information of speech utterances from the CMU-MOSEI database is analyzed. The developed system simultaneously recognizes 3 classes of sentiments (negative, neutral, positive) – multiclass task, and 6 classes of emotions (joy, sadness, anger, surprise, disgust, fear) – multilabel task.

4.1 Baseline

The first stage of experimental research is the identification of representative features from acoustic and linguistic information. To date, the use of pretrained neural networks based on Transformer is the most effective solution for many tasks [13]. Therefore, we are experimenting with various Transformer architectures, namely HuBERT [14], Wav2Vec [15], EmotionHuBERT [16] for audio and BERT [17], RoBERTa [18], ALBERT [19] – for ASR transcriptions. For each speech utterance, we extract mean and std from a vector of dimension 1 x sequence_lenght x size_vector. As a result, we get vectors of dimension 2048 (for EmotionHuBERT) and 1536 (for the rest of Transformers) for each utterance. Then these vectors are fed to the MultiOutputClassifier with Logistic Regression. The training models with the selection of hyperparameters to identify the best macro F-score on the validation set. The choice of the F-score metric is determined by the unbalanced classes in the database, as well as the importance of the

absence of predictions of false negative results. In addition, the F-score metric is used in many existing research with the CMU-MOSEI database. The use of macro F-score is necessary in order to abstract from unbalanced classes and get a more objective estimation. The results of recognition of 3 classes of sentiments and 6 binary emotions on the text set are shown in Table 1. The Average Emotion in the Table means the average between the macro F-score for each individual emotion.

Features	**Sentiment**	**Average Emotion**
Acoustic features		
HuBERT	50.6	57.9
Wav2Vec	49.8	48.8
EmotionHuBERT	*58.9*	*59.9*
Linguistic features		
BERT	57.7	53.3
ALBERT	57.6	56.7
RoBERTa	*61.9*	*59.2*
Acoustic + Linguistic features		
EmotionHuBERT + RoBERTa	*63.2*	*61.3*

Table 1: Baseline results of sentiments and emotions recognition on MultiOutputClassifier with Logistic Regression, macro F-score, %

As can be seen from Table 1, the most representative features are EmotionHuBERT and RoBERTa for acoustic and linguistic information, respectively. It can also be noted that the text modality is more informative for the analysis of sentiments, and the audio modality is for the analysis of emotions. Combining modalities makes it possible to increase the accuracy of recognition of sentiments and emotions comparing to unimodal classifiers by macro F-score = 1.4% and 2%, respectively. With the help of the conducted experiments, we establish a baseline for recognizing 3 classes of sentiments – 63.2%, and 6 binary emotions – 61.3% macro F-score on the modified CMU-MOSEI corpus. The dummy classifier recognizes sentiments and emotions macro F-score = 21.9% and 43.7%, respectively. Thus, the proposed baseline based on the MultiOutputClassifier with Logistic Regression exceeds the dummy classifier by 41.3% and 17.6% of sentiments and emotions recognition.

4.2 Approaches to modality fusion

At the second stage, we conduct experimental studies with different approaches to combining modalities. We explore different stages of fusion (early and later) and approaches of fusion – concatenation and multi-head attention (MHA). Figure 5 (Neural Network block) shows the best neural network architecture for bimodal recognition of sentiments and emotions.

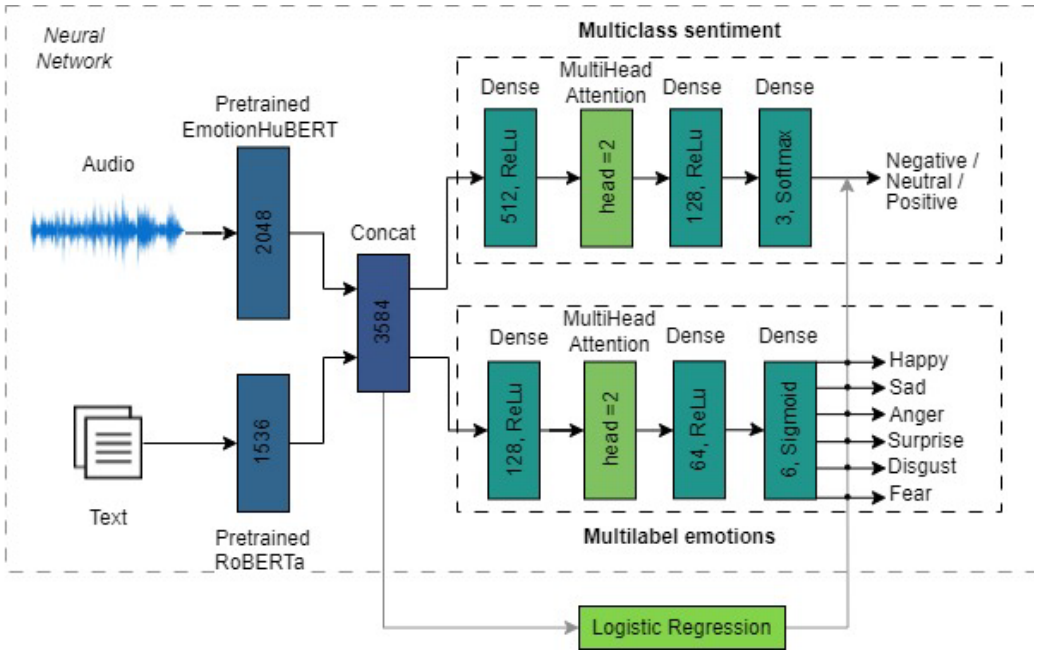
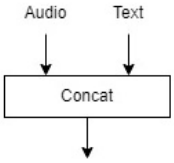


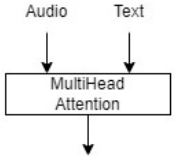
Figure 5: Bimodal multitask system architecture

We use pretrained EmotionHuBERT and RoBERTa to extract acoustic and linguistic features. The proposed MHA blocks contain 1 attention block and 2 attention heads. We use ReLU activation after each fully connected layer, except the last ones. Neural network training takes place with Adam optimizer with a learning rate of 0.01, batch size = 128. For sentiments recognition, categorical cross-entropy loss and softmax activation on the last layer are used, since a speech utterance can have only one sentiments label. Binary cross-entropy loss and sigmoid activation on the last layer are used for emotion recognition, because a speech utterance can have several binary emotion labels. Training takes place at 150 epochs, but it can stop when the loss on validation ceases to decrease during training.

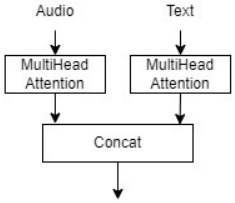
Figure 6 shows various approaches to modality fusion that were used to conduct the experiments. The early fusion of features (Fig. 6 a), b), d), e) f)) allow to immediately analyze information from two modalities [2]. Combining the modalities using MHA (Fig. 6 b), d), f)) allow to highlight the most relevant information among the 2 information flows and focus on the more important one [1]. Later fusion (Fig. 6 c), d)) allows to first highlight the relevant information for each modality, and then at later steps to combine it [10]. Using MHA (Fig. 6 e), f)) for each branch of sentiments and emotions recognition, it helps to highlight only the information that is necessary for a specific task [4].



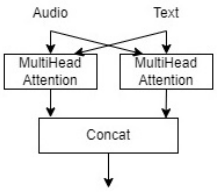
a) Early concatenation



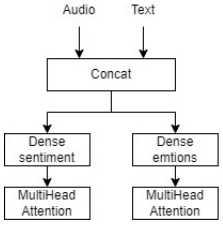
b) Early single MHA



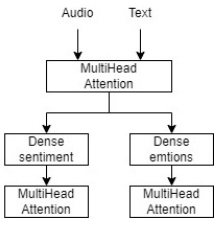
c) Early MHA for each modality



d) Early MHA for each cross-modality



e) Early concatenation with late MHA



f) Early MHA with late MHA

Figure 6: Different approaches to audio and textual modality fusion

Each approach from Figure 6 is applied and trained on the parameters described above. The results of experimental studies and conclusions on them are presented in the discussion section.

5 Experimental Results and Discussion

In Section 4, various approaches to fusion modalities using concatenation and the MHA mechanism at early and late stages are proposed. To determine the effectiveness of combining several modalities, experiments were also conducted with only one modality. The architecture of the system for unimodal multitask sentiments and emotions recognition is similar to the architecture in Figure 5, with the exception of only one modality. Table 2 presents results of experimental studies with the proposed approaches (Figure 5).

Modality fusion	Sentiment	Happy	Sad	Anger	Surprise	Disgust	Fear	Average Emotion
Unimodal Audio	58.8	69.0	62.8	57.3	47.5	66.4	47.8	58.5
Unimodal Text	61.0	64.9	55.1	63.9	47.5	70.7	47.8	58.3
Early concatenation	61.3	67.1	59.6	58.9	52.5	69.7	50.6	59.7
Early single MHA	59.1	65.3	58.4	62.4	47.5	69.3	47.8	58.5
Early MHA for each modality	56.6	68.8	61.9	62.8	47.5	67.0	47.8	59.8
Early MHA for each cross-modality	**63.0**	68.5	59.8	54.4	47.5	69.9	47.8	58.0
Early concatenation with late MHA	61.1	67.1	62.5	66.0	47.5	71.5	50.0	**60.8**
Early MHA with late MHA	61.8	68.5	58.7	55.5	47.5	70.5	47.8	58.1
Early concatenation with late MHA + LR	**63.5**	68.4	61.7	62.0	53.8	68.8	53.5	**61.4**

Table 2: Results of various fusion of modalities, macro F-score, %

Based on the results in Table 2, it can be concluded that combining the modalities makes it possible to obtain a more robust system, as well as increase in accuracy of recognizing sentiments and emotions by an average of 1-2% compared to unimodal systems. Also from the results obtained we can say, that MHA cannot be unambiguously called an effective approach to fusion modalities. For example, in our experiments, early concatenation showed a higher recognition result of emotions and sentiments, than a simple early fusion using MHA. But in general, the mechanism of attention in most cases allows to achieve higher accuracy. Also, it can be noted that applying a cross MHA to each modality is the most effective approach with respect to modality fusion using the Attention block. Using the late MHA for each task (sentiments and emotions) separately allows to achieve the highest recognition results.

The results in the Table 2 show that there is no unambiguously better approach among neural networks for recognizing emotions and sentiments at the same time. Early MHA for each cross-modality approach allowed to achieve maximum accuracy of sentiments recognition (grade 3) 63.0% macro F-score. At the same time, emotion recognition (6 binary classes) with a maximum average macro F-score of 60.8% was obtained using the early concatenation with late MHA approach.

It is also worth noting that the recognition of emotions such as surprise and fear occurs quite poorly with all the approaches considered. This may be due to the complex nature of the origin of emotions, as well as the fact that these emotions have the most unbalanced classes relative to other emotions. The emotion of disgust is always recognized better, than other emotions with all the approaches considered. Because this emotion has vivid manifestations in acoustic characteristics, as well as specific antropophones, which can manifest themselves in a linguistic modality.

To choose the most effective approach among neural networks to the recognition of emotions and sentiments, it is necessary to analyze the accuracy of recognition of each emotion separately. The disadvantage of the early MHA for each cross-modality approach is that it does not recognize emotions such as anger, surprise and fear well. However, the emotion of anger is important in the analysis of emotions, because an angry person can pose a danger to others. Therefore, it is important that automatic

systems recognize the emotion of anger as best as possible. Among neural networks, the most effective approach is Early connect with late Attention (it is shown in Figure 5 of the Neural Network block). It's disadvantage is the lower accuracy of sentiments recognition relative to other approaches. The advantage of this approach is that it predicts all emotions more reliably. However, this approach does not exceed the baseline from section 4.1.

The most effective approach for the task of bimodal multitasking of emotions and sentiments is an approach based on the neural network Early concat with late Attention and Logistic regression (Figure 5). Concatenated acoustic and linguistic features are fed to the input of the neural network and logistic regression. The two models are combined at the decision-making level. The proposed approach allows achieving 63.5% and 61.4% macro F-score for recognizing 3 sentiments classes and 6 binary emotions classes, respectively. The obtained results are 0.3% and 0.1% higher than the established baseline with Logistic regression. More detailed results of emotion and sentiments recognition using the proposed approach are presented in Table 3.

Metrics	Sentiment	Happy	Sad	Anger	Surprise	Disgust	Fear	Average Emotion
Macro precision	63.5	68.4	61.7	61.6	54.9	67.1	55.2	61.5
Macro recall	63.8	68.8	65.2	64.9	61.3	74.9	64.0	66.5
Macro F-score	63.5	68.4	61.7	62.0	53.8	68.8	53.5	61.4

Table 3: Results of emotions and sentiments recognition by early concatenation with the late Attention approach

When recognizing emotions, the multilabel task was solved, in other words, there could be several emotion labels in one phrase. Thus, it is impossible to analyze classifier errors. Sentiments recognition occurred in the multiclass task, i.e. there could be only one sentiments label in one phrase. The matrix of errors in the recognition of the sensor is shown in Figure 7.

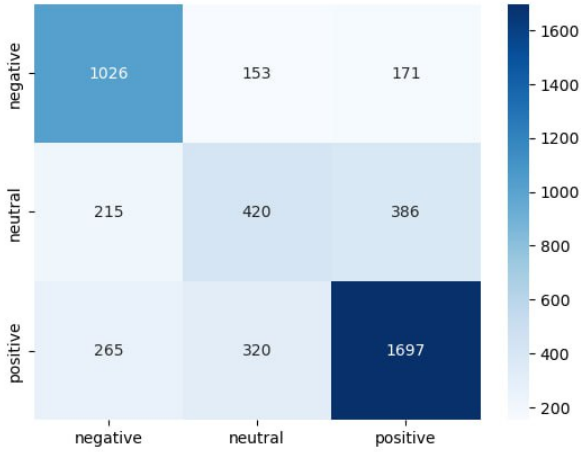


Figure 7: Matrix of sentiments recognition errors

As can be seen from Figure 7, the negative and positive classes of sentiments are recognized quite well. The neutral class is most often confused with the positive one.

As mentioned in the Related work section, all existing research in the field of emotion and sentiment recognition using the CMU-MOSEI database was conducted on the original data set. We conducted experiments on a modified data set, which makes our results completely incomparable with other works. The best accuracy of multitask recognition of emotions (6 binary classes) and sentiments (2 classes) using modal analysis (video, audio, text) F-score 78.6% and 78.8% Fare achieved in [4]. Our results (63.5% of emotions and 61.4% of sentiments) are lower than the existing ones. It is worth noting that we recognize 3 classes of sentiments when other studies carry out the recognition of 2 classes of sentiments. In addition, we only analyze audio and textual modality, while another work analyzes 3 modalities. It is also worth saying that only the baseline is set in this article, the improvement of this baseline is planned in our next studies.

6 Conclusions

The article is devoted to a bimodal multitask approach to the recognition of emotions and sentiments. The CMU-MOSEI database was used as data for experimental studies. One of the main tasks that we solved in this study is the semi-automatic preprocessing of CMU-MOSEI in order to improve data quality. We also identified representative features for acoustic and linguistic information – EmotionHuBERT and RoBERTa, respectively. Using these features, we have established a baseline for bimodal multitasking recognition of emotions and sentiments – 63.2% and 63.3% macro F-score, respectively.

In our study, we conducted experiments to identify the most effective approach to fusion (concatenation and multi-head attention) audio and text modality for the multitask of recognizing sentiments and emotions. Based on the results of the experiment, we can conclude that the use of early concatenation of acoustic and linguistic information and MHA for each task (sentiments and emotions) separately allows achieving the highest recognition results. Using EmotionHuBERT and RoBERTa as features and the MHA mechanism for each task, we achieve 61.1% and 60.8% macro F-score for a bimodal (audio and text) multitask approach to recognize 3 sentiments classes and 6 binary emotion classes.

The proposed bimodal (audio and text) approach using the Early concat with late Attention neural network and Logistic regression makes it possible to achieve 63.5% and 61.4% macro F-score for recognizing sentiments and emotions.

The prospect of further research may manifest itself in the addition of a video modality. Facial expressions, gestures, postures are also representative information in the manifestation of emotions. Therefore, it is assumed that the analysis of the video modality will help to increase the accuracy of the recognition of sentiments and emotions.

Acknowledgements

This research was supported by the Russian Science Foundation (project No. 22-11-00321, https://rscf.ru/project/22-11-00321/).

References

[1] Lee S., Han D.K., Ko H. (2021), Multimodal emotion recognition fusion analysis adapting BERT with heterogeneous feature unification, IEEE Access, Vol. 9, pp. 94557-94572.

[2] Siriwardhana S., Reis A., Weerasekera R, Nanayakkara S. (2020), Jointly fine-tuning" bert-like" self supervised models to improve multimodal speech emotion recognition, arXiv preprint arXiv:2008.06682.

[3] Ho N.H., Yang H.J., Kim S.H. (2020), Multimodal approach of speech emotion recognition using multi-level multi-head fusion attention-based recurrent neural network, IEEE Access. Vol. 8, pp. 61672-61686.

[4] Akhtar M.S., Chauhan D.S., Ghosal D., Poria S., Ekbal A., Bhattacharyya P. (2019), Multi-task learning for multi-modal emotion recognition and sentiment analysis, arXiv preprint arXiv:1905.05812.

[5] Verkholyak O., Dvoynikova A., Karpov A. (2021), A Bimodal Approach for Speech Emotion Recognition using Audio and Text, Journal of Internet Services and Information Security. Vol. 11, no. 1, pp. 80-96.

[6] Mittal T., hattacharya U., Chandra R., Bera A., Manocha D. (2020), M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues, Proceedings of the AAAI conference on artificial intelligence. Vol. 34, no. 02, pp. 1359-1367.

[7] Zadeh A.B., Liang P.P., Poria S., Cambria E., Morency L.P. (2018), Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph, Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Vol. 1, pp. 2236-2246.

[8] Velichko A., Markitantov M., Kaya H., Karpov A. (2022), Complex Paralinguistic Analysis of Speech: Predicting Gender, Emotions and Deception in a Hierarchical Framework, Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, pp. 4735-4739.

[9] Kumar A., Vepa J. (2020), Gated mechanism for attention based multi modal sentiment analysis, IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4477-4481.

[10] Choi W. Y., Song K. Y., Lee C. W. (2018), Convolutional attention networks for multimodal emotion recognition from speech and text data, Proceedings of grand challenge and workshop on human multimodal language, pp. 28-34.

[11] Ghosal D., Akhtar M. S., Chauhan D., Poria S., Ekbal A., Bhattacharyya B. (2018), Contextual inter-modal attention for multi-modal sentiment analysis, Proceedings of the 2018 conference on empirical methods in natural language processing, pp. 3454-3466.

[12] Ekman P., Freisen W.V., Ancoli S. (1980), Facial signs of emotional experience, Journal of personality and social psychology. Vol. 39, no. 6, pp. 1125.
[13] Alibaeva K., Loukachevitch N. (2022), Analyzing COVID-related Stance and Arguments using BERT-based Natural Language Inference, Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference “Dialog 2006”. Pp 8-17.
[14] Hsu W.N., Bolte B., Tsai Y.-H.H., et al. (2021), Hubert: Self-supervised speech representation learning by masked prediction of hidden units, IEEE/ACM Transactions on Audio, Speech, and Language Processing. Vol. 29, pp. 3451–3460.
[15] Baevski A., Zhou Y., Mohamed A., Auli M. (2020), Wav2vec 2.0: A framework for self-supervised learning of speech representations, Advances in neural information processing systems. Vol. 33, pp. 12449-12460.
[16] Wagner J., Triantafyllopoulos A., Wierstorf H., Schmitt M., Burkhardt F., Eyben F., Schuller B.W. (2022), Dawn of the transformer era in speech emotion recognition: closing the valence gap, arXiv preprint arXiv:2203.07378.
[17] Devlin J., Chang M., Lee K., Toutanova K. (2018), Bert: Pre-training of deep bidirectional transformers for language understanding,arXiv preprint arXiv:1810.04805.
[18] Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., Stoyanov V. (2019), Roberta: A robustly optimized bert pretraining approach,arXiv preprint arXiv:1907.11692.
[19] Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., Soricut R. (2019), Albert: A lite bert for self-supervised learning of language representations, arXiv preprint arXiv:1909.11942.