”Lviv Polytechnic” National University
Institute of Computer Sciences and Information Technologies
Department of artificial intelligence systems
Specialty 122 «Computer Science»

EXPLANATORY NOTE

to the bachelor's qualification work on the topic

«**Speech emotion recognition using machine learning methods**»

Student(s) of the group	KN-417	Dolynska K. I.
Supervisor of work	_____ *(signature)*	(Syvokon O. O.)
Consultants	_____ *(signature)*	(_____)
	_____ *(signature)*	(_____)
Head department of the AIS	_____	(Melnykova N. I.)

Lviv - 2025

”Lviv Polytechnic” National University
Institute of Computer Sciences and Information Technologies
Department of artificial intelligence systems
Specialty 122 « Computer Science »

I CONFIRM:

Head department of the AIS _____

“____” _____2023

TASK

FOR THE BACHELOR'S QUALIFICATION WORK FOR STUDENT

Khrystyna Dolynska

1. The topic of the work: «Speech emotion recognition using machine learning methods» is approved by the order of the university № 902-4-08 від 12.03.2025

2. Deadline for submission of completed work: _06.06.2025 _____

3. Initial data for (project) work: RAVDESS, TESS, MC-EIU, MELD, GoEmotions datasets; audio-to-text model Whisper; transformers superb/hubert-large-superb-er, cardiffnlp/twitter-roberta-base-emotion and bhadresh-savani/distilbert-base-uncased-emotion.

4. Content of the settlement and explanatory note: (list of issues to be developed): analytical review of literary and other sources, systematic analysis and justification of the problem, methods and means of solving the problem, practical implementation and experimental evaluation.

5. List of graphic material (with exact indication of mandatory drawings): 12 images, 15 charts and 11 tables.

6. Consultants for work, indicating the sections of the project that concern them

Chapter	Consultant	Signature, date	
		Task is issued	Task is accepted
Foreign language consultant			

7. Issue date of the task 16.09.2024

Supervisor of work _____
(signature)

Task was accepted to execution by _____
(signature)

CALENDAR OF PLAN

№	The name of the stages of the thesis	The term of execution of work stages	Note
1	Literature review	16.09.2024 – 28.10.2024	done
2	Creation of architecture	29.10.2024 – 16.12.2024	done
3	Algorithmic and software development	17.12.2024 – 10.03.2025	done
4	Experimental researches	11.03.2025 – 29.04.2025	done
5	Drawing up a note	30.04.2025 – 06.06.2025	done

Supervisor of work _____
(signature)

Student _____
(signature)

ABSTRACT

Bachelor’s degree work of the student of the group KN-417 Dolynska Khrystyna Ihorivna. The topic is "Speech emotion recognition using machine learning methods". The work is aimed at obtaining a bachelor's degree in 122 "Computer Science".

The object of research of this work is the processes of recognizing and interpreting emotions in human speech. This process arises from the interaction of various aspects of speech activity, such as intonation, tempo, expressiveness, and the use of language.

The subject of the study are methods and tools that allow us to implement the process of recognizing emotions in human speech or predicting the most appropriate emotions for a given speech context.

This goal was achieved by developing a multimodal system that combines emotion classification based on acoustic voice characteristics and analysis of the semantic content of the text. For this purpose, machine learning methods were applied, including deep neural networks and transformers, which increased the recognition accuracy and improved the analysis of the speaker's emotional state with the context of the statement. The developed system was tested based on publicly available datasets and showed high efficiency in emotion classification tasks.

As a result of the work, a software package has been created that allows automating the process of recognizing emotions in speech and interpreting them considering the textual context. The obtained results can be used as a basis for further research in the field of emotional intelligence, as well as for practical applications in the fields of education, psychology, medical diagnostics and intelligent interaction systems.

Structure and scope of work: The thesis consists of an introduction, three chapters, conclusions, a list of references comprising 38 sources, and three appendices. The main text is presented on 65 pages and includes 27 figures and 11 tables.

Keywords: modality, machine learning, classification, algorithm, method, system, information technology.

CONTENT

ABSTRACT **4**
CONTENT **5**
LIST OF ABBREVIATIONS **7**
INTRODUCTION **8**
1. ANALYTICAL SECTION **11**
1.1. Relevance and current state of research 11
1.1.1. Relevance of the topic in terms of science 11
1.1.2. Overview of existing solutions and analogs 12
1.2. Problems and challenges of automatic emotion recognition 13
1.3. Speech emotion recognition 13
1.3.1. Specifics and challenges of voice emotion recognition 13
1.3.2. Methodological approaches to recognizing emotions in the voice . 15
1.4. Recognizing emotions from the text 19
1.4.1. Specifics and challenges of text emotion recognition 19
1.4.2. Methodological approaches to recognizing emotions from text 19
1.5. Conclusions 22
2. RESEARCH SECTION **24**
2.1. Data requirements 24
2.2. Selected datasets 24
2.2.1. RAVDESS 24
2.2.2. TESS 27
2.2.3. MC-EIU 28
2.2.4. MELD 29
2.2.5. GoEmotions 31
2.2.6. Comparison of datasets 33
2.3. Requirements for the model 33
2.4. Selected approaches 34
2.4.1. CNN 34
2.4.2. HuBERT 35
2.4.3. DistilBERT 37
2.4.4. RoBERTa-base 38

2.5. Speech emotion recognition models. Implementation. 39
2.5.1. CNN 39
2.5.2. Transformers (HuBERT) 42
2.6. Text emotion classification models. Implementation. 46
2.6.1. Text pre-processing 46
2.6.2. Model training 47
2.7. Conclusions 48
3. APPROVALS AND RESULTS SECTION 50
3.1. Speech emotion recognition models. Results overview. 50
3.1.1. CNN 50
3.1.2. Transformer models (HuBERT) 52
3.1.3. Comparison of results 56
3.2. Text emotion classification models. Results overview. 57
3.2.1. Implemented models results comparison 57
3.3. Combining results 64
3.4. Architecture and logic of the system as a whole 65
3.4.1. System structure 65
3.4.2. Interaction between modules 65
3.4.3. Overview of the system interface 68
3.5. Hardware and software requirements 68
3.6. Conclusions 68
CONCLUSIONS 71
REFERENCES 73
Appendix A. 78
Appendix B. 79
Appendix C. 82

LIST OF ABBREVIATIONS

SER	-	Speech Emotion Recognition
TEC	-	Text Emotion Classification
NLP	-	Natural Language Processing
NN	-	Neural Networks
GMM	-	Gaussian Mixture Model
K-NN	-	K-Nearest Neighbor
HMM	-	Hidden Markov Model
SVM	-	Support Vector Machine
MLP	-	Multilayer Perceptron
CNN	-	Convolutional Neural Network
DCNN	-	Deep Convolutional Neural Network
RNN	-	Recurrent Neural Network
ResNet	-	Residual Network
RAVDESS	-	Ryerson Audio-Visual Database of Emotional Speech and Song
IEMOCAP	-	Interactive Emotional Dyadic Motion Capture
CHEAVD	-	Combined Human and Audiovisual Emotion Database
BES	-	Berlin Emotional Speech
PES	-	Polish Emotional Speech
SES	-	Spanish Emotional Speech
OAA	-	One-Against-All
GD	-	Gender Dependent
MFCC	-	Mel-Frequency Cepstral Coefficients
LPCC	-	Linear Predictive Cepstral Coefficients
STFT	-	Short-Time Fourier Transform
CWT	-	Continuous Wavelet Transform
DA	-	Data Augmentation
RCS	-	Random Cropping and Scaling
WGN	-	White Gaussian Noise

INTRODUCTION

Speech emotion recognition has become a critical research topic throughout the previous decade. The main difficulty exists in programming computers to detect both semantic content and emotional inflections in spoken words.

Research on speech emotions remains relevant because emotional intelligence has become vital in modern society and because this field addresses problems. The main obstacle in this challenge involves creating systems that can detect and recognize emotions within human speech. The analysis of speech emotions serves three essential purposes including communication enhancement and psychological research advancement and interface development for computer and technology interactions.

The field has gained increased interest because of both emotional intelligence development trends and practical requirements that span education and public speaking and voice assistant interactions. The rise of scientific publications alongside new platforms that analyze emotional speech content demonstrates growing interest in this topic from academic sector.

Recent advancements have highlighted the potential of speech emotion recognition in healthcare and psychological research. For instance, studies have explored its application in monitoring mental health conditions such as depression and anxiety by analyzing vocal patterns [27]. Additionally, systems have been developed to assist clinicians in assessing patients' emotional states during consultations, aiming to enhance therapeutic communication [29].

Most current systems either detect speaker emotions or offer basic speech characteristics but fail to deliver advanced emotional state detection. The development of complete solutions to enhance emotional expressiveness in speech remains either unavailable or non-functional. This research advocates creating a system which detects emotional speech tones while offering specific recommendations to enhance them.

This relevance determines the purpose of the work:

The goal of this study is to design a speech emotion recognition system based on machine learning techniques and to integrate its outputs with a system for analyzing the semantic content of spoken utterances. This multimodal approach allows for both

the detection of the speaker's emotional state and an evaluation of its consistency with the underlying semantic message, thereby opening new possibilities for enhancing the emotional expressiveness of speech.

To achieve the purpose of the study, the following **objectives** are proposed to be solved:

- Investigate and analyze various approaches to developing systems for analyzing emotions in speech and text;
- Analyze existing datasets;
- Implement a model for emotion classification based on acoustic features of the voice;
- Implement a model for predicting semantically coherent emotions;
- Develop a method for combining the outputs of both models;
- Implement a system based on the integration method combining audio- and text-based emotion classification models.

The object of the study is the processes of recognizing and interpreting emotions in human speech. This process arises from the interaction of various aspects of speech activity, such as intonation, tempo, expressiveness, and the use of language tools.

The subject of the study are methods and tools that enable the implementation of the process of recognizing emotions in human speech, as well as predicting the most appropriate emotions for a given speech context.

The social significance of the study lies in the possibility of practical implementation of emotion recognition systems in speech in such areas as education, psychological counseling, medical diagnostics, public speaking, as well as in the development of intelligent technologies capable of emotionally oriented interaction with users.

The practical significance of the study is as follows: the development of a system for recognizing emotions in speech, as well as the implementation of a system for selecting the most appropriate emotions according to the semantic context, which can be combined and used in a variety of applications. In addition, the results obtained

can be used for further research in the field of NLP, studying the impact of emotional coloring on speech perception, and developing new methods and algorithms for recognizing and analyzing emotions in speech.

The research methods are machine learning algorithms for analyzing voice and text data, methods of automatic speech recognition, as well as techniques for combining information from several modalities.

Approbation of work results. Based on the research conducted, the abstracts were published at the 82nd Student Scientific Conference. Also, a scientific article is being prepared on the topic of the bachelor's thesis.

1. ANALYTICAL SECTION

1.1. Relevance and current state of research

In recent decades, tasks related to emotion recognition have garnered growing attention within the scientific community. There has been active development in the analysis of both textual and vocal data aimed at identifying a person's emotional state. This interest is largely driven by a wide range of practical applications – from analyzing social media messages to monitoring psychological well-being during phone conversations [30] or public speaking engagements.

The publication trends in this domain show a consistent increase in the number of studies over the past 25 years, underscoring both the relevance of the topic and its potential for future advancements. This section explores the growing scientific interest in the field and provides an overview of existing applied solutions that partially address these challenges.

1.1.1. Relevance of the topic in terms of science

Emotion recognition from voice and text are interrelated but technically and conceptually different areas that have been actively researched in recent years. The analysis has shown that since the 2000s, the number of publications on these topics has been growing annually, with especially high intensity in the last five years (2020-2025). This is also confirmed by the graphs (Figs. 1.1 and 1.2), which reflect the activity of scientific publications in these areas.

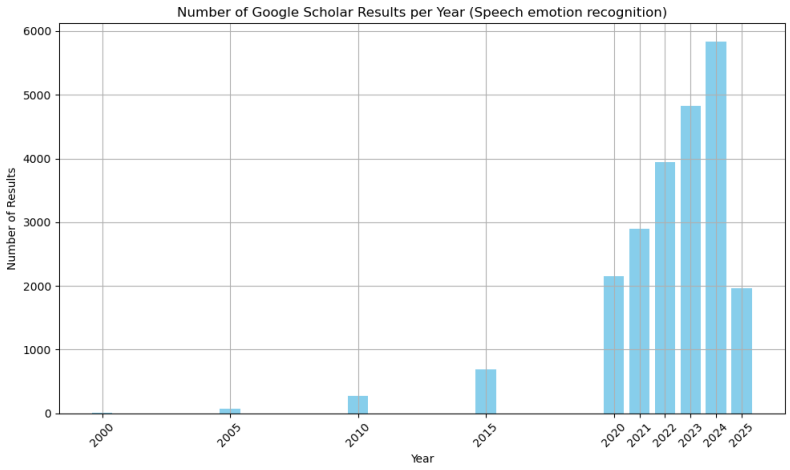

Fig. 1.1. The number of studies on the topic "Voice emotion recognition" in certain years from 2000 to 2025

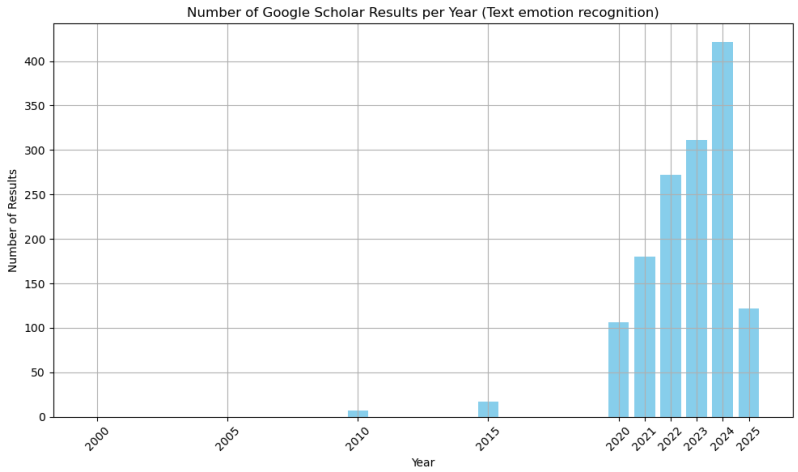

Fig. 1.2. Number of studies on the topic "Text emotion recognition" in certain years from 2000 to 2025

Despite the similar nature of the tasks, the studies often have different practical orientations: emotions in text are mostly analyzed in the context of social media or user content, while voice emotion recognition is focused on determining the speaker's psychological state [30] or the quality of his or her oral performance. This indicates a wide variety of applications of the respective technologies and emphasizes the need for specialized solutions for each of the modalities.

1.1.2. Overview of existing solutions and analogs

Market analysis shows that, despite the active development of voice emotion recognition technologies, most existing applications are focused on determining the speaker's psychological state, mood monitoring, or other auxiliary tasks. At the same time, there are very few solutions that comprehensively analyze the emotionality of speech in order to improve it. In this paper, we consider such systems as potential analogs, since the goal is not only to recognize emotions but also to provide the user with clear recommendations for improving the emotional expressiveness of his or her speech. Among the analogs found, it is worth highlighting the “Oration Master” application, which is focused on speech analysis and the overall improvement of public speaking. This system allows you to get an assessment of the emotionality of the statement, but does not provide specific recommendations for its improvement, and the process of user development takes place through the completion of courses.

In contrast to this approach, the system to be developed in this paper focuses on

in-depth analysis of the emotional coloring of a particular speech. It will be able to formulate point recommendations - what exactly and where needs to be improved - without the need to take training modules that do not guarantee a quick change in the quality of the current speech. Thus, the proposed solution has the potential to become a more effective tool for improving the emotional expressiveness of speech.

1.2. Problems and challenges of automatic emotion recognition

Automatic emotion recognition is an interdisciplinary field that combines elements of computer science, psychology, linguistics, and neuroscience. Systems of this type can work with different types of input data, each of which has its own characteristics, advantages, and limitations. The most common sources of information for emotional analysis are:

- Audio - detection of intonational signs of emotional state;
- Text - analysis of semantic and lexical content of speech;
- Video - recognition of facial expressions, gestures, and body language;
- Physiological signals (ECG, heart rate changes) - emotional reactions.

Despite the wide range of data types, automatic emotion recognition faces several common problems:

1. Subjectivity of emotions: emotions do not always have clear boundaries and can manifest themselves in different ways in different people;
2. Mixed emotions: several emotions can be manifested simultaneously in one statement or action;
3. Contextual influence: the interpretation of emotions often depends on the situation, culture, or intonation.

In addition, each type of data has its own specific challenges. In this paper, we focus on audio and text data, which are among the most common and convenient for practical applications, in the field of human speech quality improvement.

1.3. Speech emotion recognition

1.3.1. Specifics and challenges of voice emotion recognition

For a successful Speech Emotion Recognition (SER) system to work, three key

issues need to be addressed: selecting a good emotional speech database, extracting effective features, and developing reliable classifiers using machine learning methods [13].

When choosing a suitable dataset, the following criteria should be taken into account: the degree of naturalness of emotions, the size of the database, and the number of available emotions [16]. There are three main types of datasets:

- acted emotions: emotions reproduced in certain scenarios, for example, during an acting game;
- evoked emotions: emotions that arise in artificially created situations, such as reactions to music, videos, advertisements, etc.;
- natural spontaneous emotions: emotions that arise in everyday life, such as reactions to real-life events or reality TV shows.

Since the main task of this work is to train a model to detect the emotion of a person's voice during speech, the most appropriate option for training the model is to use a database with natural spontaneous emotions. Such data better reflects real expressions of emotions in speech, which will allow the model to better recognize and classify emotions in a person's voice during speech. However, the disadvantage of such recordings is that they can be distorted by background noise [12] and contain unbalanced emotional categories [16].

The next problem that needs to be solved for the system to work is the selection of effective features. Since a typical set of meaningful emotions consists of 300 different emotional states, this makes it difficult to classify them [1]. As noted in the article "Survey on Speech Emotion Recognition: Features, Classification Schemes, and Databases", this problem can be solved "[...] according to the 'Palette Theory': any emotion can be decomposed into primary emotions, just as any color is a combination of several primary colors." [4]. These primary emotions include neutral, joy, anger, disgust, fear, sadness, stress, and surprise [4].

The last challenge for implementing the SER system is the development of reliable classifiers. Numerous approaches to solving this problem have been considered in various scientific papers, including both classical statistical methods and modern deep neural networks.

1.3.2. Methodological approaches to recognizing emotions in the voice

One example of the use of classical methods is the study "Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K-Nearest Neighbor (K-NN) Techniques" [15], which describes the GMM and K-NN methods. The authors obtained two models, each of which classified certain emotions better than the other. For example, the GMM accurately recognized the emotions of anger (92%), sadness (89%), and neutral emotions (73%), while the K-NN model better recognized the emotion of happiness (90%). The authors concluded that combining the two models would improve accuracy, as each performs better at detecting different emotions.

The article "A Two-Stage Hierarchical Bilingual Emotion Recognition System Using a Hidden Markov Model and Neural Networks" [6] discusses the HMM. The authors used BES and PES datasets to train the model. The aim of their study was to prove the necessity of language knowledge to determine the type of emotion. They considered two emotion classification scenarios: combining all sentences from both datasets into one dataset and using a hierarchical approach where the language was identified first. The results of the experiments are 57.64% accuracy of emotion classification for the first approach and 93.06% for the second. Using the second approach, the authors increased the set of emotions to 6, which led to a partial loss of accuracy (89.13%), but allowed them to classify more emotions.

Another method for recognizing emotions is described in the article "Speech Emotion Recognition using Support Vector Machine" [18]. The paper considered two classification strategies using the SVM method: OAA and GD; MFCC and LPCC algorithms were used to extract features from speech utterances; two datasets were used to train the models: LDC and UGA. When analyzing the results obtained, it was found that feature detection using MFCC provided higher accuracy compared to LPCC (85.085% and 73.125%, respectively); the use of the LDC dataset increased the classification accuracy relative to the UGA dataset (90.08% and 65.97% for the training data, respectively); the accuracy of the GD classifier was higher than the OAA classifier (84.42% and 72.785%, respectively). Thus, the best combination was the use

of the GD classifier together with the MFCC algorithm and the LDC dataset.

The article "Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier" [2] describes another method of emotion recognition: MLP. The authors have complicated the model by increasing the size of the hidden layer (the input layer size was 100, and the size of the hidden layer was 750x750x750). The other training parameters the authors kept at a low level, which resulted in fast training of the model that took several minutes.

The following methods of emotion classification are used in the article "Speech Emotion Recognition Using Deep Convolutional Neural Networks Improved by the Fast Continuous Wavelet Transform" [38]. Two DCNN models were implemented and analyzed in this paper: STFT and fCWT; for DA, RCS, WGN, and no DA techniques were considered; two datasets were used to train the models: eNTERFACE05 and EMO-DB. As a result of the research, it turned out that the best combination for the emotion classification model is the fCWT classifier, the RCS5 data augmentation technique, and the eNTERFACE05 dataset with an achieved accuracy of 76.6%.

Another method is discussed in the article "Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters" [37]. The authors consider the ResNet20 model, using the VoxCeleb2 dataset for pre-training and IEMOCAP for training. To improve the results after training the model, the authors replaced the FC layer of the network and tuned the entire network to the target emotion corpus. Since the result was not better, the authors suggested that the number of parameters for training was too large for a small amount of data.

In the article "Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention" [19], the authors considered various RNN architectures. The best among them was the weighed pool with attention model, whose accuracy was 63.5%. One of the advantages of the resulting model was that it produces a result in the form of a feature vector at the utterance level. This means that the model does not identify a specific emotion but analyzes the entire sentence (or audio sequence) and classifies the full range of emotions used.

To visualize the advantages and disadvantages of each of the methods under consideration, let's draw a table:

Table 1.1. Analysis of emotion recognition methods described in the reviewed scientific papers

Method	**Source**	**Dataset**	**A set of emotions**	**Accuracy**	**Advantages**	**Disadvantages**
GMM	[15]	BES	6	66%	- distinguishes between the emotions of anger and sadness with high accuracy;	- poorly distinguishes between other emotions in the set; - as the set of emotions increases, the training time increases;
K-NN	[15]	BES	6	51,67%	- distinguishes the emotions of happiness and anger; - trains quickly;	- poorly distinguishes between other emotions in the set;
HMM	[6]	BES, PES	4	93,06%	- classifies emotions with high accuracy;	- distinguishes between a small set of emotions; - training the model is time-consuming;
HMM	[6]	BES, PES	6	89,13%	- classifies emotions with a high accuracy;	- model training is very time-consuming;
SVM	[18]	LDC, UGA	4	82,94%	- distinguishes between a set of emotions with almost equal accuracy;	- distinguishes between a small set of emotions;

MLP	[2]	RAVDESS	8	81%	- categorizes many emotions; - trains quickly;	- classification accuracy needs to be improved;
DCNN	[38]	eNTERFACE05	7	76,6%	- categorizes an impressive set of emotions; - CWT allows capturing both temporal and spectral features;	- the model requires a huge amount of data, which in turn makes it impossible to classify emotions in real time; - low accuracy of the model;
ResNet	[37]	VoxCeleb2, IEMOCAP	9	72,73%	- classifies a large set of emotions; - the model does not need to be trained again when adding new domains;	- low model accuracy;
RNN	[19]	IEMOCAP	4	63,5%	- the model can highlight relevant features of an audio track and ignore irrelevant ones, which allows preserving the dynamics of speech and understanding its connections;	- distinguishes between a small set of emotions; - low accuracy of the model;

Considering the advantages and limitations of the methods reviewed, it becomes evident that a universal solution capable of classifying a wide range of emotions with consistently high accuracy has not yet been achieved. While some approaches demonstrate strong performance for specific emotions or offer rapid training, they often fall short when applied to broader emotion sets or real-time scenarios. This highlights both a scientific and practical challenge, which is addressed in detail in Sections 2 and 3 of this work.

1.4. Recognizing emotions from the text

1.4.1. Specifics and challenges of text emotion recognition

Recognizing emotions from text is a challenging task in NLP, which consists in determining the emotional state of the author of the text. The main challenges of TEC are word ambiguity, contextual dependence, sarcasm, irony, and cultural peculiarities of emotion expression [26]. To train TEC models, corpora of texts with high-quality and detailed markup of emotional categories are required, which is a non-trivial task due to the subjectivity of emotion assessment [20].

There are different types of emotional categories in a text: basic emotions (e.g., joy, anger, sadness), complex emotions (e.g., disappointment, surprise), and psychological states (e.g., stress, anxiety) [9]. The choice of a suitable dataset significantly affects the quality and accuracy of model training.

1.4.2. Methodological approaches to recognizing emotions from text

For example, the article "Text-Based Emotion Recognition Using Deep Learning Approach" [31] presents a hybrid model that combines deep learning methods (CNN and Bi-GRU) with classical machine learning (SVM) to recognize emotions in text. The system used pre-trained word embeddings (matrix size 18210×300) and was trained on a combined corpus of ISEAR, WASSA, and Emotion-Stimulus datasets. Bi-GRU achieved the highest individual accuracy (79.46%), CNN had the best F1-score (80.76%), and the hybrid model reached 80.11% accuracy and an F1-score of 81.27%. The authors highlight the potential of such combined methods for processing diverse texts like dialogues, tweets, and reviews.

Another approach is described in the article "Convolution SSM Model for Text Emotion Classification" [11]. The paper presents a new ConvSSM model that combines the strengths of CNN (TextCNN) with the ability to model long-term dependencies using the Mamba block. The model showed better results in the task of classifying emotions from text than the basic TextCNN and TextRNN models: accuracy 65.71%, precision 73.90%, recall 63.42%, F1-score 62.27%. The authors emphasize that ConvSSM is better at recognizing six basic emotions (anger, fear, joy, love, sadness, and surprise) from Twitter data.

The article "Automatically Classifying Emotions based on Text: A Comparative Exploration of Different Datasets" [3] investigates the impact of different datasets and augmentation strategies on the effectiveness of models for recognizing emotions from text. The authors compared models on three datasets: WASSA-21, COVID-19 Survey, and GoEmotions. The results showed that combining GoEmotions with other datasets improves F1-macro for some models - in particular, ELECTRA-Large gained +9% when trained on the combined COVID-19 + GoEmotions dataset. The best results were achieved for RoBERTa (F1-macro up to 60% on WASSA-21 and 51% on COVID-19). It is also shown that manually tagged Reddit posts can be successfully classified using models trained on thematically similar data.

In the article "Context Unlocks Emotions: Text-based Emotion Classification Dataset Auditing with Large Language Models" [5] discusses a new approach to classifying emotions in text by adding context using LLMs. The authors introduce an algorithm for detecting sentences that do not have sufficient emotional context, which is often the cause of classification errors. To solve this problem, the authors propose to supplement such sentences with context by means of specially formed prompts to the GPT-4 and GPT-3.5 models. Experimental results have shown that context enrichment significantly improves classification accuracy on non-domain-oriented datasets, including SemEval 2017. For example, the accuracy of the GPT-4 model on the validation set increased from 61.2% (without context, CA) to 67.5% (with context-enriched, CAM). Similar improvements were observed for GPT-3.5.

To visualize the advantages and disadvantages of each of the methods under consideration, let's draw a table:

Table 1.2. Analysis of methods for recognizing emotions from text described in the reviewed scientific papers

Method	Source	Dataset	Set of emotions	Accuracy	Advantages	Disadvantages
Bi-GRU	[31]	ISEAR, WASSA, Emotion-Stimulus	6	79.46%	- captures the sequence, effective for dialogs	- requires large amounts of data
CNN	[31]	ISEAR, WASSA, Emotion-Stimulus	6	80,76%	- fast learning, good accuracy	- does not catch long dependencies
Hybrid CNN+Bi-GRU+SVM	[31]	ISEAR, WASSA, Emotion-Stimulus	6	80.11%	- combines the advantages of all approaches	- complex structure, harder to interpret
ConvSSM	[11]	Emotions Dataset	6	65,71%	- captures both local and global connections	- weaker results on non-Twitter data
ELECTRA-Large	[3]	COVID-19 Survey and GoEmotions	5	65%	- improvement after augmentation	- quality decreases without additional data
RoBERTa	[3]	WASSA-21 and GoEmotions	5	67%	- high stability, good metrics	- poor generalization to new domains
GPT-4 (з CAM)	[5]	GoEmotions	9	67,5%	- works well with ambiguous texts	- high computational cost

Considering the advantages and limitations of the analyzed methods, it becomes evident that a universal solution capable of accurately classifying a wide range of emotions across diverse text types has not yet been achieved. This highlights a scientific and practical challenge that is addressed in the following sections of this work.

1.5. Conclusions

This chapter provided a comprehensive analysis of the current state of research in the field of emotion recognition, focusing on both speech and text modalities. The growing interest in this area is driven by its wide practical relevance, particularly in applications such as sentiment analysis, public speaking enhancement, and psychological monitoring. Despite significant progress, current tools often fall short in delivering detailed, actionable feedback for improving emotional expression. The analysis presented here lays the foundation for addressing these gaps in subsequent sections of this work.

The review of scientific literature and market solutions revealed a notable upward trend in publications related to both voice and text emotion recognition, especially in the past five years. This reflects increasing scientific interest and technological advancements in the field. While numerous solutions exist, most focus on auxiliary tasks like mood monitoring or general performance evaluation, rather than on personalized, speech-specific emotional improvement. Systems like “Oration Master” are limited in scope and do not provide targeted feedback. In contrast, the system proposed in this study aims to fill that gap by offering precise recommendations to enhance emotional expressiveness in speech.

Despite advancements, automatic emotion recognition still faces several fundamental challenges. Emotions are inherently subjective and context-dependent, often manifesting as blends rather than discrete states. These issues are particularly pronounced when working with audio and text data, where factors such as cultural differences, background noise, and data imbalance can complicate classification. The complexity of these problems underscores the need for modality-specific approaches that balance scientific rigor with practical applicability - especially in real-time, user-

centered systems.

In the area of speech emotion recognition (SER), key considerations include selecting appropriate datasets, identifying informative acoustic features, and designing effective classifiers. Traditional methods like GMMs and SVMs continue to show merit, especially when used in combination, while deep learning models such as CNNs and RNNs offer improved performance with larger datasets. However, trade-offs remain among model accuracy, training complexity, and real-time usability. The literature review emphasizes the importance of hybrid or ensemble models tailored to spontaneous, natural emotional expressions, which better reflect real-world speech dynamics.

Emotion recognition from text also presents its own set of challenges, particularly due to the abstract and nuanced nature of language. Unlike speech, which conveys emotion through prosody and intonation, text relies on semantics, syntax, and contextual cues. While modern NLP techniques, including transformer-based models, have significantly advanced the field, they still require extensive annotated datasets and careful tuning to maintain high accuracy. Moreover, the subtlety of emotional tone in written language demands that models balance sensitivity and specificity to ensure meaningful, actionable output.

2. RESEARCH SECTION

2.1. Data requirements

Having analyzed the methods used in various scientific papers and reviewed the datasets that the authors used to achieve their goals, we will now move on to solving the problems for the successful operation of the SER system, considering the needs and requirements of this study.

Considering the information provided in the literature review, we can identify the following features that should be considered when selecting a dataset:

- Type of data (acted, evoked, natural spontaneous emotions). As mentioned above, spontaneous natural emotions are best suited for the task at hand. However, it is not easy to find a set containing this type of data.
- Number of emotions. Comparing various methods described in scientific papers, we see that the model's accuracy decreases with the number of emotions. However, classifying more emotions allows the system to more accurately determine a person's emotional state by voice. For example, instead of simply recognizing basic emotions such as happiness or anger, the system can detect smaller emotions such as surprise or delight.
- A set of emotions. Since the goal of the task is to recognize a general range of emotions, it is important to see that the dataset contains a variety of emotions, including positive, negative, and neutral, and not just a variety of ones, such as negative.
- The size of the dataset. It is also important to consider the amount of data, as it affects the ability to train the model effectively and its overall performance.
- Languages of the dataset. Datasets containing English are best suited because English is a language that is widely used internationally.

2.2. Selected datasets

2.2.1. RAVDESS

RAVDESS Emotional speech audio [17] is a dataset containing only speech

audio files from the RAVDESS dataset. These are the recordings of 24 actors, including 12 women and 12 men; each actor participated in 60 trials and voiced two lexically matched statements in a neutral North American accent. Thus, we have a total of 1440 audio files containing the following emotions expressed in speech: calm, joy, sadness, anger, fear, surprise, and disgust. Each emotional expression has two levels of intensity - normal and strong, as well as an additional neutral expression.

Each of the files in the set has a unique name. The file name consists of a 7-digit numeric identifier (for example, 03-01-06-01-02-01-12.wav). These identifiers determine the following characteristics of the recording:

- Modality: full audio system, video only, audio only
- Voice channel: speech, song
- Emotion: neutral, calm, joy, sadness, anger, fear, disgust, surprise
- Emotional intensity: normal, strong
  * for "neutral" emotion there is no strong intensity
- Text of sample: "Kids are talking by the door", "The dogs are sitting by the door"
- An attempt: 1^{st} or 2^{nd}
- Actor (from 1 to 24): odd number – men, even number – women.

Thus, you can extract all the characteristics of the recording from the name of the audio file.

The main advantages of RAVDESS Emotional speech audio are:

1. A variety of emotions: as mentioned above, this set contains 8 emotions. Compared to the other sets, this one contains a moderate number of emotions, which will allow creating a model capable of recognizing a wide range of human emotions. This set of emotions is the middle ground between overly simplified sets with only a few basic emotions and overly complex sets with more than a dozen different emotional states, which makes it difficult to train the model and apply it in real-world settings.
2. Uniform data distribution: the set contains the same number of all emotions, except for neutral, since each emotion has two types of intensity: normal and

strong, and neutral has only normal intensity. The set also contains equal numbers of voice recordings of men and women. This ensures a balanced representation of each emotion and gender, which avoids model bias during training and improves the accuracy and reliability of the model. Details of the emotion distribution can be seen in the figure below:

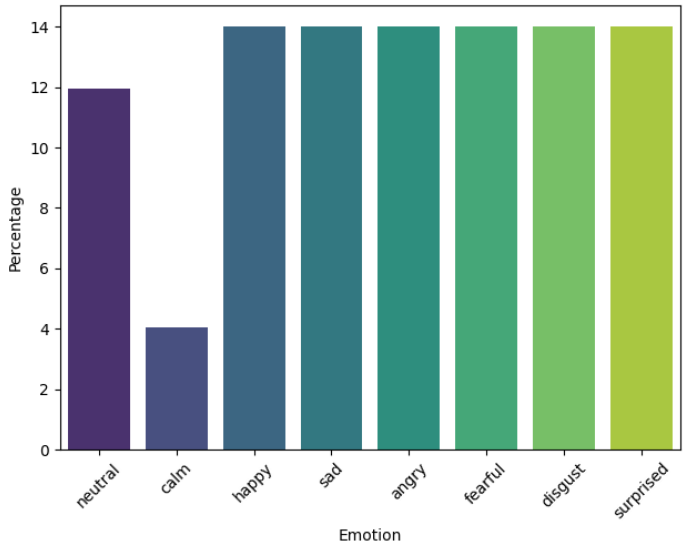

Fig. 2.1. The ratio of audio tracks for each emotion

3. High-quality audio tracks: the audio files in the set were recorded in a studio environment with high sound quality, which ensures that each emotion sounds clear and legible, which in turn increases the accuracy of emotion recognition. The figure below shows the average audio duration of each emotion:

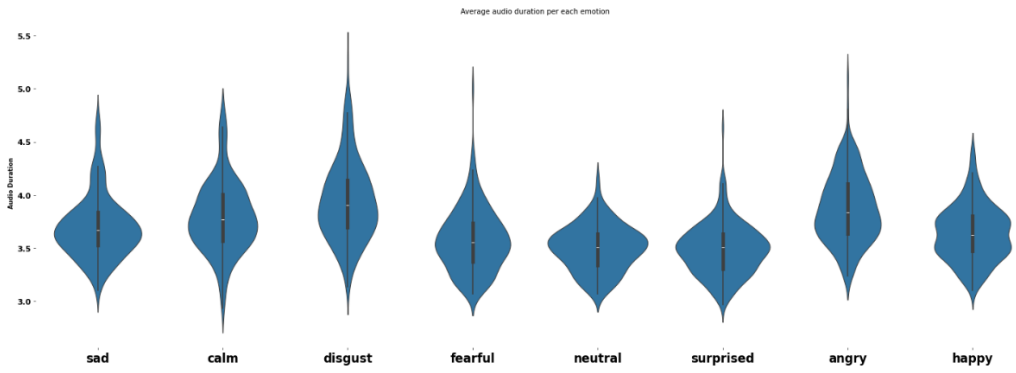

Fig. 2.2. Average audio duration for each emotion in the set

4. Moderate amount of data: The dataset contains 1440 audio files, which is a moderate amount of data to classify. It may be advisable to supplement the set with processed data from the full RAVDESS set to improve the achieved

accuracy, but for now, this data is sufficient.

To train the model to recognize emotions well, only the audio track and the emotion of the recording will be taken from the set. Since in real-world use the model will only receive audio as input and will have to output an emotion as output, there is no point in trying to use information such as gender, intensity, voice channel. The text of the recording plays a role in predicting the most appropriate emotion, as it contains context that helps interpret the content, but at the same time can distract from the true emotional tone of the speech.

2.2.2. TESS

TESS (Toronto Emotional Speech Set) [25] is a dataset containing audio recordings of speech delivered by two actresses (aged 26 and 64) to simulate seven basic emotions: anger, disgust, fear, joy, surprise, sadness, and neutral. Each of the actresses spoke 200 different words in the phrase "Say the word _", expressing each of the emotions. In total, the set contains 2800 audio files in WAV format. The dataset has been widely used in emotion recognition research due to its clear and diverse emotional expressions.

Each audio recording is placed in the appropriate folder according to the actress and the emotion she expresses. This makes it easy to navigate through the data and provides convenience for pre-processing or separating into training and test samples.

The main advantages of the TESS set are:

1. High-quality studio sound: recordings are made under controlled conditions, which guarantees clear audio and reduces noise that could affect the quality of classification.
2. Balanced emotions: each word is voiced with the same number of samples for each emotion, which avoids imbalance of classes in the model training process. Details of the emotion distribution can be seen on the Figure 2.3.
3. Optimal amount of data: 2800 audio files is enough for initial training of emotion recognition models, especially when combined with other datasets (e.g., RAVDESS described above), which allows for higher classification accuracy.

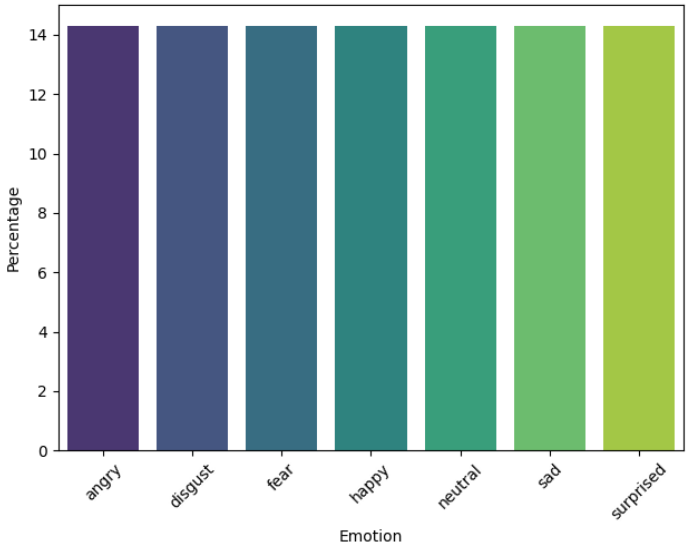

Fig. 2.3. Ratio of audio tracks for each emotion

For the purposes of this work, only audio tracks and the corresponding emotion of the recording will be used from the set. As in the case of the RAVDESS set, no information about a specific actor or phrase text will be used, as the model must learn to determine the emotion based solely on the voice.

2.2.3. MC-EIU

MC-EIU (Multimodal Conversations with Emotion-Intent Understanding) [28] is a multimodal dataset created to model emotional conversations in realistic settings. It includes video clips from 3 English-language (716 episodes) and 4 Chinese-language (119 episodes) TV series of various genres (drama, romance, crime, etc.). All clips are accompanied by subtitles, which were used to automatically extract the text and timing of the lines. Only the English part of the data was used for training.

Basic information about the set:

- Number of English dialogues: 4 013
- Total number of lines (in English): 56 012
- Annotation: emotions, intentions, speakers
- Emotions (according to Eckmann's scheme + neutral)

After extracting the English part of the labeled text of the recordings, the following data distribution was obtained: neutral - 12 488, happiness - 5 081, anger - 2 029, sadness - 1 052, fear - 902, surprise - 844 and disgust 344.

To visualize the distribution of the data, a graph of the percentage of classes in

the set was plotted:

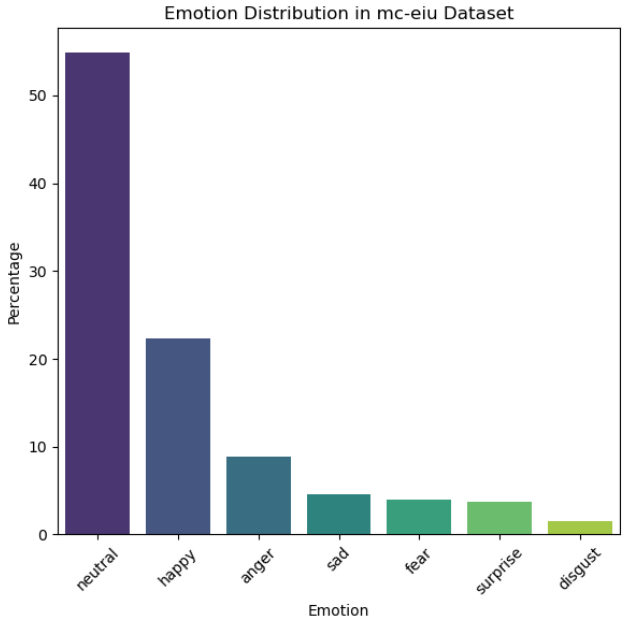

Fig. 2.4. Data distribution in the MC-EIU set

Advantages of MC-EIU:

1. Diversity of emotions: the set covers 7 emotions, including neutral, which allows the model to distinguish between both positive and negative affects.
2. Multimodality and naturalness: the data is taken from different TV series, so the dialogues are quite spontaneous and natural, close to real everyday conversations.
3. Data volume and diversity: more than 56 thousand (22.5 thousand after processing) English-language lines from dialogues of different genres provide a broad context for model training.
4. Contextual integrity: all dialogues contain at least two rounds of interaction between two speakers, which provides a deeper context of the conversation.

Disadvantages of MC-EIU:

1. Class imbalance: the number of utterances labeled with neutral and happy emotion far outweighs the less represented classes, such as "disgust" or "fear". This can cause a bias in the model during training and requires the use of compensation methods (e.g., class weighting or sample balancing).

2.2.4. MELD

MELD (Multimodal EmotionLines Dataset) [32] is a multimodal dataset, an

extension of the previous EmotionLines set, designed for recognizing emotions in multi-speaker dialogues. Unlike MC-EIU, which focuses mainly on dialogues between two participants, MELD covers multi-party conversations, which adds complexity to the classification, but better reflects natural communication scenarios.

In MELD, each sentence is accompanied by three modalities: text, audio, and video. The annotation was based on Eckman's basic emotion scheme with the addition of a neutral emotion, similar to the approach in MC-EIU.

Basic information about MELD:

- Number of dialogs: 1,039 for training, 114 for validation, and 280 for test
- Number of utterances: 9,989 for training, 1,109 for validation, and 2,610 for test
- Emotion labels (7 categories): neutral, happiness, anger, sadness, fear, surprise and disgust.

All the lines were labeled by at least 3 annotators, and cases of complete disagreement were excluded. For emotional labeling, a collective voting approach was used to ensure consistency.

Advantages of MELD:

1. Multimodality: like MC-EIU, MELD includes text, audio, and visual information.
2. Multi-spoken: Unlike many sets with paired dialogues (e.g. MC-EIU), MELD contains complex multi-spoken interactions where it is important to consider context and interpersonal relationships.
3. Clear structuring: the dataset is already divided into train, test, and validation parts, which simplifies integration into deep learning pipelines.
4. Volume: more than 13,000 replicates - almost twice as many as in the previous set, which allows the model to learn from richer scenarios.

Disadvantages of MELD:

5. Unbalanced: As in the set discussed above (MC-EIU), the number of examples in different emotional classes varies significantly. For example, "neutral" and "joy" are represented much more frequently than "disgust" or "fear", which creates a problem of class bias when training models. This

requires additional strategies such as stratified sampling, balancing, etc.

To visualize the imbalance of the data, we plotted the distribution of data by class in each set:

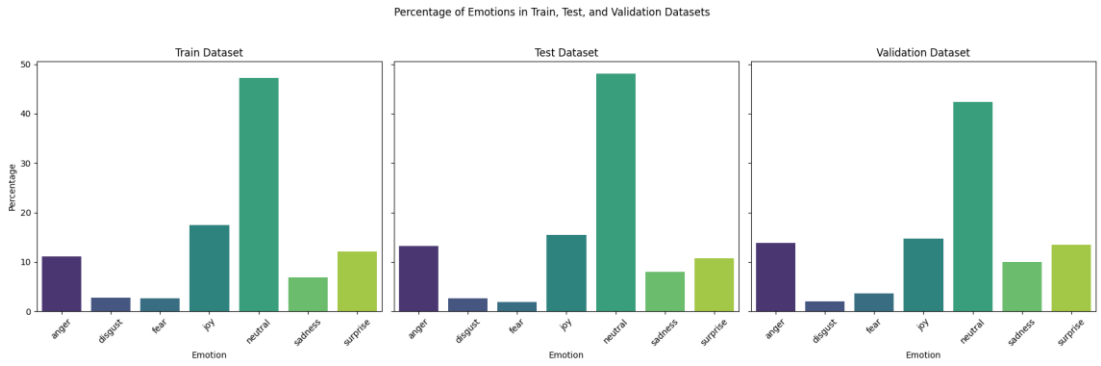

Fig. 2.5. Data distribution in the MELD train, test, and validation sets

2.2.5. GoEmotions

GoEmotions [8] is a large dataset for recognizing emotions in texts created by Google researchers based on comments from Reddit. Unlike MC-EIU and MELD, which have a dialogic structure, GoEmotions is a monologue set consisting of individual user texts, mostly short remarks. Its main feature is a wide variety of emotions: 27 emotional categories plus neutral, covering a wide range of feelings.

Basic information about GoEmotions:

- Total number of examples: 58 009
- Number of emotions: 27 main ones + neutral
- Number of unique annotators: 82
- Number of annotators per example: 3 or 5
- Most examples have one label (83%), and in 94% of examples at least two annotators agreed on at least one emotion.

Emotion labels were selected from a predefined list, but raters were allowed to select several if emotions were expressed simultaneously. Only recordings with one emotion were selected for this set.

For the purposes of this work, 7 emotions were selected as the most relevant for classification and used in the voice recognition module: neutral - 55 298, happiness - 4 329, anger - 5 202, sadness - 3 827, fear - 1 778, surprise - 3 472 and disgust 2 914.

Advantages of GoEmotions:

1. Largest among the reviewed sets: with more than 58 thousand examples, GoEmotions surpasses MELD and MC-EIU in terms of volume, which allows for deep model building.
2. Wide coverage of emotions: The 27 categories cover both basic emotions, making the set valuable for more complex emotional expression studies.
3. High quality markup: The data were marked up by native English speakers from India, with careful instructions and a mechanism for marking up difficult examples.

Disadvantages of GoEmotions:

1. Imbalance: As with the MC-EIU and MELD sets, GoEmotions shows a significant imbalance between classes. For example, neutral has almost ten times more examples of fear or disgust.
2. To demonstrate the imbalance of the data, we plotted the distribution of the data in the set between the classes:

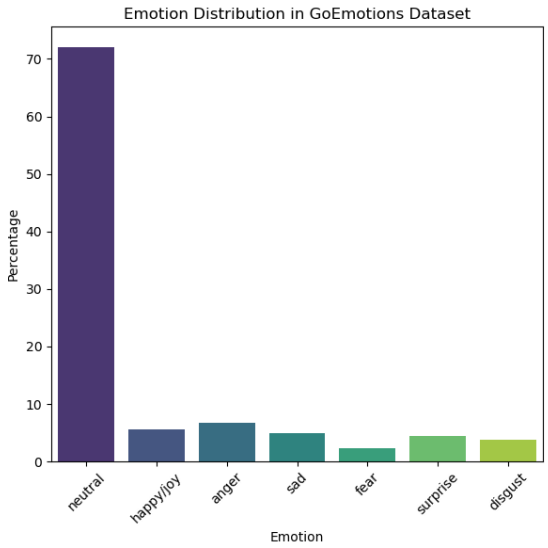

Fig. 2.6. Distribution of data in the GoEmotions set

3. Single-speaker structure: unlike MELD or MC-EIU, which contain the context of the dialog, GoEmotions consists of individual utterances, which makes it difficult to consider the context of emotions in ambiguous utterances.
4. Not natural origin of the texts: unlike the sets discussed above, the sentences are taken from the real Reddit environment, which on the one hand ensures spontaneity and vibrancy of emotional expression, but on the other hand does

not contain recordings of live conversations between people.

2.2.6. Comparison of datasets

To better understand the features of each dataset, let's build a table comparing all the datasets that are planned to be used in the study:

Table 2.1. Comparison of the sets used in this study

Dataset	RAVDESS	TESS	MC-EIU	MELD	GoEmotions
Data type	evoked	evoked	acted	acted	-
Number of emotions	8	7	7	7	28
The size of the dataset	1440	2800	45 007	1433	58 009
Language	English	English	English	English	English
SER task	+	+	+	+	-
TEC task	-	-	+	+	+

As shown in Table 2.1, the datasets vary in size, number of emotions, and suitability for specific tasks, providing a diverse foundation for both speech and text-based emotion classification.

2.3. Requirements for the model

Within the emotion recognition task of this study, two main areas were identified: emotion recognition in voice (audio) and emotion classification in text. Each of these areas has both common and specific requirements for the model.

Common requirements:

- Classification accuracy. Regardless of the type of input data (audio or text), the model must accurately classify emotions with the ability to scale up to 6-8 classes.
- Possibility of retraining. The models should have an architecture that allows for fine-tuning on new data sets, including those with different language or acoustic specifics (e.g., Ukrainian).
- Optimal performance. Due to the possible processing of large amounts of

data or integration into a web interface, it is desirable that the models demonstrate good performance on both GPU and CPU.

Requirements for models for audio (SER):

- Processing of spectrograms or raw audio. The model must accept mel-spectrograms or raw audio signals as input.
- Resistance to noise and speech variations. Considering different accents, intonations, and speech speed, the model should capture both local (changes in intonation) and global emotion signs.
- Compatibility with small amounts of annotated data. It is desirable that the model supports training with a limited number of labeled examples.

Requirements for models for text (TEC):

- Work with short texts. As the input is expected to be passages of 3-5 words in length, the model must be capable of extracting meaningful semantic and emotional information from limited context.
- Contextual understanding. The model should consider not only keywords but also the context in which they are used (for example, irony or sarcasm).

2.4. Selected approaches

Let's move on to choosing emotion recognition approaches. After analyzing the methods and tools described in various literary sources, several approaches were considered to implement the emotion recognition models.

2.4.1. CNN

CNN architecture is a combination of three components:

1. Convolutional layers: contain a certain number of filters applied to the input data. Each filter scans the input data using the dot product method to create a certain number of feature maps in one convolutional layer.
2. Pooling layers: used to reduce or shrink the feature maps of objects. There are several schemes of dimensionality reduction: max pooling, min pooling, mean pooling, average pooling, etc.
3. Fully connected layers: used to extract global features that are fed to the SoftMax classifier to determine the probability for each class.

CNN arranges all these layers in a hierarchical structure: convolutional layers (CL), pooling layers (PL), and then fully connected layers (FCL), followed by the SoftMax classifier. [22] For a better understanding of the CNN architecture, here is an example of the simplest CNN architecture. Its visualization can be seen on Fig. 2.7.

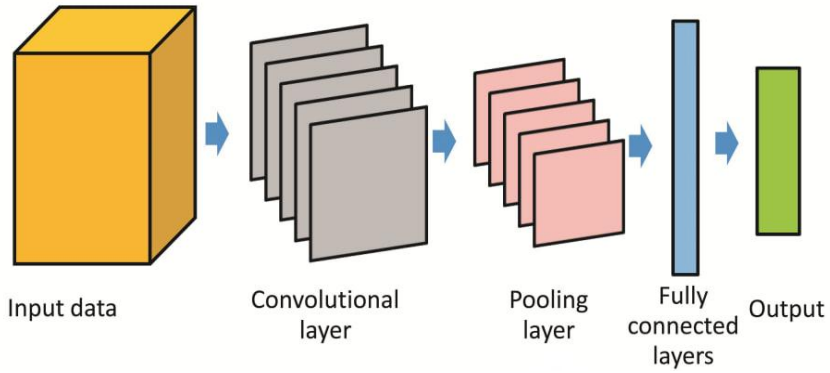

Fig. 2.7. CNN with 1 convolutional, 1 connecting and a fully connected layer [21].

The advantages of using this approach are:

- Extraction of high-level features from low-level raw pixel information. That is, the CNN can automatically detect local patterns in the input mel-spectrograms, such as changes in amplitude and frequency, which are important for emotion recognition;
- Due to the pooling layers, CNNs are robust to scale changes and shifts in the input data, which makes the model more flexible to different variants of audio signals;
- CNNs can adapt to complex and multidimensional input data, such as mel-spectrograms, due to their ability to reduce dimensionality through convolutional and pooling layers.

These advantages make CNNs a good choice for emotion recognition tasks based on audio data.

2.4.2. HuBERT

HuBERT [36] is a model that uses a transformer architecture to process speech data and learn from large amounts of unlabeled data in a semi-supervised environment. The model combines two main components:

- Pre-training: learns from raw audio signals by creating hidden units

through unsupervised clustering of audio fragments.

- Transformer: uses the standard transformer architectural block, which consists of a multi-layered self-attention mechanism. This allows the model to consider the global context of speech, which is important for recognizing emotional patterns in long audio signals.

The HuBERT model works in such a way as to gradually refine its hidden units during training on audio data, which allows for an improved representation of semantic and acoustic features of speech. It is important to note that HuBERT does not require a large amount of labeled data - the model adapts well to tasks where there are no fully annotated sets. Figure 2.8. illustrates the HuBERT approach, where the model predicts the hidden cluster assignments of masked audio frames (e.g., y_2, y_3, y_4), generated through one or more iterations of k-means clustering.

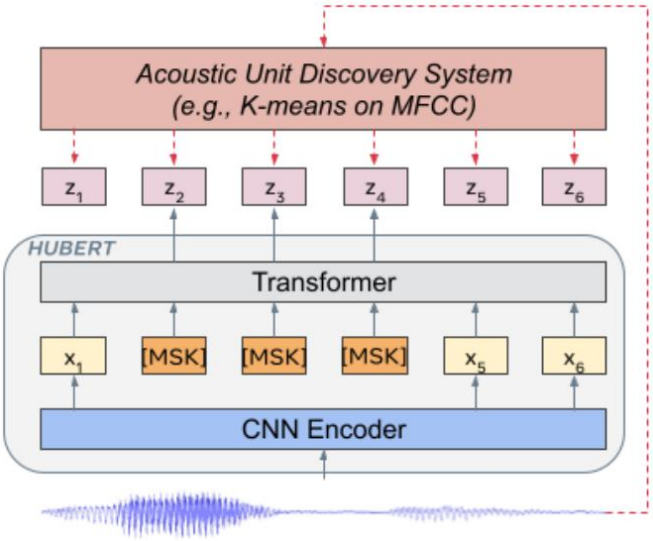

Fig. 2.8. HuBERT approach [36]

Advantages of HuBERT in the context of this study:

- Ability to work with a small amount of annotated data: Thanks to pre-training on large volumes of raw audio, the model can achieve high accuracy, which makes it especially easy to build your own dataset.
- Ability to work with raw audio data: HuBERT can use large amounts of unannotated audio data for pre-training, which reduces the need for fully labeled sets.
- Deep understanding of the speech context: the use of transformers allows

HuBERT to capture both local and global acoustic and semantic features of speech, which is important for correct emotion recognition.

- Flexibility to different languages and accents: due to pre-training on a variety of audio signals, the model demonstrates high resistance to speech variations, which makes it versatile for use in different contexts and will ensure a smooth transition from the English-language dataset to the Ukrainian-language one.

Thus, the HuBERT model can be considered a good choice for emotion recognition tasks based on audio data.

2.4.3. DistilBERT

distilbert-base-uncased-emotion [7]: a fine-tuned model based on DistilBERT [35], created for the task of classifying emotions in text. It was trained using the Hugging Face Trainer on the Emotion Dataset, which contains examples of texts with corresponding emotional labels. DistilBERT is a simplified and accelerated version of BERT obtained by knowledge distillation during pre-training. In the process of distillation, the model learns from a large BERT model, retaining most of its capabilities, but with fewer parameters.

Basic information:

- Architecture: DistilBERT (based on BERT, but lighter)
- Input: English-language text (tokenized as uncased, i.e. case insensitive)
- Size: Approximately 66 million parameters (40 percent smaller than BERT-base)
- Format: PyTorch via Hugging Face Transformers
- Fine-tuning: training lasted 8 epochs, learning rate was 2e-5, batch size was 64
- Number of classes: 6 or more depending on the task, can be adapted to 7 emotions such as neutral, happy, angry, sad, surprised, fear, and disgust.

Advantages of the classification task:

1. Compactness and speed. DistilBERT is much lighter than the classic BERT, which allows it to run on the CPU and reduce the inference time. This is

important for high-performance systems or mass text classification.

2. High accuracy with lower complexity. Despite the reduction in size, the model retains up to 97 percent of BERT quality in speech understanding, which is sufficient to accurately detect basic emotions.
3. Readiness for adaptation. The model is already adapted to the task of emotional classification. It can be easily retrained or retrained on new data, including a set of 7 target emotions, such as GoEmotions or MELD.
4. Integration with Hugging Face. Full compatibility with the transformers library allows you to quickly integrate the model into pipelines, visualize results, use tokenizers and metrics with minimal effort.

2.4.4. RoBERTa-base

twitter-roberta-base-emotion [34] is a model customized on RoBERTa-base that was created for the task of classifying emotions in social media texts, including tweets. It was pre-trained on approximately 58 million tweets and then retrained specifically for emotion classification within the TweetEval benchmark. TweetEval is a set of tasks that covers various aspects of NLP for Twitter, including emotion recognition.

Main features:

- Architecture: RoBERTa-base (an improved version of BERT trained on a larger corpus without NSP tasks)
- Input: English text adapted to the specifics of short messages on Twitter
- Pre-training volume: approximately 58 million tweets
- Format: PyTorch with Hugging Face Transformers library
- Fine-tuning: performed on TweetEval benchmark data for the emotion recognition task
- Number of classes: several basic emotions (e.g. happiness, anger, sadness, fear, love, surprise), can be adapted to other types

Advantages for the task of classifying emotions from text:

1. Optimization of social media texts. The model has been trained on tweets, so it recognizes emotions better in informal, abbreviated, or slang messages typical of modern online communication.

2. High quality thanks to RoBERTa. RoBERTa's architecture provides better text representation than the original BERT due to more thorough pre-training. This increases the accuracy of emotional classification.
3. Compatibility with Hugging Face. The model can be easily integrated into modern NLP pipelines through the transformers library and supports fast testing, fine-tuning, and quality assessment.
4. Readiness for retraining. Although the model is already trained on typical emotions, it can be easily adapted to a set of 7 emotions, such as those in GoEmotions.

2.5. Speech emotion recognition models. Implementation.

To recognize emotions from voice signals, two approaches were implemented: a model based on a convolutional neural network (CNN) and a model based on the HuBERT transformer. The transformer models were trained on different corpora - separately on RAVDESS, as well as on the combined set of RAVDESS and TESS.

2.5.1. CNN

One of the first approaches chosen to solve the problem of recognizing emotions from voice signals was a convolutional neural network (CNN). This type of approach has proven itself in image processing tasks, so it was adapted to audio using mel-spectrograms, a visual representation of an audio signal that allows detecting spatial patterns characteristic of different emotions. The CNN model was trained to classify the spectrograms into eight emotional states based on their frequency and time features.

2.5.1.1. Data distribution

Let's start by preparing the data for use in model training. After downloading the dataset, the data was divided into three parts: train, test, and validation. Two actors were randomly selected for the test and validation datasets, including one person of each gender (1 male, 1 female). The remaining 20 actors were assigned to the train set. This distribution of data can be explained by the following factor: when classifying emotions in the real world, the model will "hear" the voice and mannerisms of a particular person for the first time, and therefore, using the recordings of one actor in both the training and test or validation parts is inappropriate.

It is also worth considering the uniformity of the data distribution: uneven data distribution will negatively affect the model's ability to recognize emotions evenly. It is necessary to add records of each type of emotion to the test and validation parts to be able to determine the accuracy of recognizing each emotion.

The details of the emotion distribution for each set can be seen in Figure 2.9.

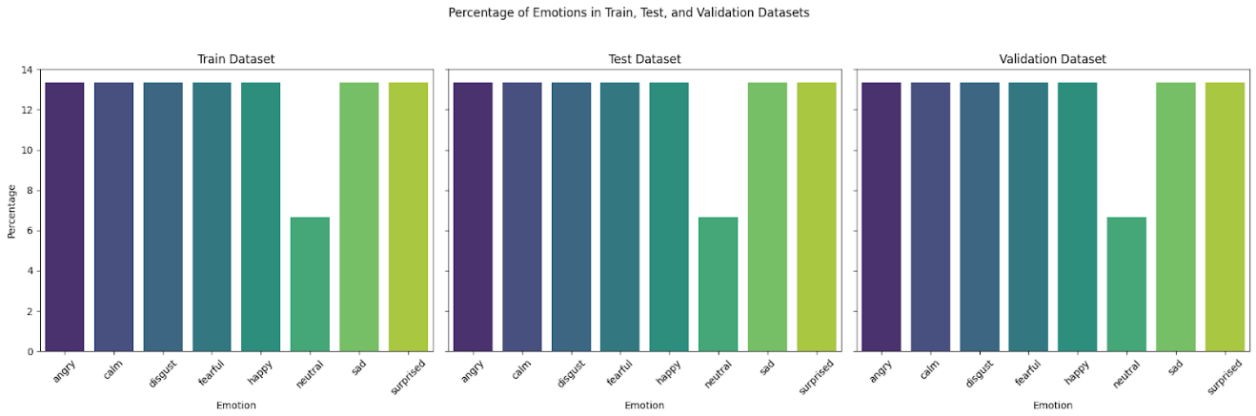

Fig. 2.9. Data distribution of each emotion for the train, test, and validation sets.

2.5.1.2. Audio pre-processing

After dividing the data into separate sets, the name of each recording was used to extract its features and characteristics, including: the path to the location where the recording is stored, the modality of the recording, its voice channel, the emotion reproduced, the emotional intensity, the expression, the attempt number, and the number and gender of the actor. All these details can make it easier for the model to recognize voice emotions in general, but it is important to note that the input to the human speech analysis system is only audio. Therefore, we will train the model using only the audio recording and the detected emotion as input.

So, after removing all unnecessary features, the only ones left in the sets are the path to the audio recording and the detected emotion. The recordings themselves, which are located along the paths specified in the set, were converted into mel-spectrograms, which is one of the most informative representation of audio for the CNN model. Visualization of spectrograms helps to understand the differences between emotional states at the frequency response level.

Data augmentation was applied to improve the generalizability of the model. Two additional variants were created for each recording: a time shift and addition to a constant length and a change in speed and pitch or addition of noise.

This allowed us to increase the size of the training set from 1200 to 3600 recordings, which significantly improves the model's resistance to variations in sound characteristics.

2.5.1.3. Model training

Let's move on to building the model architecture. In addition to the basic layers of the CNN model described earlier in this Section, we added the Batch Normalization and Dropout layers, as their use contributes to a stable and efficient learning process, preventing overfitting and improving the overall generalizability of the model.

Thus, the model consists of four main blocks, each of which contains a convolutional layer (Conv2D), a normalization layer (BatchNormalization), a ReLU activation layer (Activation), a pooling layer (MaxPooling2D), and a random dropout layer (Dropout). Convolutional layers are used to extract useful features from the input data, normalization to stabilize the learning process, ReLU activation to add nonlinearity to the model, maximal pooling to reduce the dimensionality of the data, and random dropout layers to prevent overfitting.

After these blocks, the data is aligned and passed through two fully connected layers (Dense), which are designed to detect complex patterns in the data. The last layer uses the softmax activation function to derive probabilities for each emotion.

To visualize the architecture of the resulting model, you can see the result of model summarization (model.summary) in Appendix B.

The model training process will include the following steps:

- Model initialization: creating a model using defined architecture. In our case, the model will expect input data in the form of two-dimensional Mel-spectrograms.
- Compile the model: use the Adam optimizer to optimize the model weights and the loss function metric "categorical_crossentropy" to measure the model error during training.
- Training of the model: training will be performed on the set of initial training data and data obtained after augmentation. The training process will use methods such as EarlyStopping and ModelCheckpoint to prevent overfitting and ensure that the model that gives the best results on the

validation data is retained.

- Model Validation: Testing the model on a validation dataset. This will help ensure that the model not only performs well on the training data but can generalize its findings to new data that it has not seen before.
- Testing the model: testing the model on a test dataset to see how it performs on data it has never seen before.

The number of epochs is 100, the batch size is 32.

2.5.2. Transformers (HuBERT)

The second approach to solving the emotional classification problem is transformers, particularly the HuBERT (Hidden-Unit BERT) architecture, which is highly efficient in processing speech signals. Unlike CNN, which works with audio converted into spectrograms, HuBERT operates directly with the raw audio signal, which allows the model to automatically learn important speech features at several levels of abstraction. This approach ensures better generalization and the ability to capture subtle intonation shades characteristic of different emotions. The model was trained on both the separate RAVDESS dataset and the combined RAVDESS + TESS corpus.

2.5.2.1. General scheme of model architecture

2.5.2.1.1. Data distribution

In this study, two transformer models were implemented for speech-based emotion classification. Two different corpora were used to train these models:

- A model based on the RAVDESS corpus
- Model based on the combined RAVDESS and TESS corpora

Accordingly, the distribution of the data was slightly different depending on the chosen corpus. A detailed description of the data distribution is provided below separately for each of the corpora.

2.5.2.1.2. Audio pre-processing

Before feeding audio into the HuBERT (superb/hubert-large-superb-er) transformer model, several important steps are required to ensure that the input data is correct and conforms to the format on which the model has been pre-trained.

The audio preprocessing process consists of the following steps:

- Loading a signal with a sampling rate of 16 kHz: The superb/hubert-large-superb-er model was pre-trained on 16 kHz audio, so all input files must be converted to this format. To do this, we used the FFmpeg tool to transcode audio to .wav format with the specified sample rate and the Librosa library to download audio files and convert them to the required format as numeric arrays with a frequency of 16 kHz. This ensured full compatibility with the model and prevented errors in further processing. In addition, this frequency reduces the computational load compared to higher frequencies (e.g., 44.1 kHz), which is important when working with a large amount of audio.
- Transformation to Hugging Face Dataset: The discretized audio files, along with their respective labels, are organized into a datasets.Dataset structure from the Hugging Face library.
- Audio processing using Wav2Vec2FeatureExtractor: The Wav2Vec2FeatureExtractor object is used to convert audio to a format suitable for HuBERT. It automatically performs signal amplitude normalization, conversion to a float32 array, padding, or trimming of long recordings.
- Batching and saving encoded data: To speed up training and reduce the memory load when loading all encoded data simultaneously, the audio data is grouped into batches (packets) of fixed size (256 samples), which are processed through the feature extractor and saved to disk as tensors.
- Generating labels for training: For each audio sample, a pre-generated numerical label corresponding to the emotion class is extracted.

After completing all the steps, the final dataset is generated, which contains the following fields: input_values, attention_mask, and label. It is ready for submission to the DataLoader and further use for training the HuBERT-based model.

2.5.2.1.3. Model training

The model training process includes the following steps:

- Loading the model: Initialization of the HubertForSequenceClassification model with Hugging Face, specifying the number of classes.
- Creation of data loaders: Creating train_loader, val_loader, test_loader with the specified parameters: batch_size=4, shuffle for the training set, and collate_fn for processing variable length audio input signals and creating attention masks.
- Determination of the loss function: To account for class imbalance, class weights were calculated based on the frequency of each emotion in the training set. These weights are used in the CrossEntropyLoss function, which allows the model to focus on less represented classes.
- Optimizer: The AdamW optimizer was used for training with an initial learning factor of 1e-5, which is well suited for transformers.
- Prediction function: Prediction function that calculates the probabilities through softmax and selects the class with the highest probability as the result.
- Model evaluation: The evaluation function, which puts the model into the evaluation mode, calculates the loss and accuracy on the validation set.
- Training cycle: The model was trained for 8 epochs. For each epoch:
    - a training pass was performed: forward propagation, backward propagation, and updating of the weights by the optimizer;
    - the current accuracy and loss for each battle were calculated;
    - after each epoch, validation on the set was performed, and if the accuracy increased, the model was saved as the best (ravdess-tess-hubert-model-best.pth).
- Saving the results: Upon completion of the training, the collected losses (train_loss, val_losses) and accuracies (train_accuracies, val_accuracies) were saved for further analysis. The file with the best model was also downloaded.

For better visualization of implemented logic there is a flowchart in Appendix B (Figure B.1).

2.5.2.2. Building transformer models based on different corpora

2.5.2.2.1. RAVDESS. Data distribution.

For data partitioning for the transformer model based on RAVDESS alone, we used the approach described earlier for data partitioning for the CNN model (see Section 2.5.1.1). The only difference is that in this case the records were divided into two sets: train and test. In particular, the train set included 22 actors, and the test set included 2 (1 man and 1 woman whose recordings were not included in the train set).

The sizes of the sets were as follows: train set – 1320 and test set 120 copies.

The distribution of data in the resulting sets can be seen in Figure 2.10.

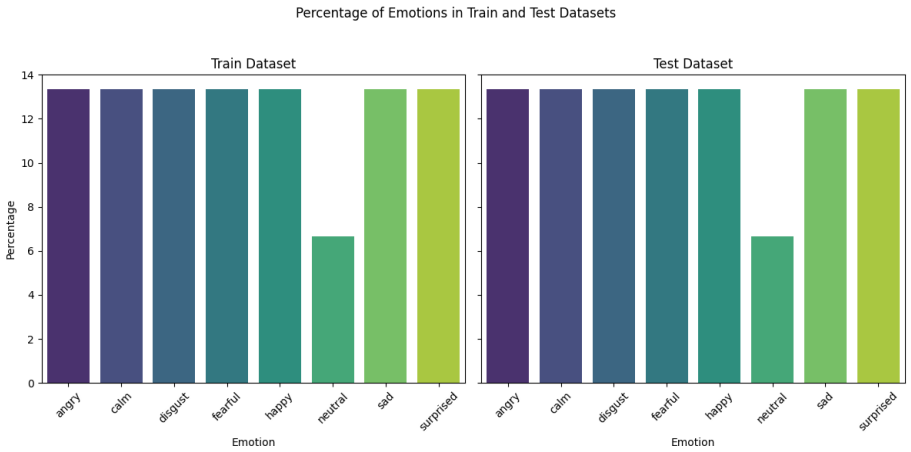

Fig. 2.10. Data distribution of each emotion for the training and test sets.

2.5.2.2.2. RAVDESS & TESS. Data distribution.

To build a transformer model based on RAVDESS and TESS, the same approach was used to distribute the RAVDESS data set as for the CNN model. However, for this dataset, all audio files with the "calm" emotion were preliminarily removed to match the number of classes with the TESS dataset, which has 7 classes (without "calm").

The division of the TESS set was organized as follows: first, keywords ("time", "near", etc.) were extracted from the audio file names, and then 40 unique words were randomly selected to build the train, validation, and test sets. Of these, 20 were assigned to the validation set, another 20 to the test set, and the rest to the train set. This approach ensures that the same words do not appear simultaneously in different subsets of the data.

After pre-processing both datasets, for each subset from both sources (train, validation, and test), were generated the corresponding dataframes, which included

only the path to the audio file and the emotion label. After that, the corresponding subsets (train, validation, and test) were combined by concatenating the dataframes from RAVDESS and TESS. Thus, single fused sets were formed for further training, validation, and testing of the transformer model.

The resulting set sizes are as follows: train set – 3280, validation set – 384 and test set 384 instances.

The distribution of data in each of the combined sets can be seen in Figure 2.11.

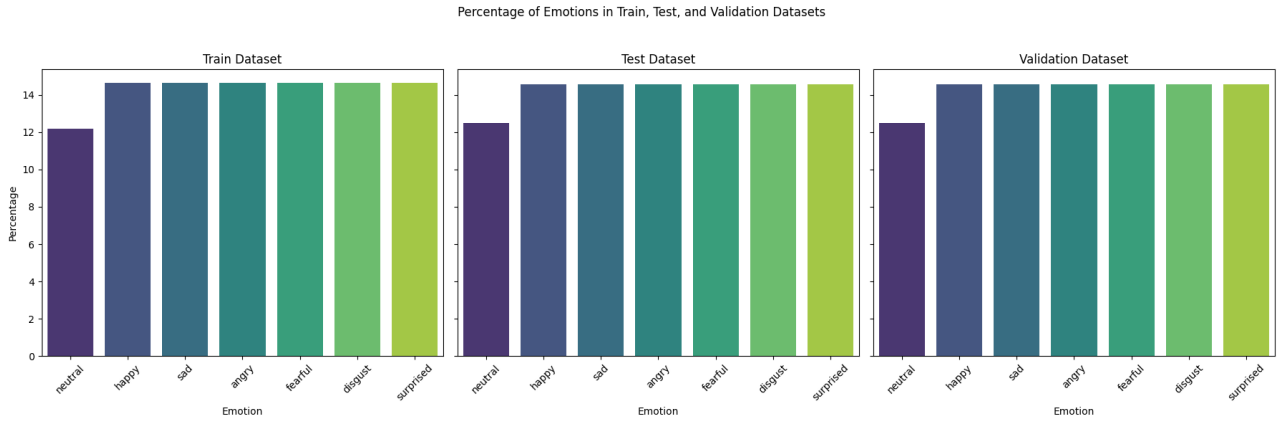

Fig. 2.11. Data distribution of each emotion for train, test, and validation sets.

2.6. Text emotion classification models. Implementation.

To classify the most appropriate emotions from text data, two approaches were implemented: a conventional multi-class classification and a two-stage system that first distinguishes between neutral and emotionally colored sentences and then classifies non-neutral emotions with another model. Both approaches were tested using different transformer models and datasets, including MELD, MC-EIU, and GoEmotions. This approach made it possible to compare the results obtained from both a simple multi-class architecture and a cascade structure in the face of an unbalanced distribution of emotions in the data.

2.6.1. Text pre-processing

Three datasets were used to train the models: MELD, MC-EIU, and GoEmotions. The models were obtained by training different transformers on these datasets. Despite the differences between them, the text preprocessing was performed according to a single scheme that included the following steps:

- Data loading: Each of the datasets were loaded and divided into train,

validation, and test data.

- Emotion binarization: In the binary classification task, all emotions other than neutral were combined into one class (non-neutral), while neutral was left as a separate class.
- Balancing the classes: To avoid a significant imbalance in the training samples, the number of examples of the most represented classes was reduced when necessary.
- Multi-class labeling of emotions: For the multi-class task, emotional categories were converted into numerical labels. Synonymous or related emotions (e.g., sad and sadness) were combined under common labels. All datasets were brought to a single format with a fixed number of classes.
- Text tokenization: All texts were tokenized using a pre-trained tokenizer (based on Hugging Face AutoTokenizer). We applied padding and truncation to the specified length, as well as conversion to tensor format. The tokenized samples, along with the corresponding labels, were grouped into batches of a fixed size (256 samples) and saved as .pt files for further use.
- Preparation of DataLoaders: DataLoader objects were created for each of the subsets (training, validation, test). A special function collate_fn was used to generate a batch with the necessary fields: input_ids, attention_mask, and labels.

2.6.2. Model training

The process of training a model to solve the task of classifying emotions from text includes the following steps:

- Model initialization: Depending on the task at hand - binary or multi-class classification - a transformer model architecture with the required number of output classes was chosen.
- Setting up the optimizer and scheduler: The AdamW optimizer with a fixed learning rate was used to optimize the model parameters. We also used a scheduler that linearly decreases the learning rate during training.

- Model training: The training process was performed in epochs. At each step, the model was put into training mode, data batches were processed, losses were calculated, backtracking and weights were updated. Training losses and accuracy were calculated in parallel.
- Model validation: At the end of each epoch, the model was validated in the evaluation mode without updating the weights. Accuracy metrics were calculated, a confusion matrix was built, and a full classification report was generated.
- Early stopping and model saving: An early stopping mechanism was also implemented, which stopped training if the validation loss did not improve for a specified number of epochs. The best version of the model was saved for further use.
- Model evaluation on test data: After the training was completed, the best model was loaded, and the final evaluation was performed on the test dataset.

For better visualization of implemented logic there is a flowchart in Appendix B (Figure B.2).

2.7. Conclusions

This chapter provided a structured overview of the theoretical and practical foundations for implementing emotion recognition. It established a solid base for future development by analyzing data, models, and implementation methods across both speech and text modalities.

Defining data requirements highlighted the need for datasets that balance authenticity, diversity, and size. English-language corpora were prioritized due to their accessibility and compatibility with pre-trained models.

The dataset review underscored complementary strengths. Acted speech datasets like RAVDESS and TESS offered clarity for baseline training, while MELD, MC-EIU, and GoEmotions introduced real-life complexity and emotional diversity.

Model requirements differed by modality. Audio models prioritized noise robustness and intonation awareness, while text models required context sensitivity and

the ability to handle short or informal inputs.

CNNs and transformer architecture were chosen for their suitability. CNNs excelled in processing mel-spectrograms, while models like HuBERT, DistilBERT, and RoBERTa effectively captured emotional patterns in raw audio and short texts.

Speech emotion recognition began with a CNN trained on mel-spectrograms from RAVDESS. It used frequency and timing patterns to classify emotions and was fine-tuned through extensive preprocessing and evaluation. HuBERT-based models improved generalization by learning from raw waveforms. Training on RAVDESS and TESS, these models bypassed manual feature extraction and aligned closely with HuBERT's strengths.

Text emotion recognition used both multi-class and two-stage classifiers. Standardized preprocessing and data balancing were applied to MELD, MC-EIU, and GoEmotions, and transformer models were trained with early stopping and scheduling to ensure stable performance.

3. APPROVALS AND RESULTS SECTION

3.1. Speech emotion recognition models. Results overview.

3.1.1. CNN

Analyzing the results of model training on the augmented data of the train set, we see that the highest model training accuracy achieved was 0.91 and the classification accuracy of the validation set reached the maximum result (1.0). Figure 3.1. shows that the training was interrupted at epoch 49.

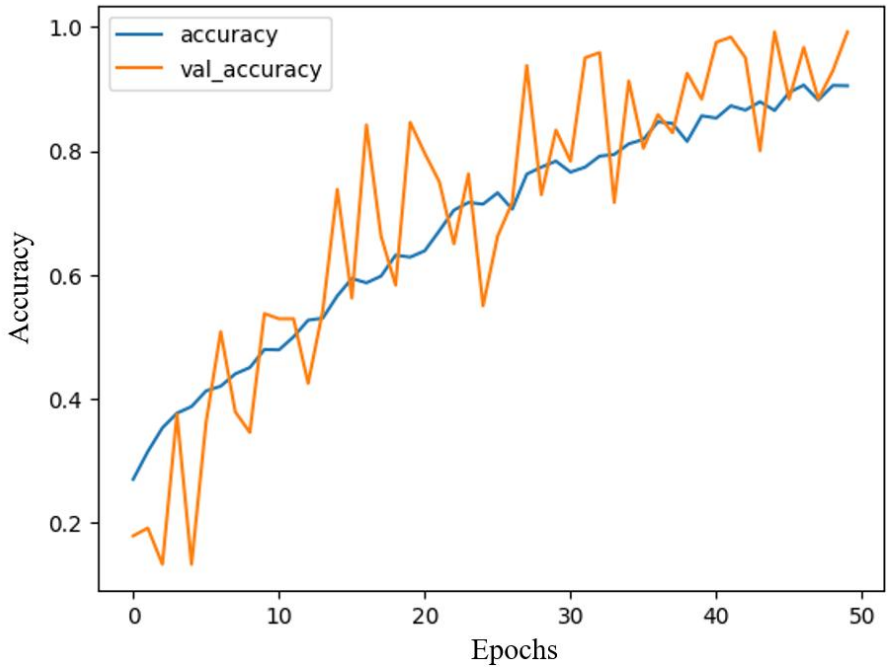

Fig. 3.1. The result of the model training accuracy on the set of augmented data at different epochs

Now let's look at the metrics shown in Figure 3.2. These are the results obtained by testing the model on the augmented data. As you can see, the overall accuracy achieved is 0.81. The trained model distinguishes well between the emotions of surprise, calmness, happiness, neutral, and disgust, but poorly between the emotions of fear, sadness, and anger. Notably, the highest F1-score was achieved for the “happy” and “surprised” classes, indicating that the model can confidently recognize these more distinct emotional expressions. On the other hand, “angry” and “sad” emotions yielded the lowest precision and F1-scores, suggesting confusion with semantically or acoustically similar categories. The relatively high recall for the “sad” class (0.92), despite its low precision, reflects a tendency of the model to overpredict sadness, which

may lead to false positives. The macro and weighted averages being close to the overall accuracy indicate a balanced performance across the dataset, although class imbalance remains a potential factor in the observed variations. These results emphasize the importance of refining feature representation and further augmenting underrepresented emotion classes.

	precision	recall	f1-score	support
calm	0.95	0.75	0.84	24
disgust	0.83	0.83	0.83	24
neutral	0.89	0.67	0.76	24
surprised	1.00	0.88	0.93	24
fearful	0.71	0.83	0.77	24
sad	0.61	0.92	0.73	12
angry	0.58	0.62	0.60	24
happy	0.92	1.00	0.96	24
accuracy			0.81	180
macro avg	0.81	0.81	0.80	180
weighted avg	0.83	0.81	0.81	180

Fig. 3.2. Metrics for testing a model trained on an aggregated dataset

The confusion matrix built for the model trained on the augmented dataset can be seen in Figure 3.3.

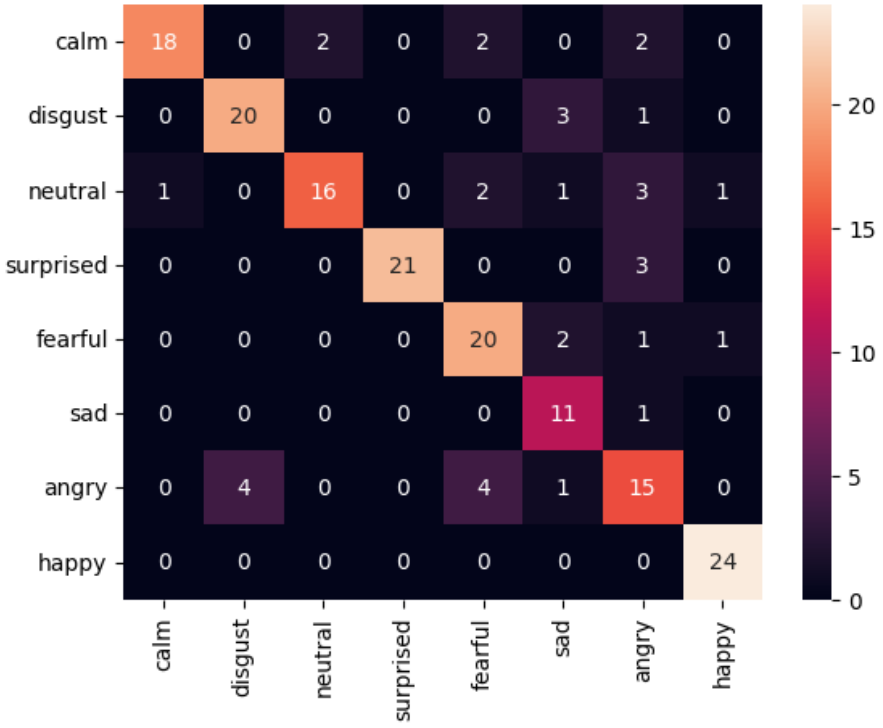

Fig. 3.3. Emotion recognition confusion matrix of the model trained on the augmented dataset

All metrics for this model can be viewed on Table 3.1.

Table 3.1. CNN model metrics

Train loss	Train accuracy	Validation loss	Validation accuracy	Test loss	Test accuracy
0.0716	91.1%	0.0009	100%	0.69	81.4%

In general, the model showed high accuracy in emotion classification, due to balanced data distribution, augmentation, and the use of deep architecture. The best results were achieved for emotions with distinct acoustic characteristics (such as joy, surprise, fear, and disgust), while neutral or similar-sounding emotions (such as neutral and sadness) could sometimes be confused. But due to high train and perfect validation accuracy it seems that model is overfitting on the data, or there are the same samples in these sets.

3.1.2. Transformer models (HuBERT)

3.1.2.1. RAVDESS

Results and evaluation of the model:

Training of the transformer model on the RAVDESS dataset showed a gradual improvement in accuracy over epochs. Figure 3.4. shows the metrics for testing the resulting model.

Fig. 3.4. Comparison of the training accuracy of the model trained on the RAVDESS set at each epoch

All metrics for this model can be viewed on Table 3.2.

Table 3.2. HuBERT/RAVDESS model metrics

Train loss	Train accuracy	Test loss	Test accuracy
0.774	80.5%	0.098	80%

At the first epoch, the accuracy was about 33.6%, and by the fifth epoch it increased to 80.5%, which indicates that the model was trained effectively. The overall accuracy of the model on the test set reached 80%, which is quite a good result for emotion classification in audio.

The confusion matrix of the model trained on the RAVDESS set is shown in Figure 3.5.

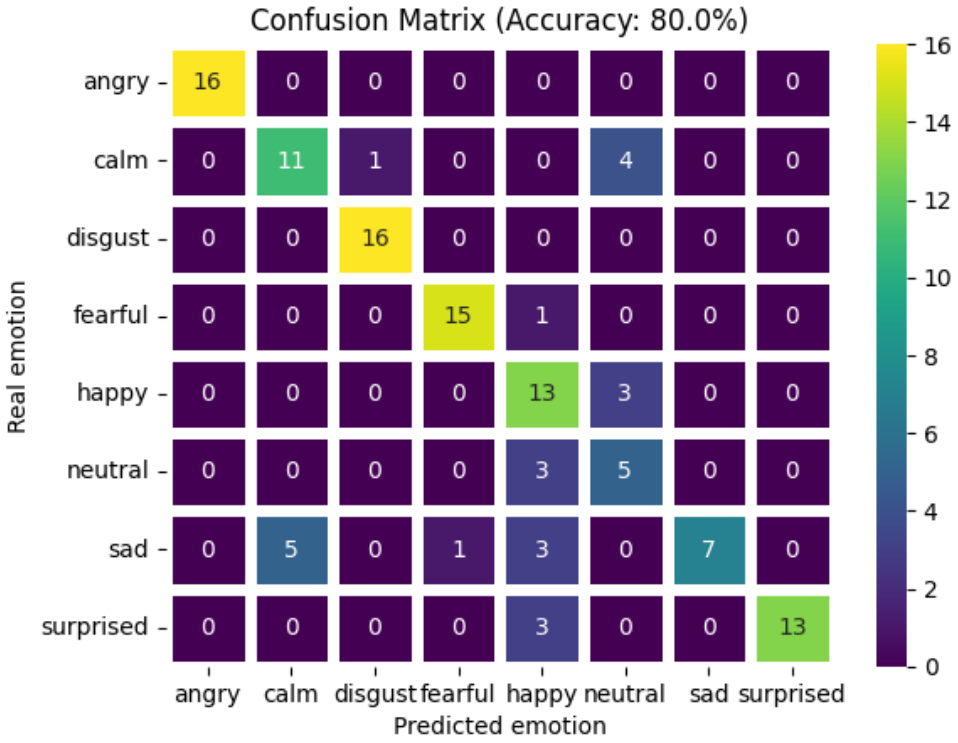

Fig. 3.5. Emotion recognition confusion matrix of the transformer model trained on the RAVDESS dataset

The analysis of this matrix shows that the model most accurately recognizes the emotions "angry", "disgust", "fearful", and "surprised", showing clear diagonal values (for example, 16 samples of "angry" and "disgust" were classified correctly). At the same time, the model is more likely to make mistakes when classifying the emotions "calm", "sad", and "neutral". The greatest confusion is observed between "calm" and

"sad", where several samples of one category are misclassified as another.

3.1.2.2. RAVDESS & TESS

Results and evaluation of the model:

The results of training the transformer model on the combined RAVDESS and TESS datasets are much better. Figure 3.6. shows the results of the model's classification accuracy for train and validation data at each epoch. As we can see, both train and validation accuracy increase with each epoch. From this dynamic, we can conclude that the model does not get too used to the training data.

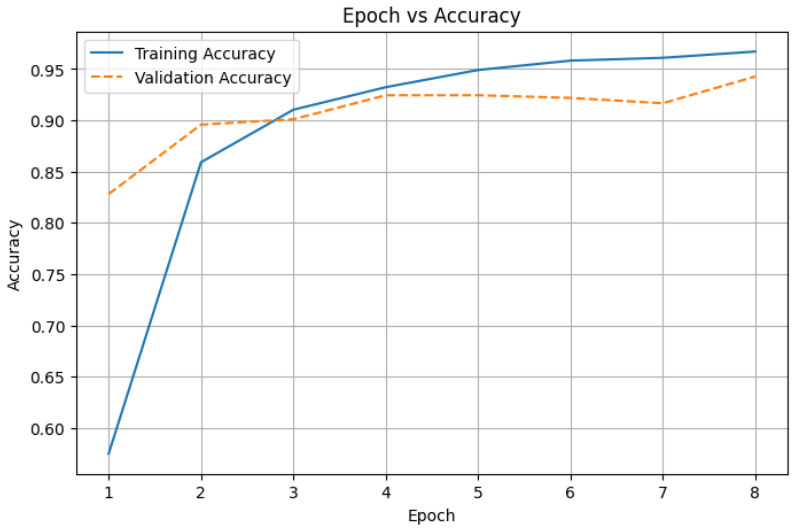

Fig. 3.6. Comparison of the train and validation accuracy of the model trained on the RAVDESS and TESS datasets at each epoch

All metrics for this model can be viewed on Table 3.3.

Table 3.3. HuBERT/RAVDESS&TESS model metrics

Train loss	Train accuracy	Validation loss	Validation accuracy	Test loss	Test accuracy
0.4315	96.7%	0.2533	94.3%	0.1290	96.4%

The training accuracy reached 96.7%, and the validation accuracy was 94.3%. On the test set, the model demonstrated an accuracy of 96.4% with a loss of 0.1290, which indicates a high level of generalization.

Figure 3.7. shows the model testing metrics after 8 epochs of training. This classification report confirms that the model recognizes all emotions well, including neutral, happiness, sadness, anger, fear, disgust, and surprise, with f1-scores in the

range of 0.92-0.99. The emotions "disgust" and "surprised" stand out with an f1-score of 0.99.

Classification Report:

	precision	recall	f1-score	support
neutral	0.89	1.00	0.94	48
happy	0.96	0.88	0.92	56
sad	1.00	0.91	0.95	56
angry	0.93	1.00	0.97	56
fear	0.97	1.00	0.98	56
disgust	1.00	0.98	0.99	56
surprised	1.00	0.98	0.99	56
accuracy			0.96	384
macro avg	0.96	0.96	0.96	384
weighted avg	0.97	0.96	0.96	384

Fig. 3.7. Metrics for testing the transformer model trained on the RAVDESS and TESS sets

The confusion matrix built for the resulting model can be seen in Figure 3.8.

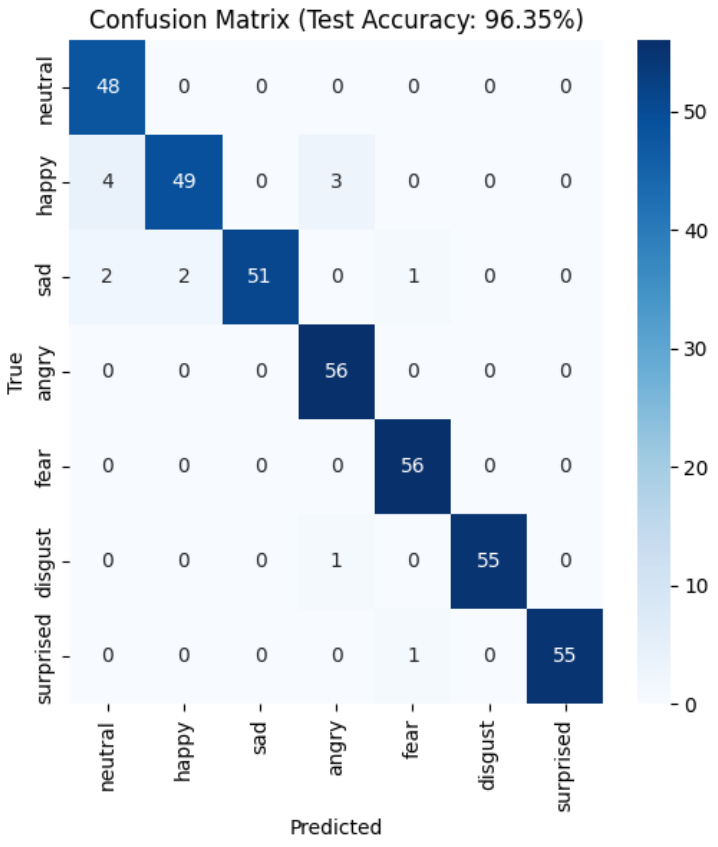

Fig. 3.8. Emotion recognition confusion matrix of the transformer model trained on the RAVDESS and TESS sets

Comparison of the results shows that the addition of the second TESS dataset has significantly improved the quality of emotion classification. Increasing the volume

and diversity of training data contributes to better generalization and accuracy. At the same time, the confusion matrix confirm that the main errors are mainly related to similar emotional states, which is quite typical for the emotion recognition task.

3.1.3. Comparison of results

Let's compare the results of training different models on the RAVDESS and TESS datasets using different approaches. The table below summarizes the key accuracy metrics for the CNN and transformer (HuBERT) models trained on both the RAVDESS and the combined RAVDESS and TESS datasets.

Table 3.4. Comparison of the results obtained.

Model	CNN	HuBERT	HuBERT
Dataset	RAVDESS + augmentation	RAVDESS	RAVDESS & TESS
Number of classes	8	8	7
Train accuracy	91%	81%	96.7%
Validation accuracy	100%	-	94.3%
Test accuracy	81%	80%	96.4%

As can be seen from the table, the HuBERT transformer model trained on the combined RAVDESS and TESS datasets demonstrated the best results in all key metrics, with the test accuracy reaching 96.4%. This demonstrates the model's high ability to generalize information and effectively recognize emotions in various audio data. The CNN model, which worked with the RAVDESS augmented set, also showed good results, including high training accuracy, but its testing accuracy was much lower. The HuBERT transformer model trained on RAVDESS alone performed mediocrely, indicating that it would benefit from the addition of additional data from the TESS set.

Thus, the use of combined and sufficiently diverse datasets together with modern

transformer architecture allowed us to achieve the highest accuracy of emotion classification.

Let’s now analyze the most common classification errors of the best-performing model, based on its confusion matrix (Fig. 3.8):

- Happy - Sad: The model occasionally confused happy with sad (4 records), and sad with happy (2 records). The reason for this is that these emotions might share similar acoustic patterns in moderate intonations, especially when not expressed with strong prosody.
- Happy - Angry: 3 examples of happy were classified as angry, possibly due to similarities in pitch or strong intensity in expressive speech (loud excitement vs. controlled anger).
- Sad - Neutral: 2 samples of sad were predicted as neutral, and 2 neutral as sad. This happened, as both emotions can be spoken in flat tone, normal intensity or with reduced prosodic variation.
- Disgust - Angry: 1 disgust sample was predicted as angry, as both these speech recordings can involve low-pitched expressions and quite strong intensity.
- Surprised - Fear: One surprised sample was confused with fear, which may occur due to similarities in high-pitch vocal bursts, strong intensity or breathiness.

Overall, the number of misclassifications is small and centered on classes with subtle acoustic overlaps, reflecting challenges even for human annotators. Nevertheless, these cases suggest opportunities for improvement, such as:

- Incorporating emotion intensity labels
- Fine-tuning with more ambiguous samples
- Using attention visualization for interpretability

3.2. Text emotion classification models. Results overview.

3.2.1. Implemented models results comparison

Since many models were trained, differing only in the type of transformer and the data used, it makes no sense to describe each of them in detail. Instead, we will

compare the results in the comparison tables.

Table 3.5 presents the performance of binary classification models trained on the MELD dataset using two transformer architectures. RoBERTa-base outperformed DistilBERT in both training and validation phases, suggesting stronger generalization during development. However, both models achieved nearly identical test accuracy, indicating that despite RoBERTa's advantage during training, the final performance on unseen data was comparable.

This outcome may imply that the additional complexity of RoBERTa does not provide a significant edge in generalizing to real-world data for this specific task. Furthermore, the marginal test accuracy difference hints that performance bottlenecks might stem more from the dataset characteristics than from model capacity alone.

Table 3.5. Comparison of the results of models trained for binary classification (neutral and all others as non-neutral).

Transformer	RoBERTa-base	DistilBERT
Used datasets	MELD	MELD
Additional actions with data	Normalization (4710 instances for each class)	Normalization (4710 instances for each class)
Train accuracy	92,5%	85,3%
Validation accuracy	80%	78,9%
Test accuracy	77%	77,5%

Table 3.6 shows the results of classifying six non-neutral emotions using DistilBERT with and without normalization. The overall performance across both setups was relatively low, with modest gains in test accuracy when normalization was applied. These outcomes suggest that distinguishing between multiple emotional

classes - especially in unbalanced or complex datasets - remains a challenging task even with transformer-based models.

Moreover, while normalization improved the balance of the training data, it did not lead to significant improvements on the validation set, indicating possible limitations in generalization. The noticeable gap between training and test performance further suggests the presence of overfitting, emphasizing the need for more advanced regularization or data augmentation techniques.

Table 3.6. Comparison of the results of models trained to classify 6 emotions (non-neutral).

Transformer	DistilBERT	DistilBERT
Used datasets	MELD & MC-EIU	MELD & MC-EIU
Additional actions with data	-	Normalization (599 copies for each class)
Train accuracy	70,23%	61,41%
Validation accuracy	45,6%	46,5%
Test accuracy	50,9%	51,8%

Table 3.7 compares models trained to classify seven emotions using different datasets. The model trained on GoEmotions achieved the highest test accuracy, demonstrating the benefit of using a large and diverse dataset. Models trained on MELD or MELD combined with MC-EIU performed worse, which highlights the impact of dataset selection and balance on multi-class emotion classification accuracy. The lower validation and test accuracies for the MELD and MELD & MC-EIU models indicate that even with higher training accuracy, the models struggled to generalize. This may suggest overfitting or insufficient diversity in the training data. Interestingly,

the model trained on the combined MELD & MC-EIU dataset showed the highest training accuracy, yet the lowest test performance, emphasizing that dataset size alone does not guarantee better generalization. The superior results of the GoEmotions-based model also point to the importance of dataset labeling quality and emotional granularity. Finally, these findings underscore the challenges in scaling emotion classification to more classes, particularly when working with imbalanced or context-sensitive datasets.

Table 3.7. Comparison of the results of models trained to classify 7 emotions.

Transformer	DistilBERT	DistilBERT	DistilBERT
Used datasets	MELD	MELD & MC-EIU	GoEmotions
Additional actions with data	-	-	Selecting random 5500 instances for a neutral emotion
Train accuracy	80,78%	89,68%	87,12%
Validation accuracy	55,3%	52,6%	61,6%
Test accuracy	58,6%	53,8%	65,1%

For the binary classification task (Table 3.5), the best results in validation were shown by the RoBERTa-base model, which achieved 80% accuracy, while the test accuracy was 77%. The DistilBERT-based model was inferior in both training and validation, although it showed slightly higher accuracy on the test set is 77.5%.

In the case of classifying the six emotions (Table 3.6), the models performed poorly overall. Although data normalization improved the results on the test set (up to 51.8%), the overall accuracy level remained low, which indicates the complexity of the task in a multi-class environment.

Finally, for the classification of seven emotions (Table 3.7), the model trained on the GoEmotions dataset using DistilBERT showed the best test accuracy of 65.1%. This outperforms the results of models trained on MELD or its combination with MC-EIU.

As a result, this latter model was chosen for further use, as it provided the highest accuracy in classifying emotions from text among all tested options.

All metrics for this model can be viewed on Table 3.8.

Table 3.8. DistilBERT/GoEmotions model metrics

Train loss	Train accuracy	Validation loss	Validation accuracy	Test accuracy
0.341	87.12%	1.4669	61.6%	65.1%

The train accuracy reached 87.12%, and the validation accuracy was 61.6%. On the test set, the model achieved an accuracy of 65.1%, which indicates a notable generalization gap and highlights the difficulty of accurately classifying emotions in short, context-limited text samples.

Figure 3.9. shows the results of the model's classification accuracy for train and validation data at each epoch.

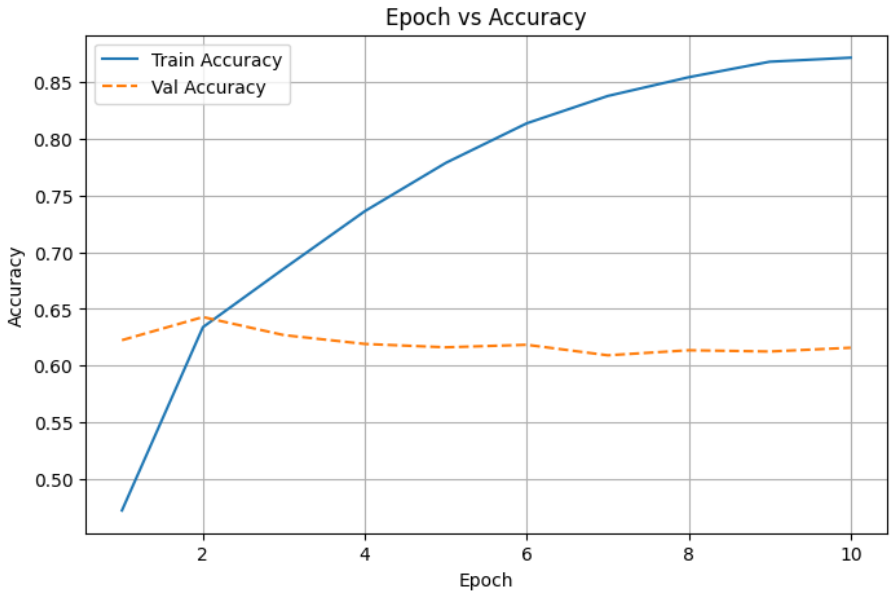

Fig. 3.9. Comparison of the train and validation accuracy of the DistillBERT model trained on GoEmotions dataset

Figure 3.10. shows the model metrics after 10 epochs of training.

Classification Report:

	precision	recall	f1-score	support
neutral	0.63	0.50	0.55	550
happy/joy	0.76	0.79	0.78	433
anger	0.63	0.68	0.65	521
sad/sadness	0.66	0.69	0.68	383
surprise	0.67	0.70	0.69	347
fear	0.66	0.72	0.69	178
disgust	0.50	0.53	0.52	291
accuracy			0.65	2703
macro avg	0.65	0.66	0.65	2703
weighted avg	0.65	0.65	0.65	2703

Fig. 3.10. Metrics for testing the DistillBERT/GoEmotions model

The model achieved a weighted average F1-score of 0.65 across all seven emotion classes, indicating moderate overall performance. It performed best on the happy/joy class with an F1-score of 0.78, showing strong precision (0.76) and recall (0.79). Emotions such as disgust and neutral were more challenging, with lower F1-scores of 0.52 and 0.55, respectively. The macro average values (around 0.65–0.66) suggest balanced performance across classes despite variations in support. This implies the model handles dominant and underrepresented classes with similar efficiency, though certain emotional nuances remain difficult to capture.

The confusion matrix built for the resulting model can be seen in Figure 3.11.

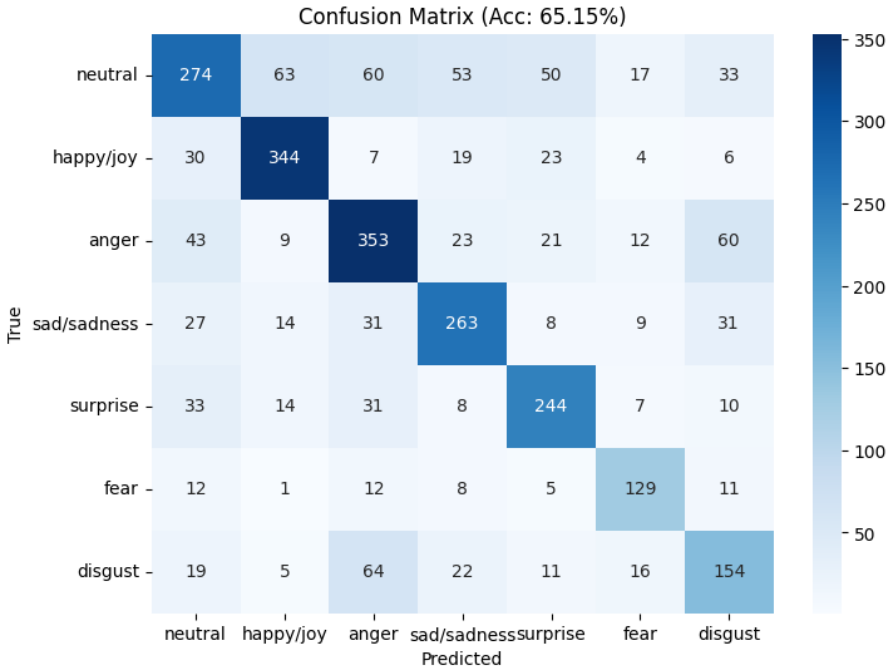

Fig. 3.11. Emotion recognition confusion matrix of the DistillBERT model trained on the GoEmotions dataset

Let’s now analyze the most common classification errors of this model, based on its confusion matrix (Fig. 3.11):

Despite being the best performer for the seven-class task, the model frequently confused semantically or tonally similar emotions. Some of the most often error trends include:

- Neutral misclassifications: The neutral class was often misclassified as happy, sad, or surprise, and conversely. This likely stems from the subtle emotional cues in text that can resemble neutrality or be heavily context dependent.
- Anger - Disgust: A notable number of instances of anger and disgust were confused with each other. These emotions share overlapping lexical cues (negative or intense adjectives), which can easily mislead the model.
- Sadness confusion: Sad/sadness was frequently misclassified as anger or neutral, indicating that sadness-related language lacks distinct markers in short text inputs or gets overridden by stronger negative sentiment terms.
- Surprise entanglement: The surprise category overlapped with almost all others, especially neutral, anger, and happy. Without sufficient conversational or situational context, it's difficult for the model to recognize surprise reliably.
- Fear and disgust overlap: These lower-represented categories had many scattered predictions, indicating weak representation in the training data or ambiguity in textual patterns.

A likely reason for these widespread misclassifications is the lack of conversational or situational context in the dataset. Since many inputs are short and decontextualized, the model might rely heavily on emotionally charged keywords. For example, sentences with words like "angry" or "furious" might be classified as anger, while phrases with words like "stinky" or "gross" could trigger disgust, even when the actual emotion differs.

These patterns suggest that future improvements could involve:

- Incorporating longer or multi-turn context (e.g., previous utterances)

- Leveraging attention heatmaps to identify misleading tokens

3.3. Combining results

After pre-processing the audio and text data and classifying emotions by separate models, we implemented a late fusion process. This approach allows us to combine the predictions of two independent models - one that analyzes audio and the other that works with text - and form a consistent emotional interpretation of the input fragment.

Since the audio and text were divided synchronously into passages of the same content during the pre-processing stage, when combining the results, it is enough to simply compare the corresponding elements from both arrays of emotional labels. If both models predict the same emotion, it is considered final. In case of discrepancies, the emotion from the audio is marked as less accurate, and the result of the text model is presented as recommended for interpretation.

The flow of the data can be seen on Figure 3.12, where A stands for audio recording, a_i - audio chunk, t_i - text chunk, l_i - emotion label of the i-th chunk, e_i - emotion of the i-th chunk.

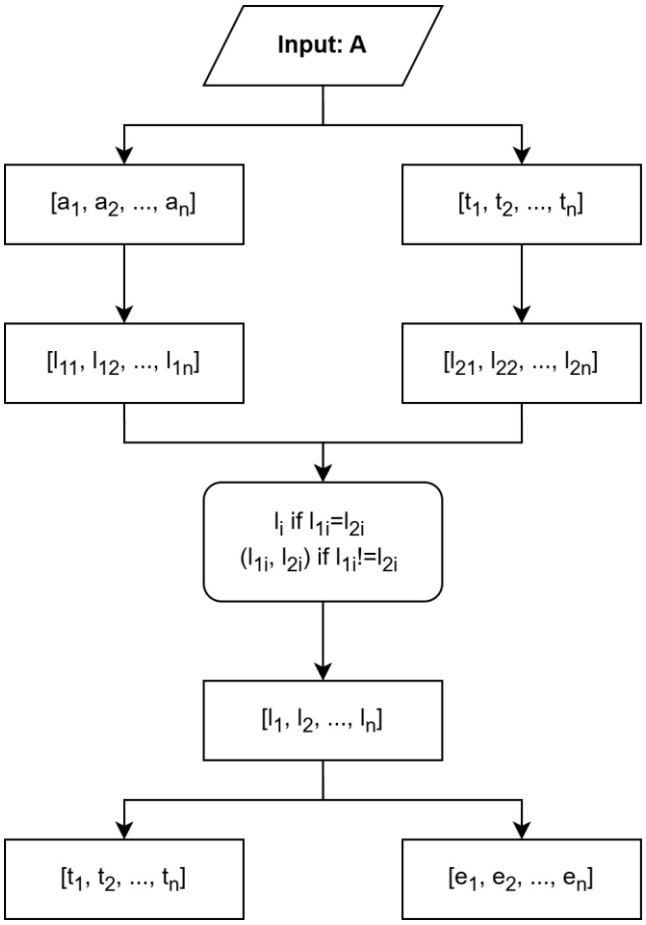

Fig. 3.12. Flow of the data

3.4. Architecture and logic of the system as a whole

The Gradio framework was used to implement a project that could demonstrate the work of both models.

3.4.1. System structure

The system is implemented in the form of a sequential processing chain, each stage of which performs a separate function of converting the input audio file to the result. The general structure is presented in the form of a flowchart (Fig. 3.13), which illustrates the sequence of interaction of the main components of the system.

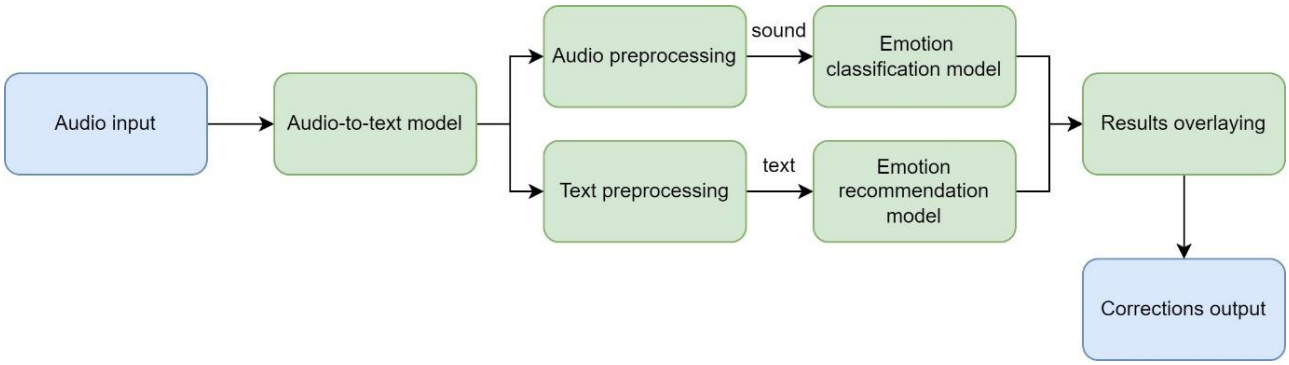

Fig. 3.13. Diagram of the system logic

After receiving the input audio, the pre-processing module is launched, which is responsible for the coordinated formation of text and audio fragments. For this stage, we used a ready-made audio-to-text model from OpenAI - Whisper. Next, the generated chunks are sent to separate classification models - one for audio and the other for text. Each model works independently, returning its own emotional state estimates.

The final stage is the module for comparing and matching results, where emotional labels obtained from different modalities are compared. At this level, the logic of the final choice or correction of emotion is implemented. Thanks to the previously agreed division of fragments, the classification results are easily combined, after which the final representation is generated - a text with the appropriate emotional markup, ready to be displayed to the user.

3.4.2. Interaction between modules

The system's topology has a linear-modular structure, which provides for the sequential passage of incoming data through a set of logically linked processing stages.

The system components are organized in the form of clearly distinguished modules, each of which is responsible for a specific stage of information processing or analysis.

In general, the system implements a sequential processing topology in which data passes through the stages one by one. Although the system has two classifiers - for audio and text - they do not work in parallel at the back-end level. First, pre-processing takes place, and then the models are run one by one.

Main components and their interconnection:

- Input module:
  - Receives an audio file from the user and initiates processing.
- Pre-processing module (Whisper + segmentation):
  - Performs speech transcription and divides audio/text into fragments.
- Module for recognizing emotions from audio:
  - Processes audio chunks and generates a set of emotion labels.
- Emotion Recognition Module for Text:
  - Processes text chunks and generates a set of emotion labels.
- Fusion module:
  - Compares emotions from both sources.
  - Generates the final analysis result according to predefined logic: if the emotions match, they are accepted as correct, if they differ, the result of the audio model is marked as such that needs to be corrected by the user in the speech.
- Output module (interface):
  - Displays text with appropriate emotional labels.

This approach to topology allows us to control each stage of processing and simplifies system debugging and scaling. In the future, individual modules can be optimized or implemented in parallel without changing the overall architectural scheme.

To complement the description of the system’s architecture and clarify the interaction between its modules, a BPMN diagram has been developed:

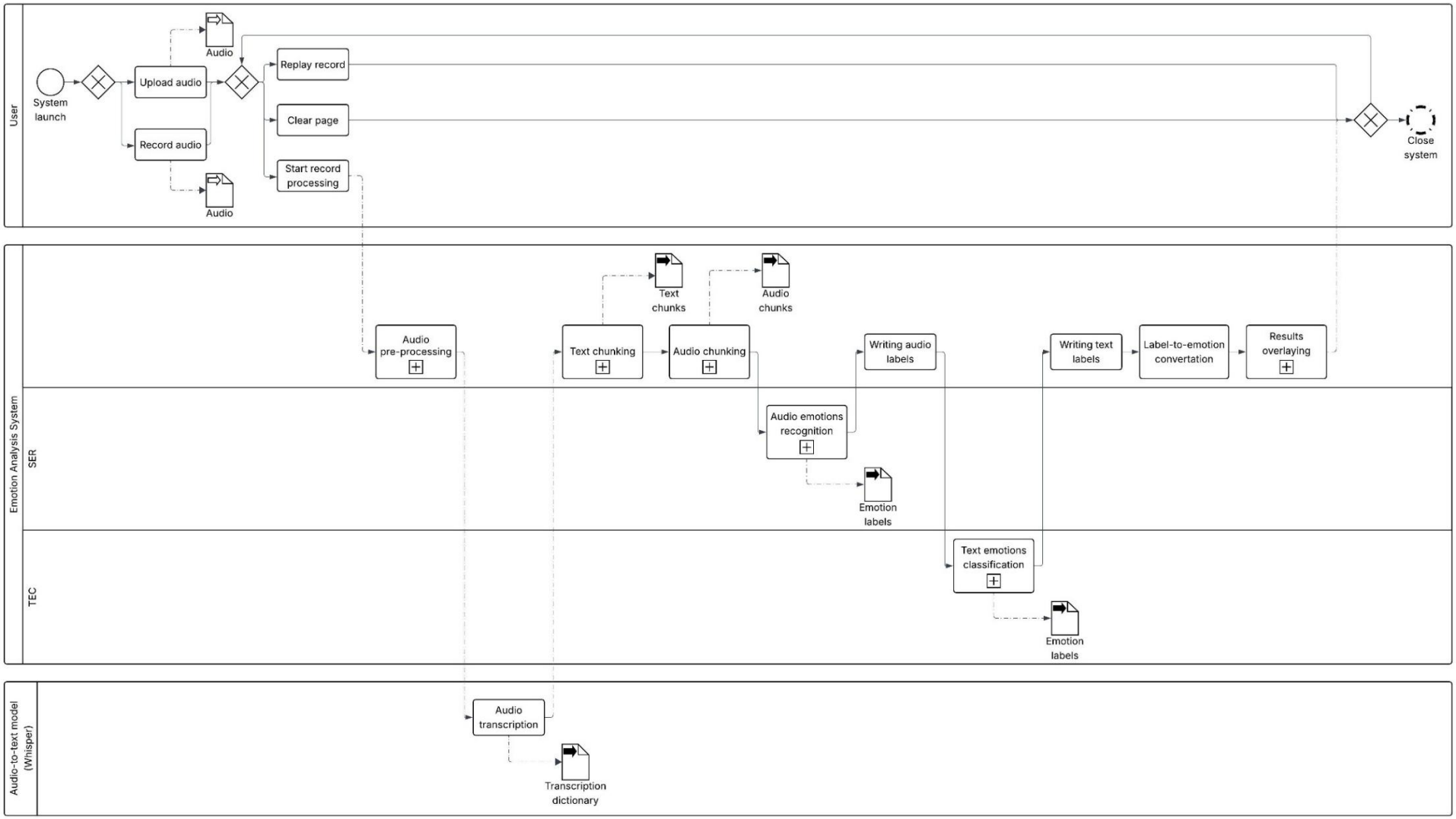

Fig. 3.14. System workflow BPMN diagram

3.4.3. Overview of the system interface

Let's move on to the interface of the implemented system. The user can upload an existing audio file or make a recording manually on the page. The uploaded recording can be re-listened to. After the user clicks the "Process and show results" button, the emotion analysis system starts working. The result looks like this: chunks of text extracted from the audio, and emotions above them. If the emotion recognized from the voice and the emotion predicted from the text are the same, one emotion will be displayed above the text passage, if not, 2 emotions will be displayed as a recommendation to change the recognized emotion to the recommended one.

You can see the user interface of the system in Appendix C.

3.5. Hardware and software requirements

The Google Colab environment was used to accomplish the tasks. A GPU T4 graphics processor was used to train the models, which ensured faster computing and more efficient processing of audio and text data.

Requirements for Gradio project:

- Python 3.11.4
- Download libraries from requirements.txt (The link of the repository is on the Appendix A)

3.6. Conclusions

In this section, the conducted experiments and implementation steps have demonstrated the feasibility and effectiveness of multimodal emotion recognition using both audio and textual inputs. By training and comparing different models, integrating their predictions, and deploying them in a functional interface, the system has been designed to operate robustly and adaptively. The inclusion of diverse datasets, normalization techniques, and fusion strategies has significantly contributed to the overall performance. The combination of deep learning and pre-trained transformer models allowed the system to capture emotional nuances with high precision across different modalities.

Speech Emotion Recognition:

The results from training CNN and HuBERT-based models on speech data

revealed that the choice of architecture and training data directly influenced performance. The CNN achieved a training accuracy of 91% and a perfect validation score but dropped to 81% on the test set, indicating potential overfitting. Meanwhile, the HuBERT transformer, when trained solely on RAVDESS, exhibited moderate results, achieving 80% accuracy. However, when the TESS dataset was added, the model achieved a significant improvement, with test accuracy reaching 96.4%. These outcomes have confirmed that data diversity and transformer-based modeling strongly benefit speech emotion classification tasks.

Text Emotion Classification:

Evaluation of text-based emotion classification models has shown that while transformer architectures such as RoBERTa-base and DistilBERT can generalize well, their success heavily depends on dataset selection and preprocessing. RoBERTa outperformed DistilBERT during training and validation, though both reached comparable test accuracies around 77%. For multi-class tasks, performance was generally lower, with models struggling to distinguish between non-neutral emotions. The best performance, 65.1% test accuracy, was observed with a DistilBERT model trained on the GoEmotions dataset, emphasizing the advantage of using large and balanced datasets. Ultimately, this model was selected for integration due to its superior performance in classifying diverse emotions in textual data.

The implemented fusion strategy helped leverage the strengths of both audio and text models to generate a more accurate emotional prediction. During testing, it became evident that the synchronization of segments from both modalities was effective for comparison. Whenever both models agreed, the emotion was accepted as final; when they differed, the text model's output was prioritized. This method provided a simple yet functional solution for handling discrepancies, and future refinements may incorporate confidence scores or weighted decisions. The late fusion approach proved to be a reliable mechanism for harmonizing multimodal predictions into a coherent interpretation.

The final emotion recognition system was designed using a modular and scalable architecture. Each processing stage - from audio input to final emotion labeling - was logically and independently developed to allow flexibility and maintainability.

Whisper was employed to transcribe audio into text, enabling parallel analysis by the two models. Upon receiving outputs from both models, the system compared and merged predictions before presenting the final emotional markup. The entire system was deployed using the Gradio framework, allowing for real-time user interaction. This architecture not only facilitated testing and debugging but also laid a foundation for further integration of other modalities or user-specific customization.

Each module of the system was developed to fulfill a specific role, forming a linear pipeline where data flowed smoothly from one stage to the next. The modular design supported independent development and testing, improving fault tolerance and scalability. Audio and text were processed in parallel by their respective models, and the merging module reconciled their outputs with minimal latency. These design choices ensured that the system could be maintained and upgraded efficiently. The interactive interface built on Gradio allowed users to upload audio and receive emotion-labeled text in real time, demonstrating the successful integration of all components into a unified, usable product.

CONCLUSIONS

While completing this bachelor's thesis on voice emotion recognition using machine learning methods, the following tasks were accomplished:

- Various approaches to creating systems for analyzing emotions in speech and text were investigated and analyzed.

  A review of current approaches demonstrated the growing role of deep learning models in emotion recognition, especially transformer-based architectures in both speech and natural language processing tasks. This theoretical groundwork informed the practical design decisions made in the project.

- Appropriate datasets for model training were selected: RAVDESS and TESS for voice data, MC-EIU, MELD, and GoEmotions for text data.

  The datasets were chosen based on their diversity, annotation quality, and relevance to real-world emotional expression. This selection provided a balanced foundation for training models capable of generalizing well across various modalities and emotional categories.

- A model for classifying emotions based on acoustic characteristics of the voice, including CNN architecture and the modern HuBERT transformer model, was implemented.

  Several models were tested, and the HuBERT model outperformed others, achieving the best results. Trained on the combined RAVDESS and TESS datasets, HuBERT reached a classification accuracy of 96.4%. This confirms the effectiveness of self-supervised transformer-based models for capturing subtle emotional cues in vocal signals.

- A model for predicting the appropriate emotion, considering the content of the text, using the transformer architectures DistilBERT and RoBERTa-base, was also developed.

  Although text-based models did not perform as well as the voice-based model, they still achieved reasonable results. The best performance came from DistilBERT trained on the GoEmotions dataset, with an accuracy of

65.1%. These results show the challenge of textual emotion classification, particularly due to the subjectivity and ambiguity in interpreting emotion from written language.

- A method for combining the results of both models to improve the overall quality of speech emotion analysis was implemented.

  The fusion of voice and text emotion outputs was explored to enhance prediction reliability. By combining insights from both modalities, the system can compensate for limitations in either modality alone, resulting in more accurate and robust emotional assessments.

- A full-fledged interactive system was implemented that allows analyzing the emotions of a user's speech based on their audio and, optionally, on the corresponding text.

  The final prototype integrates both models into a user-friendly web interface developed using the Gradio framework. Users can upload or record audio, view detected emotions from both modalities, and receive an overall interpretation. This serves as a proof-of-concept demonstrating how complex machine learning systems could potentially be translated into accessible tools, pending further validation.

The work illustrates the feasibility of developing such systems within the scope of undergraduate research and provides a foundation for future development and validation efforts. Potential future directions include:

- adapting the system to Ukrainian-language data;
- improving audio chunking and merging algorithms;
- expanding the emotion set and addressing class imbalance;
- optimizing performance using GPU acceleration;
- enabling real-time or parallelized processing;
- exploring practical applicability in domains like education, psychological assistance, or smart assistants, subject to further validation and user studies.

REFERENCES

1. Akshay S. Utane, S. L. Nalbalwar. EMOTION RECOGNITION through SPEECH. 2nd National Conference on Innovative Paradigms in Engineering and Technology (NCIPET 2013). NCIPET, 1 (November 2013), 0-0.
2. Alnuaim AA, Zakariah M, Shukla PK, Alhadlaq A, Hatamleh WA, Tarazi H, Sureshbabu R, Ratna R. Human-Computer Interaction for Recognizing Speech Emotions Using Multilayer Perceptron Classifier. J Healthc Eng. 2022 Mar 28;2022:6005446. doi: 10.1155/2022/6005446. PMID: 35388315; PMCID: PMC8979705.
3. Anna Koufakou, Jairo Garciga, Adam Paul, Joseph Morelli and Christopher Frank. Automatically Classifying Emotions based on Text: A Comparative Exploration of Different Datasets. 2023. https://doi.org/10.48550/arXiv.2302.14727
4. Ayadi, M.M., Kamel, M.S., & Karray, F. Survey on speech emotion recognition: Features, classification schemes, and databases. 2011. Pattern Recognit., 44, 572-587.
5. Daniel Yang, Aditya Kommineni, Mohammad Alshehri, Nilamadhab Mohanty, Vedant Modi, Jonathan Gratch and Shrikanth Narayanan. Context Unlocks Emotions: Text-based Emotion Classification Dataset Auditing with Large Language Models. 2023. https://doi.org/10.48550/arXiv.2311.03551
6. Deriche, M., Abo absa, A.H. “A Two-Stage Hierarchical Bilingual Emotion Recognition System Using a Hidden Markov Model and Neural Networks”. Arab J Sci Eng 42, 5231–5249 (2017). https://doi.org/10.1007/s13369-017-2742-5.
7. Distilbert-base-uncased-emotion. https://huggingface.co/bhadresh-savani/distilbert-base-uncased-emotion
8. Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, Sujith Ravi. GoEmotions: A Dataset of Fine-Grained

Emotions. 2020. https://doi.org/10.48550/arXiv.2005.00547.

9. Ekman, P. (1992). An Argument for Basic Emotions. Cognition and Emotion, 6(3-4), 169–200. https://doi.org/10.1080/02699939208411068
10. Fan, Weiquan & Xu, Xiangmin & Xing, Xiaofen & Chen, Weidong & Huang, Dongyan. (2021). LSSED: a large-scale dataset and benchmark for speech emotion recognition.
11. Jiaxin Shi, Mingyue Xiang. Convolution SSM model for text emotion classification. 2024. Proceedings Volume 13210, Third International Symposium on Computer Applications and Information Systems (ISCAIS 2024); 1321021 (2024) https://doi.org/10.1117/12.3034918
12. Kamińska, D., Sapiński, T. & Anbarjafari, G. Efficiency of chosen speech descriptors in relation to emotion recognition. J AUDIO SPEECH MUSIC PROC. 2017, 3 (2017). https://doi.org/10.1186/s13636-017-0100-x.
13. Kerkeni, Leila & Serrestou, Youssef & Raoof, Kosai & Cléder, Catherine & Mahjoub, Mohamed & Mbarki, Mohamed. (2019). Automatic Speech Emotion Recognition Using Machine Learning. 10.5772/intechopen.84856.
14. Kerkeni, L.; Serrestou, Y.; Mbarki, M.; Raoof, K. and Mahjoub, M. (2018). Speech Emotion Recognition: Methods and Cases Study. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence - Volume 1: ICAART; ISBN 978-989-758-275-2; ISSN 2184-433X, SciTePress, pages 175-182. DOI: 10.5220/0006611601750182.
15. Lanjewar, Rahul & Mathurkar, Swarup & Patel, Nilesh. (2015). Implementation and Comparison of Speech Emotion Recognition System Using Gaussian Mixture Model (GMM) and K- Nearest Neighbor (K-NN) Techniques. Procedia Computer Science. 49. 50-57. 10.1016/j.procs.2015.04.226.
16. Lieskovská, E.; Jakubec, M.; Jarina, R.; Chmulík, M. A Review on Speech Emotion Recognition Using Deep Learning and Attention Mechanism. Electronics 2021, 10, 1163. https://doi.org/10.3390/electronics10101163.

17. Livingstone SR, Russo FA (2018) The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. PLoS ONE 13(5): e0196391. https://doi.org/10.1371/journal.pone.0196391.
18. Manas Jain, Shruthi Narayan, Pratibha Balaji, Bharath K P, Abhijit Bhowmick, Karthik R and Rajesh Kumar Muthu, “Speech Emotion Recognition using Support Vector Machine”, 2020.
19. Mirsamadi, Seyedmahdad & Barsoum, Emad & Zhang, Cha. (2017). Automatic Speech Emotion Recognition Using Recurrent Neural Networks with Local Attention. 10.1109/ICASSP.2017.7952552.
20. Mohammad, S. M. (2012). #Emotional Tweets. Proceedings of the First Joint Conference on Lexical and Computational Semantics, 246–255. https://doi.org/10.3115/v1/S12-1025
21. Monkam, Patrice & Qi, Shouliang & Ma, He & Gao, Weiming & Yao, Yudong & Qian, Wei. (2019). Detection and Classification of Pulmonary Nodules Using Convolutional Neural Networks: A Survey. IEEE Access. PP. 1-1. 10.1109/ACCESS.2019.2920980.
22. Mustaqeem; Kwon, S. A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition. Sensors 2020, 20, 183. https://doi.org/10.3390/s20010183.
23. Niko Laskaris. How to apply machine learning and deep learning methods to audio analysis. 2019.
24. Olga Chernytska. Complete Guide to Data Augmentation for Computer Vision. 2021. Towards Data Science.
25. Pichora-Fuller, M. Kathleen, Dupuis, Kate. Toronto emotional speech set (TESS). 2020. doi/10.5683/SP2/E8H2MF
26. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. Information Fusion, 37, 98–125. https://doi.org/10.1016/j.inffus.2017.02.003
27. Rafiul Islam, Md. Taimur Ahad, Faruk Ahmed, Bo Song, Yan Li. Mental

Health Diagnosis From Voice Data Using Convolutional Neural Networks and Vision Transformers. 2024. https://doi.org/10.1016/j.jvoice.2024.10.010.

28. Rui Liu, Haolin Zuo, Zheng Lian, Xiaofen Xing, Björn W. Schuller, Haizhou Li. Emotion and Intent Joint Understanding in Multimodal Conversation: A Benchmarking Dataset. 2024. https://doi.org/10.48550/arXiv.2407.0275.
29. Samaneh Madanian, David Parry, Olayinka Adeleye, Christian Poellabauer, Farhaan Mirza, Shilpa Mathew, Sandy Schneider. Automatic Speech Emotion Recognition Using Machine Learning: Digital Transformation of Mental Health. 2022. https://aisel.aisnet.org/pacis2022/45
30. Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, Sandra L. Schneider. Speech emotion recognition using machine learning — A systematic review. Intelligent Systems with Applications, Volume 20. 2023. https://doi.org/10.1016/j.iswa.2023.200266.
31. Santosh Kumar Bharti, S Varadhaganapathy, Rajeev Kumar Gupta, Prashant Kumar Shukla, Mohamed Bouye, Simon Karanja Hingaa and Amena Mahmoud. Text-Based Emotion Recognition Using Deep Learning Approach. 2022. Computational Intelligence and Neuroscience, Volume 2022, Article ID 2645381, 8 pages. https://doi.org/10.1155/2022/2645381
32. Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. 2019. https://doi.org/10.48550/arXiv.1810.02508\.
33. T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi and E. Ambikairajah, "A Comprehensive Review of Speech Emotion Recognition Systems," in IEEE Access, vol. 9, pp. 47795-47814, 2021, doi: 10.1109/ACCESS.2021.3068045.

34. Twitter-roBERTa-base for Emotion Recognition. https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion
35. Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2019. https://doi.org/10.48550/arXiv.1910.01108
36. Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, Abdelrahman Mohamed. HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units. 2021. https://doi.org/10.48550/arXiv.2106.07447
37. Y. Xi, P. Li, Y. Song, Y. Jiang and L. Dai, "Speaker to Emotion: Domain Adaptation for Speech Emotion Recognition with Residual Adapters," 2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Lanzhou, China, 2019, pp. 513-518, doi: 10.1109/APSIPAASC47483.2019.9023339.
38. Zwol, Björn & Langezaal, Mathijs & Arts, Lukas & Gatt, Albert & van den Broek, Egon L.. (2023). Speech Emotion Recognition Using Deep Convolutional Neural Networks Improved by the Fast Continuous Wavelet Transform. 10.3233/AISE230012.

Appendix A.

Link to the system code repository:

khrystynadol/diploma_gradio_app

Link to the repository with the code used in the research:

khrystynadol/diploma_source_code

Appendix B.

Visualization of architecture of the CNN model for the SER task using the model.summary command:

```
Model: "model_melspec"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_1 (InputLayer)        [(None, 60, 94, 1)]       0
 conv2d (Conv2D)             (None, 60, 94, 32)        1312
 batch_normalization (BatchNormalization)  (None, 60, 94, 32)  128
 activation (Activation)     (None, 60, 94, 32)        0
 max_pooling2d (MaxPooling2D)  (None, 30, 47, 32)      0
 dropout (Dropout)           (None, 30, 47, 32)        0
 conv2d_1 (Conv2D)           (None, 30, 47, 32)        40992
 batch_normalization_1 (BatchNormalization)  (None, 30, 47, 32)  128
 activation_1 (Activation)   (None, 30, 47, 32)        0
 max_pooling2d_1 (MaxPooling2D)  (None, 15, 23, 32)    0
 dropout_1 (Dropout)         (None, 15, 23, 32)        0
 conv2d_2 (Conv2D)           (None, 15, 23, 32)        40992
 batch_normalization_2 (BatchNormalization)  (None, 15, 23, 32)  128
 activation_2 (Activation)   (None, 15, 23, 32)        0
 max_pooling2d_2 (MaxPooling2D)  (None, 7, 11, 32)     0
 dropout_2 (Dropout)         (None, 7, 11, 32)         0
 conv2d_3 (Conv2D)           (None, 7, 11, 32)         40992
 batch_normalization_3 (BatchNormalization)  (None, 7, 11, 32)  128
 activation_3 (Activation)   (None, 7, 11, 32)         0
 max_pooling2d_3 (MaxPooling2D)  (None, 3, 5, 32)      0
 dropout_3 (Dropout)         (None, 3, 5, 32)          0
 flatten (Flatten)           (None, 480)               0
 dense (Dense)               (None, 64)                30784
 dense_1 (Dense)             (None, 256)               16640
 emotion_output (Dense)      (None, 8)                 2056
=================================================================
Total params: 174280 (680.78 KB)
Trainable params: 174024 (679.78 KB)
Non-trainable params: 256 (1.00 KB)
_________________________________________________________________
```

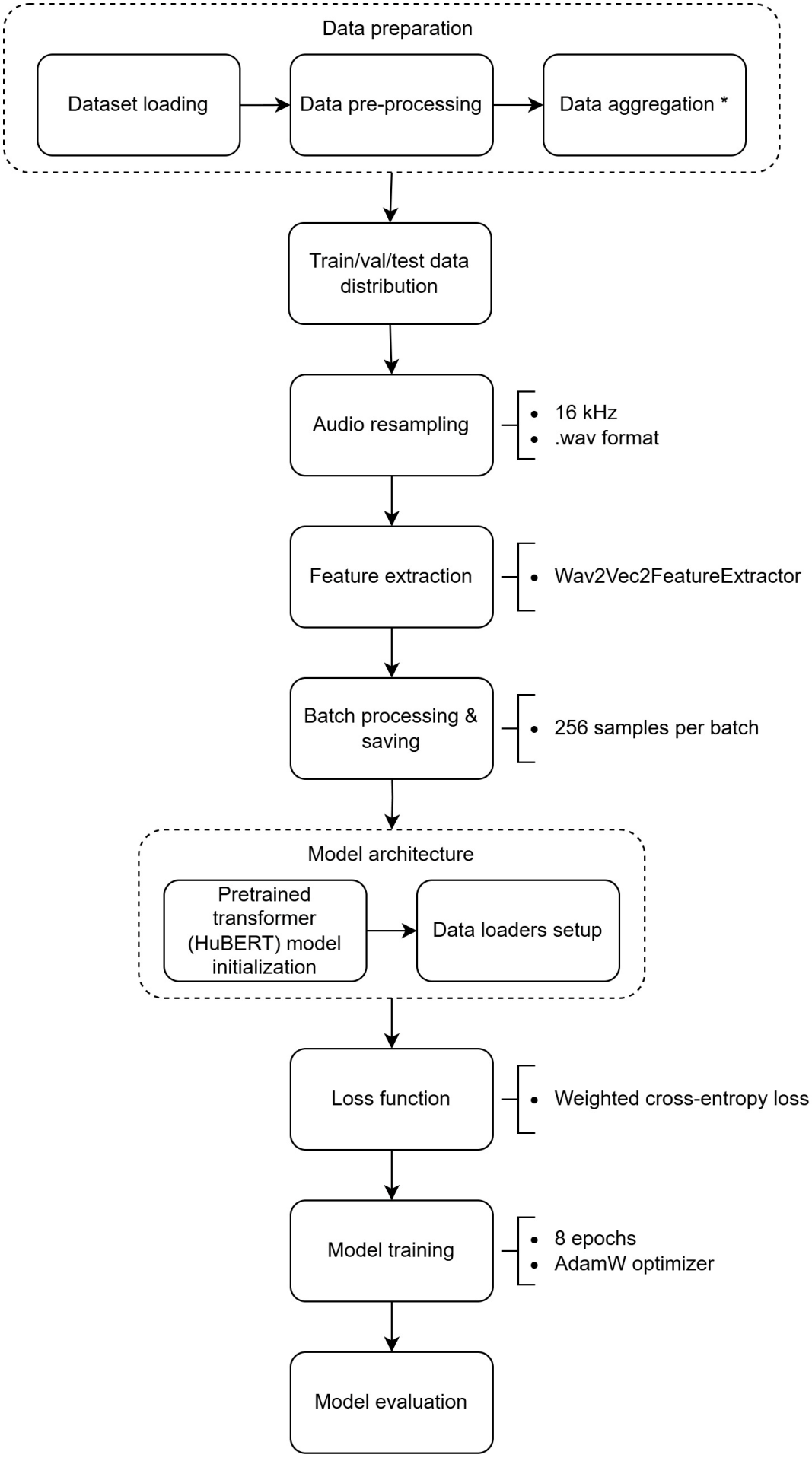

Fig. B.1. SER HuBERT transformer model flow

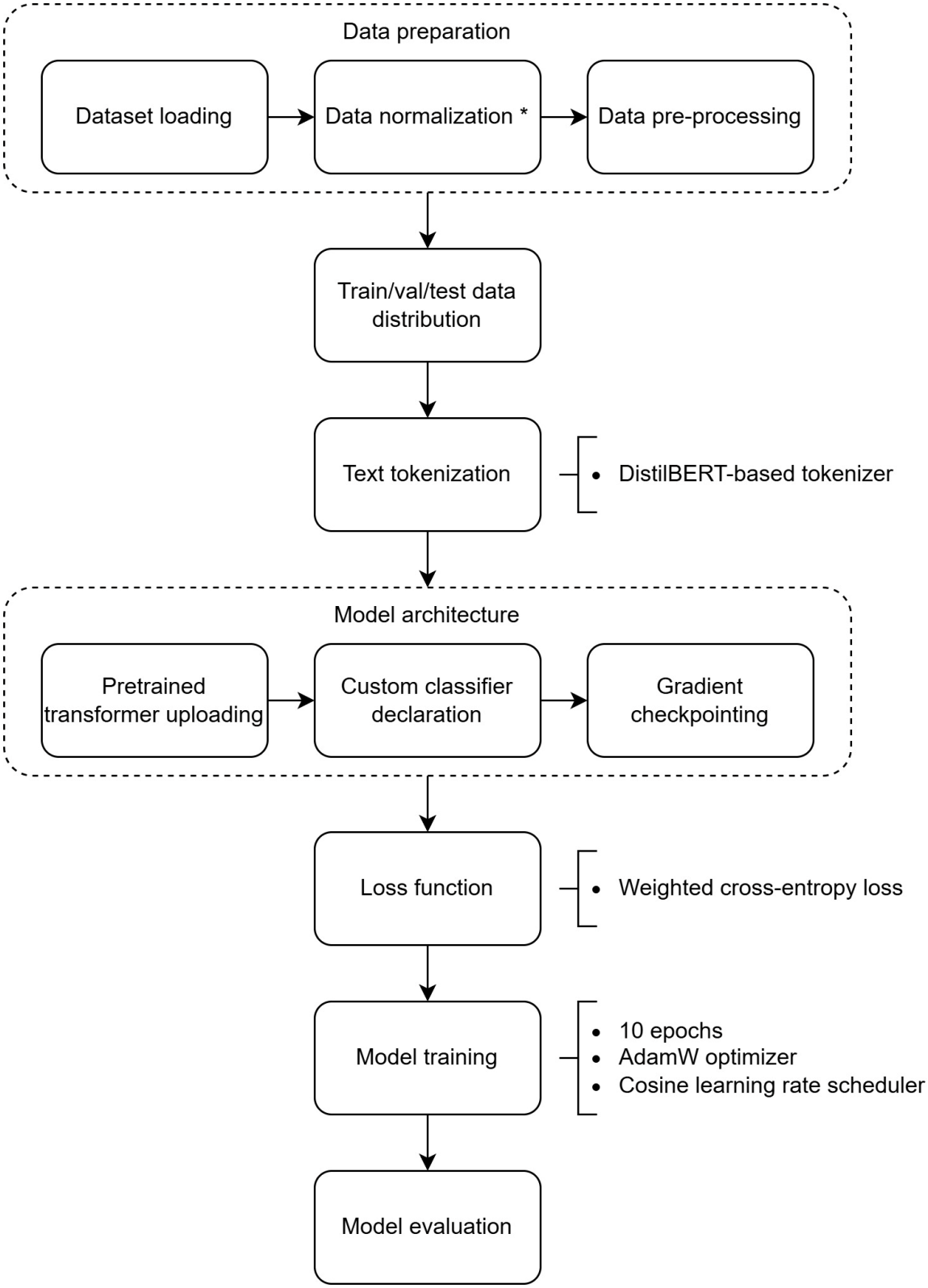

Fig. B.2. TEC transformer model flow

Appendix C.

The interface of the developed project looks like this:

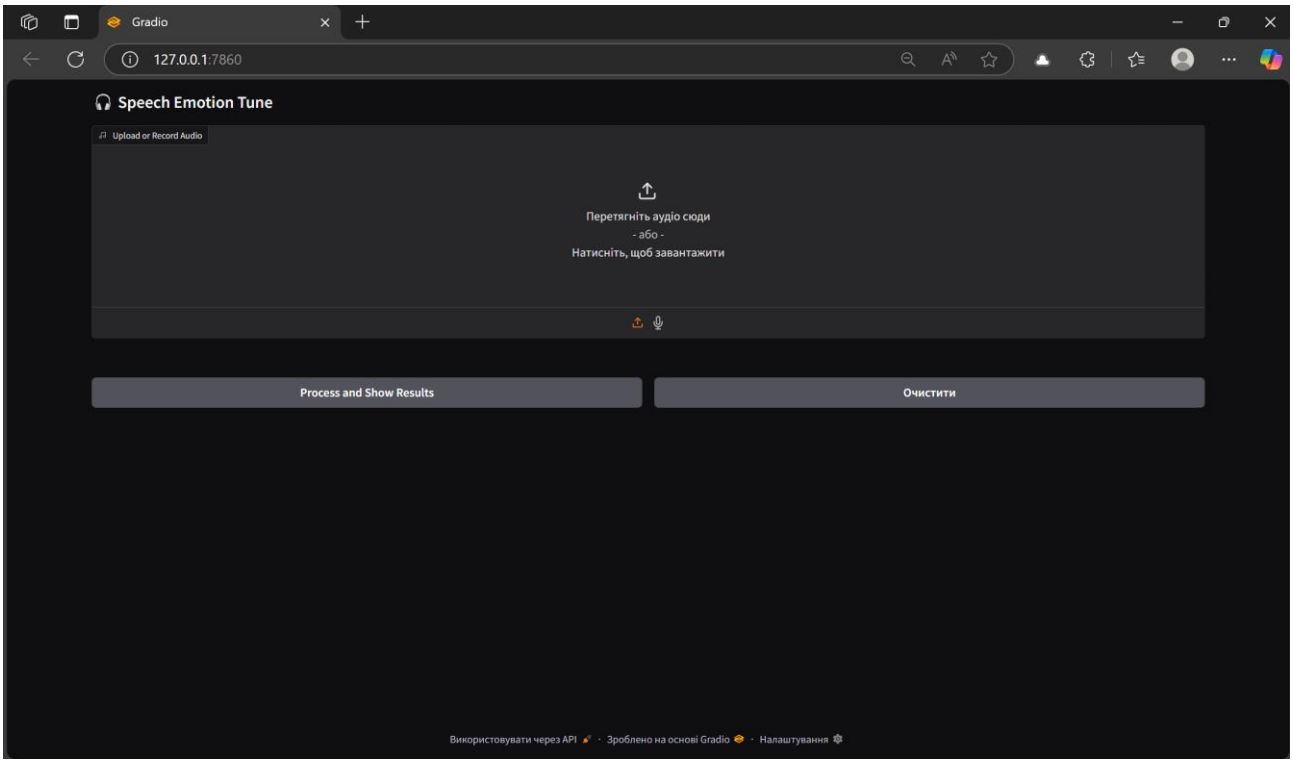

Fig. C.1. Initial view of the page

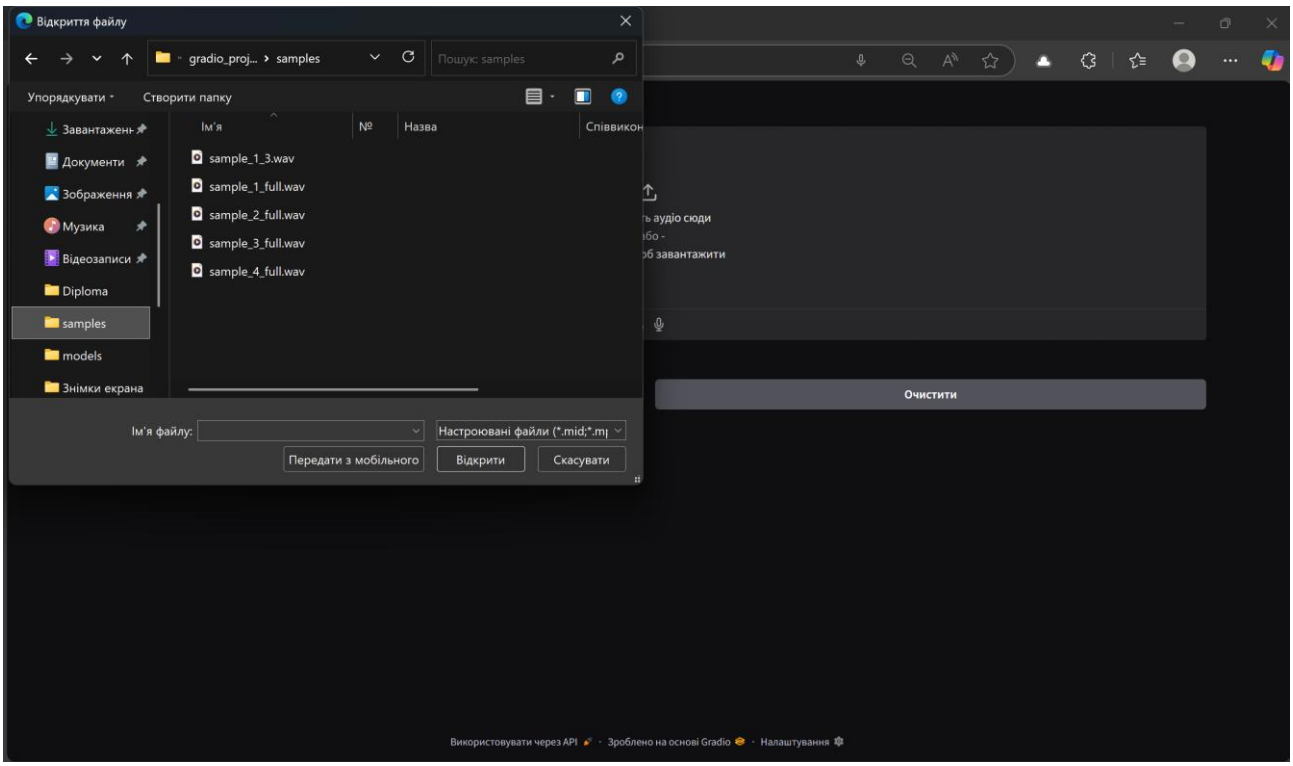

Fig. C.2. Uploading a record to the system

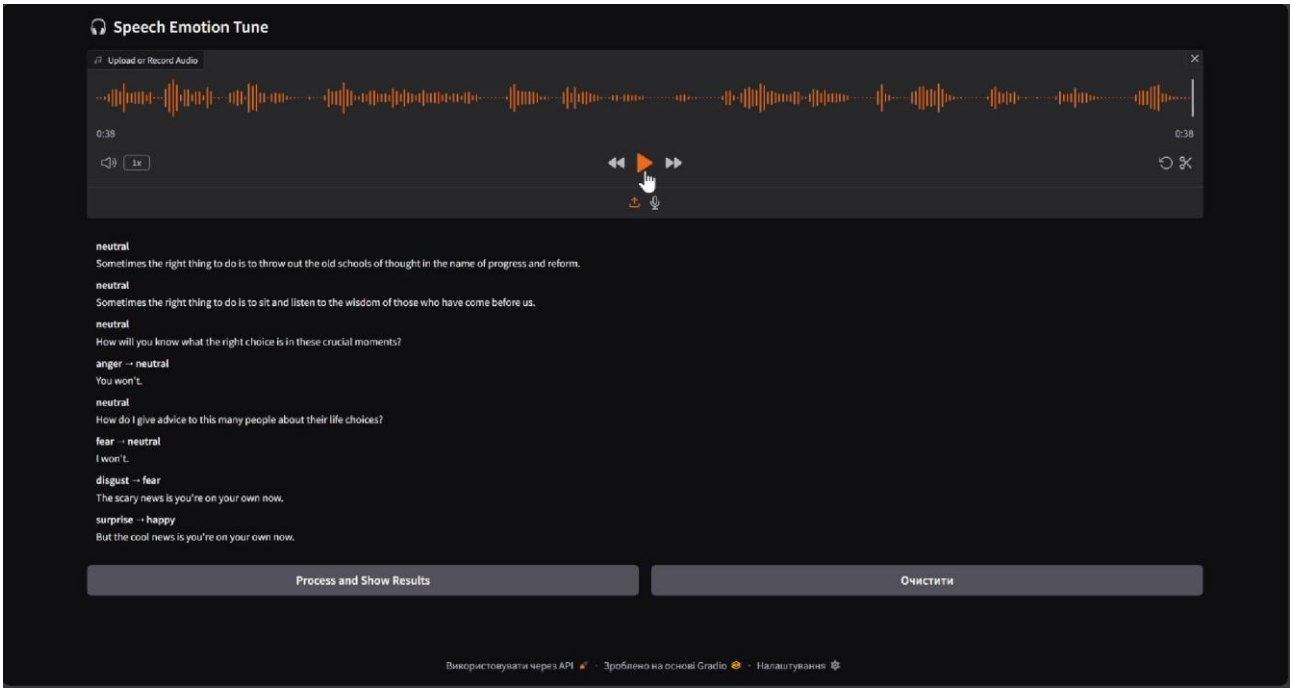

Fig. C.3. The result of the analysis of the speech excerpt

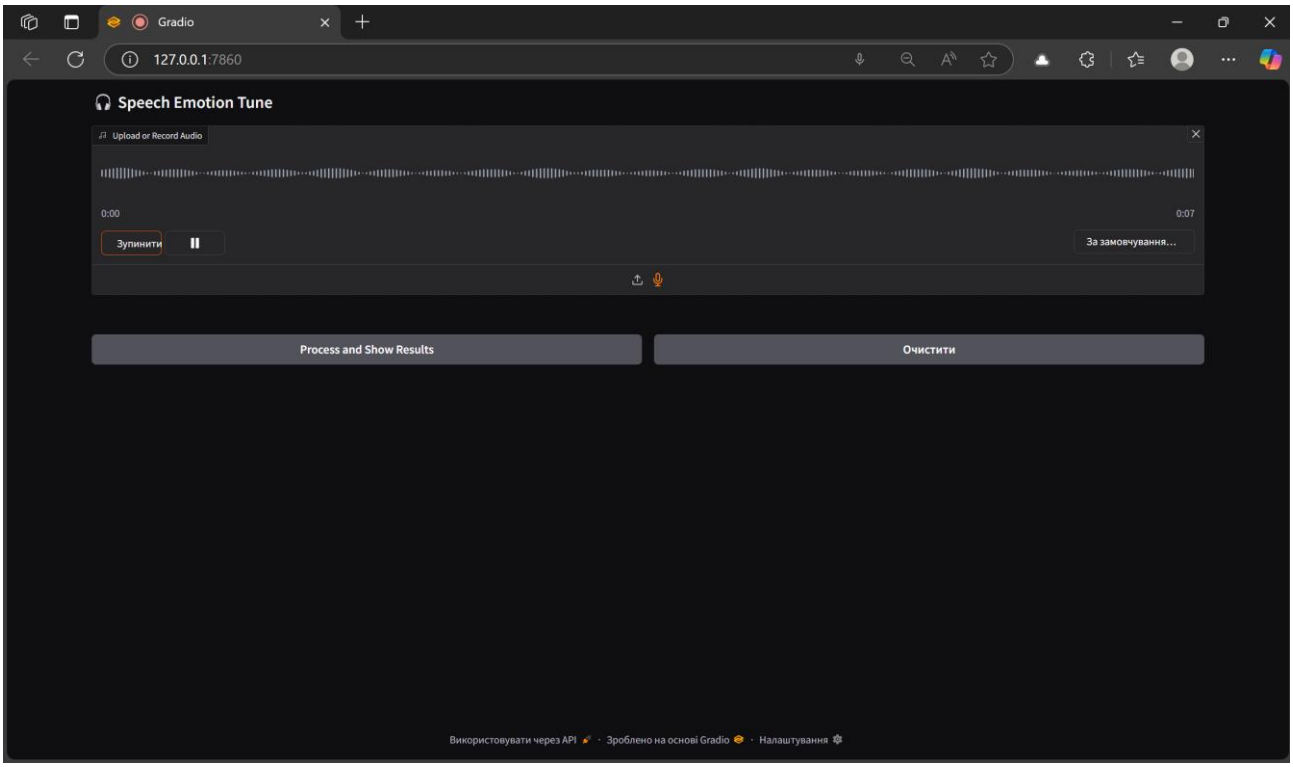

Fig. C.4. Recording an excerpt of the speech

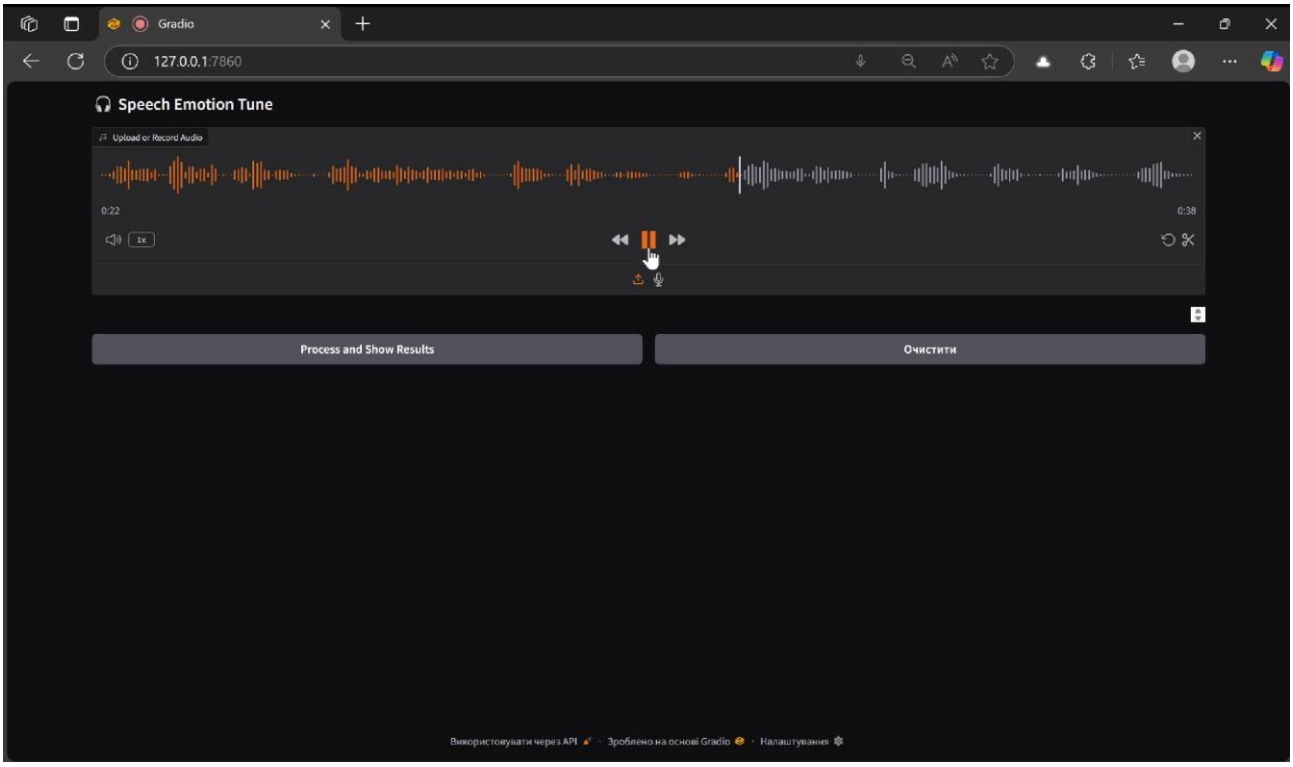

Fig. C.5. Listening to a recorded excerpt of the speech

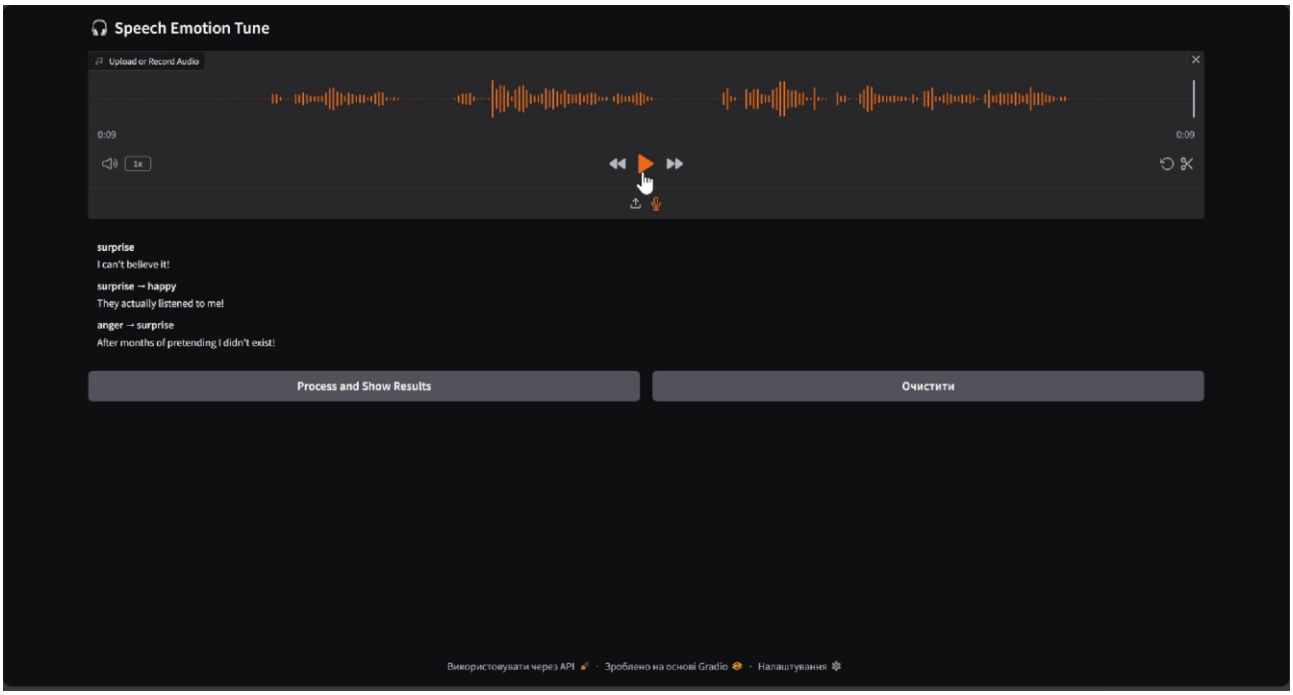

Fig. C.6. The result of the analysis of the recorded speech excerpt