UKRAINIAN CATHOLIC UNIVERSITY

MASTER THESIS

Speech Sentiment Classification in a Ukrainian-Russian Environment

Author:
Liudmyla PATENKO

Supervisor:
Oleksii IGNATENKO

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

in the

Department of Computer Sciences
Faculty of Applied Sciences

Lviv 2024

Declaration of Authorship

I, Liudmyla PATENKO, declare that this thesis titled, "Speech Sentiment Classification in a Ukrainian-Russian Environment" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

UKRAINIAN CATHOLIC UNIVERSITY

Faculty of Applied Sciences

Master of Science

Speech Sentiment Classification in a Ukrainian-Russian Environment

by Liudmyla PATENKO

Abstract

The process of sentiment classification involves categorizing human speech into one or more classes based on the emotional information expressed by the speakers. This study is focused on the development of a Speech Sentiment Classification (SSC) system designed to classify sentiment in a multi-lingual environment, including the Ukrainian language, while addressing the challenge of data scarcity. The research presents and evaluates three distinct approaches to this problem: a text-only classifier utilizing a Large Language Model (LLM), an audio-only classifier, and a bi-modal fusion approach that combines both text and audio features. The results indicate that the bi-modal fusion approach achieved an accuracy of 85% and an F1 score of 0.85 for binary classification of negative versus neutral sentiment.

Acknowledgements

First and foremost, I want to thank my supervisor, Oleksii Ignatenko, for his invaluable guidance throughout this project. I am also grateful to Yurii Paniv and Olexandr Korniienko for their valuable ideas and insights, which have greatly contributed to the success of this project.

I would like to extend my thanks to Stream Telecom for providing me with the opportunity to work on such an exciting project and for their ongoing support and feedback during the project.

I am also thankful to Ruslan Partsey for organizing the smooth process of the diploma project and for his guidance and support along the way.

Lastly, I am grateful to everyone involved in creating and executing the Ukrainian Catholic University Data Science program, and I would like to specially mention Daryna Petrenko for her constant support and inspiration.

Contents

Declaration of Authorship ii

Abstract iii

Acknowledgements iv

1 Introduction 1

2 Related works 3
2.1 Background . . . 3
2.1.1 Emotion dimensions . . . 3
2.1.2 Research datasets . . . 3
2.2 Training models from scratch . . . 3
2.2.1 Classical ML . . . 3
2.2.2 Deep Neural Networks . . . 5
2.3 Language transferability . . . 5
2.4 Fusion of audio and text models . . . 6
2.5 Pre-trained Self-Supervised models . . . 6
2.5.1 Finetuning . . . 7
2.6 Weakly-Supervised Learning . . . 8
2.7 Summary . . . 8

3 Datasets and evaluation 10
3.1 Russian-Ukrainian datasets . . . 10
3.1.1 Labeled data . . . 10
3.1.2 Raw data . . . 12
3.2 Evaluation metrics . . . 12

4 Metodology 13
4.1 Research Gap and Problem Formulation . . . 13
4.1.1 Research Gap . . . 13
4.1.2 The Problem Setting . . . 13
4.1.3 Research Goal . . . 13
4.2 Research Setting and Approach to Solution . . . 14
4.2.1 Approach to Solution . . . 14
Text-only baseline approach . . . 14
Acoustic model . . . 15
4.2.2 Bi-modal fusion . . . 16
4.3 Conclusion . . . 17

5 Experiments 18
5.1 Text model . . . 18
5.1.1 ASR model evaluation for transcription extraction . . . 18
5.1.2 Zero-Shot classifiers . . . 19

5.1.3 Classifier training 20
5.2 Acoustic model 21
5.2.1 ComParE audio features 21
Feature selection 21
5.2.2 Wav2Vec2 embeddings 22
No tuned 22
5.3 Bi-modal fusion 23
5.4 Execution time evaluation 24
5.5 Discussion 24

6 Conclusions **26**
6.1 Results 26
6.2 Future work 26

Bibliography **27**

List of Figures

2.1 Mapping categorical emotions into VAD "coordinates" (Bălan et al., 2019) 4
2.2 XLS-R model with Wav2Vec2.0 architecture created by Facebook (Mohamed et al., 2022) 7

3.1 Initial emotion and language distribution 11
3.2 Merged labels distribution 11

4.1 Diagram of a baseline 15
4.2 Diagram of an early fusion pipeline 16
4.3 Diagram of a late fusion pipeline 17

List of Tables

2.1 Short overview of the popular SER datasets. Columns **NA** and **NE** correspond to a number of actors and a number of emotions. Modalities A, V, and T correspond to audio, visual, and text components. . . . 4
2.2 SOTA results on IEMOCAP datastes. Unweighted accuracy is reported since weight accuracy is not present for all the works. Emotions are translated as following: a-anger, s-sadness, h-happiness, e-excitement, n-neutral, o-other. "Att" in model names is shortened to "Attention" 9

5.1 WER for Ukrainian and Russian texts. Lower WER indicates better performance. 19
5.2 Labels tuning on annotated text for MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model 19
5.3 UK/RU/ENG labels for annotated text for MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model 19
5.4 LLM classifiers evaluation on ASR transcription and translation 20
5.5 Logistic Regression classifier results (5-fold CV) using last hidden state of LLMs 20
5.6 Cross validation results for Logistic regression training using opensmile ComParE features 21
5.7 Wav2Vec2.0 XLS-R Facebook model embeddings evaluation: conversation-level 22
5.8 Wav2Vec2.0 XLS-R Facebook model embeddings evaluation: utterance-level 23
5.9 Evaluation of the fusion of 'cardiffnlp /twitter-roberta-base-sentiment-latest' hidden states (using ASR English translation) with audio features. The original score for the text component is in the 1st row, and the original F1 score for the audio component is in the last column . . 23
5.10 Evaluation of the fusion of 'MoritzLaurer /mDeBERTa-v3-base-mnli-xnl' hidden states (using ASR transcription) with audio features. The original score for the text component is in the 1st row, and the original F1 score for the audio component is in the last column 24
5.11 System components execution time for 1 audio file with duration 50 seconds 25
5.12 Models comparison in terms on accuracy, F1 score and execution time 25

List of Abbreviations

AI	**A**rtificial **I**ntelligence
ASR	**A**utomatic **S**peech **R**ecognition
CNN	**C**onvolutional **N**eural **N**etworks
CV	**C**ross-**V**alidation
DNN	**D**eep **N**eural **N**etworks
FFT	**F**ast **F**ourier **T**ransform
GMM	**G**aussian **M**ixture **M**odel
HF	**H**ugging **F**ace
HMM	**H**idden **M**arkov **M**odel
HS	**H**idden **S**tate
KNN	**k-** **N**earest **N**eighbors
LLD	**L**ow-**L**evel **D**escriptor
LLM	**L**arge **L**anguage **M**odel
LR	**L**ogistic **R**egression
LSTM	**L**ong **S**hort-**T**erm **M**emory
MFCC	**M**el-**F**requency **C**epstral **C**oefficients
ML	**M**achine **L**earning
NLI	**N**atural **L**anguage **I**nference
NLP	**N**atural **L**anguage **P**rocessing
SER	**S**peech **E**motion **R**ecognition
SOTA	**S**tate-**O**f-**T**he-**A**rt
SSC	**S**peech **S**entiment **C**lassification
SSL	**S**elf-**S**upervised **L**earning
SVM	**S**upport **V**ector **M**achine
UA	**U**nweighted **A**ccuracy
VAD	**V**alence **A**rousal **D**ominance
WA	**W**eighted **A**ccuracy
WER	**W**ord **E**rror **R**ate

Chapter 1

Introduction

Human speech is one of the main channels through which people pass information to the world. Emotions are a hidden part of human communication. It is often necessary not only to process what the human said but also how it was said. In today's business landscape, call centers serve as a first-line connection between customers and businesses. One key aspect of customer communication is the ability to recognize and respond to their emotions, especially when they indicate conflict or dissatisfaction. SSC technology can provide the potential for businesses to improve by identifying the emotional responses of their customers. In view of Ukraine's bilingual history, there is a need for an SSC model that can detect conflicts, based on the primary emotion of the conversation, particularly in call center settings in a multi-lingual environment, including the Ukrainian language.

SSC focuses on classifying emotions on a positive-negative scale. It is a subfield of the much broader field of Speech Emotion Recognition (SER), which describes emotions in various dimensions. To capture the most recent and promising approaches, alongside SSC field, we will review State-Of-The-Art (SOTA) works from the SER field, as they can be easily applied to SSC task.

One of the most effective ways to achieve good results in sentiment classification task is by using DNN (Deep Neural Network) models. Typically, these models involve a preprocessing step to extract hand-crafted acoustic features which are then passed as input to CNN (Convolutional Neural Network) (Badshah et al., 2017). Alternately, the model can take audio directly on input, as implemented in Trigeorgis et al., 2016; Tzirakis, Zhang, and Schuller, 2018, using a combination of LSTM (Long Short-Term Memory) and CNN models. The highest accuracy for DNN-based models is achieved by fusing text and audio components together, as demonstrated by Wu, Zhang, and Woodland, 2021; Morais et al., 2022; Atmaja, Sasou, and Akagi, 2022.

However, such models show good results only on the similar data they were trained on, but the research datasets available for training are usually acted. That means that narrated emotions could be exaggerated or feel artificial. This results in bad generalization and poor results in real-world scenarios. Additionally, research datasets are mostly available for commonly spoken languages like English, German, Spanish, and Chinese. The reason for such data scarcity lies in the efforts required for the annotation. It includes selecting the exact start and end points of the audio track, annotating emotions, which can be subjective, and optionally annotating the text that was spoken.

Self-supervised models are designed to solve the issue of labeled data scarcity since they can be pre-trained on large amounts of unlabeled audio. Then such models may be fine-tuned for SER or SSC task (Wang, Boumadane, and Heba, 2021; Wagner et al., 2023). Another approach would be to create weak SER labels for the

available audio data, for example, by using existing ASR (Automatic Speech Recognition) corpora and then train models in a supervised way.

In this work we aim to investigate two key advancements in SER field: bi-modal fusion with text and usage of self-supervised pre-trained models. Furthermore, our primary goal is to develop an end-to-end system for conflict detection in conversations. In this work we will mainly concentrate on classification of overall sentiment of the conversation. As a firs step we will consider that conflict is occurred when the conversation has negative sentiment.

Chapter 2

Related works

In this section, we will review the most promising research works that can address the issue of sentiment analysis for human conversation in a multi-lingual context.

2.1 Background

2.1.1 Emotion dimensions

Emotions in speech can be classified in different ways. One common method is to present them as either positive or negative sentiment, or as categorical classes such as anger, sadness, happiness, or neutrality. Alternatively, emotions can be described using a 3-dimensional scale of valence-arousal-dominance (VAD) (**Bradley**), where valence determines the emotional vector on a positive-negative scale, arousal determines the strength of the emotion, and dominance indicates the degree of control over the emotion. These three emotion representations are interrelated. For instance, categorical classes can be mapped to the VAD 3D coordinate system, and valence can be used to measure sentiment (Wagner et al., 2023) (Figure 2.1). Many datasets in SER are annotated to present emotions in one of these systems or can describe both categorical classes along with VAD mapping.

2.1.2 Research datasets

The primary research dataset used for SER evaluation is **IEMOCAP** (Busso et al., 2008). Categorical emotions include anger, sadness, happiness, disgust, fear, surprise, frustration, excitement, and neutral states. Usually, researchers validate their models on 4 classes: anger, sadness, happiness+surprise, and neutral, due to their sufficient quantity and relatively equal distribution. More information on the IEMOCAP dataset and other popular datasets can be found in Table 2.1. As with many of the research SER corpora, the IEMOCAP dataset is available upon request from the publishing institution.

2.2 Training models from scratch

2.2.1 Classical ML

Based on a survey by Akçay and Oğuz, 2020, statistical models together with audio features have traditionally been used for SER tasks. One example of such a model is the Gaussian Mixture Model (GMM). GMM is a probabilistic model that represents a mixture of multiple Gaussian distributions. In the context of SER, GMMs can be utilized to model the distribution of acoustic features associated with different

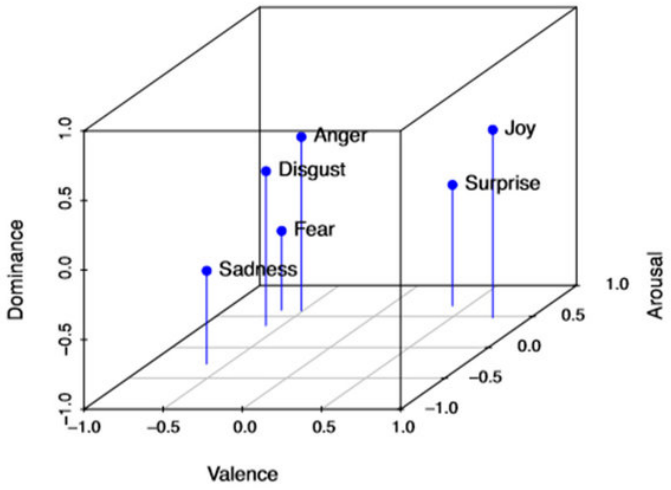


FIGURE 2.1: Mapping categorical emotions into VAD "coordinates" (Bălan et al., 2019)

Dataset Name	Lang	Length (h)	NA	NE	Nature	Modality
IEMOCAP Busso et al., 2008	EN	12	10	9	Acted	A,V,T
RECOLA Ringeval et al., 2013	FR	9.5	46	5	Elicited	A,V
MSP-Podcast Lotfian and Busso, 2017	EN	100 (v1.7)	N/A	9	Natural	A,T
CMU-MOSEI Zadeh et al., 2018	EN	66	1000	6	Natural	A,V,T
EmoDB Burkhardt et al., 2005	DE	0.5	10	7	Acted	A

TABLE 2.1: Short overview of the popular SER datasets. Columns **NA** and **NE** correspond to a number of actors and a number of emotions. Modalities A, V, and T correspond to audio, visual, and text components.

emotions. During training, the model estimates optimal parameters of the Gaussian distribution independently by maximizing the expectation of seeing the train example in the distribution given a set of acoustic features. This process is repeated for each class in the dataset. During testing, the likelihood of the test example is computed for each model to determine its most probable class (Metallinou, Lee, and Narayanan, 2008).

Except for GMMs for SER tasks have also been used such Machine Learning methods as Hidden Markov Models (HMM), Support Vector Machines (SVM), k-Nearest Neighbour (KNN), Decision Tree and Naïve Bayes Classifier(**Wani**). As an example, Sahu, 2019 reports 63% accuracy using Logistic Regression and SVM classifier on IEMOCAP dataset.

2.2.2 Deep Neural Networks

Deep Neural Networks (DNN) are known to be more efficient than traditional ML models because of their ability to capture complex patterns. DNN-based models can be developed as end-to-end models, where raw audio data is used as input, or may include additional steps such as audio preprocessing (e.g. noise cancellation, silent removal), transcribing, and retrieving features (both audio and text embeddings) (Trigeorgis et al., 2016).

Convolutional Neural Network (CNN) is a type of DNN architecture that is highly effective in extracting meaningful features from spectrotemporal representations of speech signals. Although CNNs were originally designed to work with images, they can also be used for SER tasks by converting spectrograms extracted from audio into images and using them as input to the network (Badshah et al., 2017). Additionally, 1D CNNs can be used to work with sequential data. In a study of Wagner et al., 2023, a 14-layer CNN network was proposed as a baseline SER model, consisting of 6 two-layer convolutional blocks, each followed by max pooling.

Recurrent neural network architectures, such as LSTM (Long Short-Term Memory), can be useful in capturing dependencies in audio data over time. To leverage both CNN and LSTM architectures, Trigeorgis et al., 2016 proposed an end-to-end model that uses raw audio instead of hand-crafted audio features. The model is built from two 1-d CNN blocks, which extract latent features, and LSTM model is used to extract contextual information. The study found a correlation between the activation of the CNN cells and prosodic audio features. This model architecture was improved in Tzirakis, Zhang, and Schuller, 2018 by increasing the number of CNN and LSTM layers while maintaining the proper kernel size of CNN to capture features. The current result is SOTA for RECOLA dataset.

2.3 Language transferability

The common approach when working with low-resourced languages, such as Ukrainian, is to train the model on the dataset publicly available for some popular language and then validate it on the small dataset for the target language (Iosifov et al., 2022). The research goal of this work was to find transferability between different languages. Authors report that among 7 datasets (English, German, French, Chinese, Farsi, Estonian, and Urdu), Chinese along with English were the least transferable, and the

model trained on Farsi provided the best results for the other languages: 77% accuracy when 2 emotions (angry and neutral) were evaluated and 36% accuracy for 4 emotions.

2.4 Fusion of audio and text models

The impact of the text dimension in the SER task was a point of significant research interest. Studies conducted in Wu, Zhang, and Woodland, 2021; Atmaja, Shirai, and Akagi, 2019 have reported that text-only models result in better accuracy than audio-only models. However, these studies confirm that the fusion of audio and text (Wu, Zhang, and Woodland, 2021; Atmaja, Shirai, and Akagi, 2019) improves general performance by 5-7%. See Table 2.2 for more detailed comparison.

Such fusion models may have a simple architecture but provide sufficient results. The model proposed in Atmaja, Shirai, and Akagi, 2019 consists of an LSTM layer for text embedding and a Dence layer for audio features. The model results in 75.5% UA on IEMOCAP dataset (without CV). To achieve this level of accuracy, authors retrieved 34 different audio features of time and spectral domains, Mel-Frequency Cepstral Coefficients (MFCC), and chromas. Also, they used preprocessing techniques to reduce noise and delete silence.

Given that the context of the phrase in conversation is important to determine its emotion, Wu, Zhang, and Woodland, 2021 proposed to combine together audio features, word embeddings, and BERT model with a context window to capture information. The best result of 78.4% UA with 5-fold CV authors achieved using [-3,3] context window.

The drawback of fusion audio with text modality is that it is required to use additional Speech-to-Text model in order to transcribe text into audio. It consumes additional resources, and when poorly chosen (with high error rate or with frequent hallucinations) may even worsen the results of acoustic model.

2.5 Pre-trained Self-Supervised models

Previously discussed DNN-based acoustic models were built and trained from scratch for the SER task. Another class of the models are based on transformer architecture (Vaswani et al., 2017): the raw audio input is passed to the CNN feature encoder to extract latent speech representations, which are fed into a contextual encoder consisting of several transformer layers. The models are trained in a self-supervised way by masking some part of the audio and trying to reconstruct it, minimizing loss. This way, models can learn useful language representations directly from the large quantity of audio input data without relying on external labeled annotations. Than the pre-trained models can be further tuned for the SSC task. Wav2Vec 2.0 (Baevski et al., 2020) and HuBERT (Hsu et al., 2021) are the example of such models. They have similar transformer-based architecture but differ in the the definition of targets and loss functions during pre-training.

Wav2Vec 2.0 model uses a quantization module that takes latent feature vectors from the CNN network as input. By applying Gumbel-Softmax, the module learns discrete speech units. These quantized units then serve as the target for minimization, achieved by computing the contrastive loss between them and the output of the encoder. The drawback of such an approach is that the quantization module is trained along with the model itself and requires precise parameter tuning.

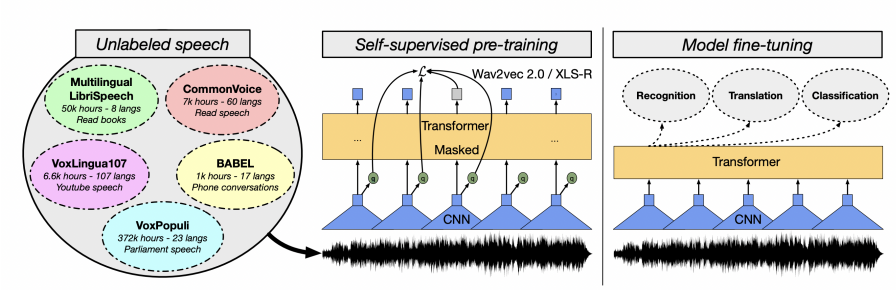


FIGURE 2.2: XLS-R model with Wav2Vec2.0 architecture created by Facebook (Mohamed et al., 2022)

HuBERT self-supervised training strategy suggests the use of a separate acoustic model and cross-entropy loss. In the first step, MFCC features are extracted from the audio segments, which then are clustered using k-means to N clusters. Obtained clusters are used later during training as targets for the transformer encoder.

In the context of our work, Wav2Vec2.0 has a main advantage over the other self-supervised models currently available due to the incorporation of Ukrainian audio data within its training dataset. The example of such multilingual model, based on Wav2Vec2.0 architecture is XLS-R (Mohamed et al., 2022) (see Figure 2.2).

2.5.1 Finetuning

In Wang, Boumadane, and Heba, 2021, the authors explore fine-tuning strategies of pre-trained self-supervised models HuBERT and Wav2Vec 2.0. Authors achieved SOTA results with both fully fine-tuned and partially fine-tuned (with frozen CNN layers) HuBERT models, with the latter showing a 1% improvement, reaching 72.56% UA on a 10-fold CV. Experiments with the additional pre-training model on 960h of the ASR dataset showed a slight accuracy drop. Authors suggest a loss of prosodic information during additional ASR training.

The usage of a pre-trained self-supervised model for SER tasks in the context of emotion valence (sentiment scale) is also described in Wagner et al., 2023. MSP-Podcast corpus was used for fine-tuning, and cross-corpus evaluation was performed on IEMOCAP and MOSI datasets. The authors highlighted two significant findings:

- Fine-tuned transformer-based models are better generalized and robust, as confirmed by obtaining good results on cross-corpus evaluation. This is an important statement since previous works focused on achieving the best results on a specific dataset. Authors suggest that transformer-based architectures are better suited to real-world applications than DNN-based.

- Models implicitly learn linguistic information from the input audio signal, as demonstrated by the experiment with synthesized emotionless data, where models were able to provide a good result on the valence scale. As a drawback of this feature, its performance towards English drops when the model is trained in multi-lingual data.

2.6 Weakly-Supervised Learning

One approach to increase model performance when available labeled data is limited would be to create weak labels, for example, by using existing ASR (Automatic Speech Recognition) corpora and then train models in a supervised way.

ASR systems are widely used for the automatic generation of transcription for audio content. An example of such model is Whisper ASR model (Radford et al., 2023). It was pre-trained on 680,000 hours of labeled audio data available on the internet, which is 10 times bigger than the training corpora used for pre-trained self-supervised models reviewed in the next section. Also, one-third of the training data was taken from different non-English languages (including Ukrainian).

In a study Wu, Zhang, and Woodland, 2021, researchers used an ASR system to extract transcription from the audio, which was then used as input to the SER bimodal model in the training and inference stages. Authors reported a drop of 3.7% when ASR transcription is used instead of the reference text. However, Li et al., n.d. demonstrated an almost linear dependency between the Word Error Rate (WER) of the ASR model and the UA of the SER system on English datasets. The higher the WER, the lower the accuracy.

The limitation of such weakly-supervised learning approach lies in the number of models required to preprocess the audio before the obtained text label and audio sample can be used in the SSC model: ASR, speaker diarization, audio-to-text alignment, and, possibly, speaker recognition. The more steps we have in the pipeline, the higher the possibility of accumulating errors.

2.7 Summary

Recent advancements in SER offer the possibility of improving performance and robustness in the context of the Ukrainian language. Techniques like bi-modal fusion, which incorporates text data alongside audio, can potentially improve the identification of emotions that might not be expressed in speech with high intensity. With the usage of pre-trained ASR systems, text components can be extracted from audio for further bi-modal fusion. Additionally, leveraging pre-trained self-supervised models trained on large unlabeled datasets can enhance the model's ability to learn complex relationships within the data, potentially leading to more robust emotion classification.

Model	Modality	Emotions	Test setting	UA
Bi-modal fusion				
Self-Attention + BERT[-3,3] Wu, Zhang, and Woodland, 2021	A+T	h, s, n, a	5-fold CV	**78.41**
LSTM + Dense Atmaja, Shirai, and Akagi, 2019	A + T	e, s, n, a	80/20 split	75.49
Self-Attention + BERT[-3,3] Wu, Zhang, and Woodland, 2021	A+T	h, s, n, a, o	5-fold CV	75.6
Self-supervised models				
huBERT + W2V2 Morais et al., 2022	A	h+e, s, n, a	5-fold CV	77.76
Attention-Guided-WavLM-Large-v2 Ioannides et al., 2023	A	h, s, n, a	5-fold CV	74.32
RNN w/ attention + SpecAugment Lu et al., 2020	A	h+e, s, n, a	10-fold CV	72.56
Text-only model				
BERT[-3,3] Wu, Zhang, and Woodland, 2021	T	h, s, n, a, o	5-fold CV	71.88
LSTM + Attention Atmaja, Shirai, and Akagi, 2019	T	e, s, n, a	80/20 split	68.01
Deep Neural Networks				
CNN14 Wang, Boumadane, and Heba, 2021	A	h+e, s, n, a	10-fold CV	55.8

TABLE 2.2: SOTA results on IEMOCAP datastes. Unweighted accuracy is reported since weight accuracy is not present for all the works. Emotions are translated as following: a-anger, s-sadness, h-happiness, e-excitement, n-neutral, o-other. "Att" in model names is shortened to "Attention"

Chapter 3

Datasets and evaluation

This chapter focuses on reviewing the dataset provided by Stream Telecom company, which will be used to develop a sentiment classification system. We will review the text and audio annotations, label distribution, and explain our choices for merging the labels. Additionally, we will discuss the audio properties of the data.

We also acknowledge the availability of a larger, unlabeled dataset, which presents opportunities for pseudo-labeling and further expansion of our system. To ensure the accuracy and effectiveness of our models, we will use cross-validation techniques and report metrics such as accuracy, precision, recall, F1 score, and Word Error Rate (WER).

3.1 Russian-Ukrainian datasets

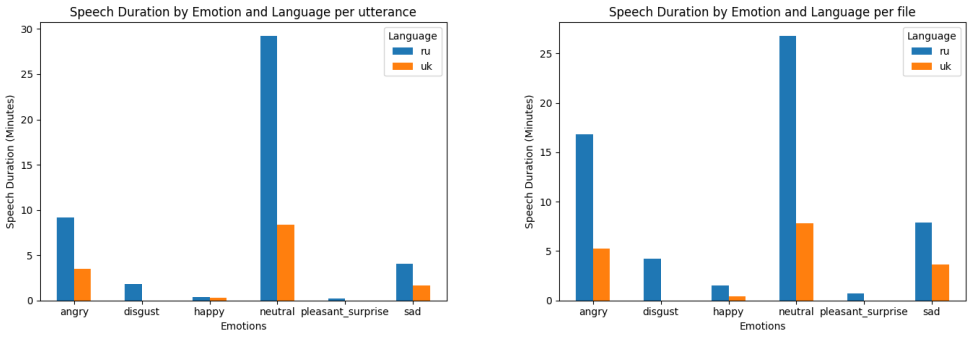
3.1.1 Labeled data

We will conduct experiments using a dataset provided by a Stream Telecom company and annotated in Mrozek and Danylov, 2021. The audio is recorded with 8kHz sample rate, 16 bits per sample and 1 channel. The dataset contains 90 audio files, and each sample is a record of a different customer. The audio files contain conversations between an operator and a customer in Russian or Ukrainian languages. The average duration of each conversation is 50 seconds. The dataset includes 6 emotion labels for each utterance as well as overall file emotion and ground-truth transcription. The emotion labels are the following: "pleasant_surprise", "happy", "neutral", "sad", "angry" and "disgust". Data labels distribution is imbalanced (see Figure 3.1):

- Emotion labels are highly skewed towards neutral class and less towards negative emotions. There are almost no positive conversations present in the data.
- There is an imbalance in the language component, with approximately two-thirds of the conversations being in Russian and one-third in Ukrainian.

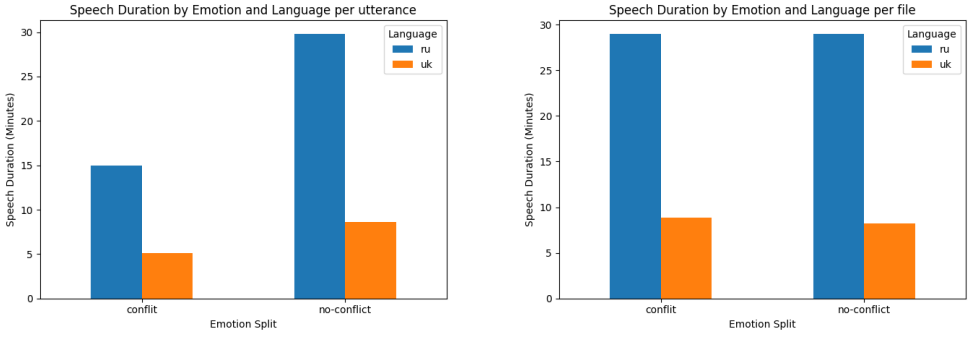
To eliminate emotion bias we mapped emotions into 2 classes: "conflict" and "no-conflict" based on sentiment (valence) component of each emotion. We merged "happy" and "pleasant_surprise" to "neutral" class to indicate "no-conflict" conversation. "sad", "angry" and "disgust" labels were combined together to "conflict" label. New label distribution is illustrated in Figure 3.2. We received balanced class distribution per language for file level, as for utterance level, there is still bias towards "no-conflict" class, since speech of call center operator is always marked as "neutral".

This dataset already has been preprocessed to remove unrelated sounds such as dial tones, voice mail agents, silence, and music, as well as sensitive information related to the customer and the company. Additionally, we resampled audio from 8kHz to 16kHz before starting the experiments with acoustic classifiers to have consistency with pre-trained models.



(A) utterance level (B) file level

FIGURE 3.1: Initial emotion and language distribution



(A) utterance level (B) file level

FIGURE 3.2: Merged labels distribution

3.1.2 Raw data

Stream Telecom company also provided us with 3000 audio files recorded from 03/2020 to 11/2020. The recordings share the same audio properties as previously described dataset. The data is not labeled and not preprocessed, but it may be further used for pseudo-labeling.

3.2 Evaluation metrics

For classification model comparison, we will use accuracy, precision, recall, and F1-score metrics with n-fold cross-validation.

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$F_1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

To report ASR model performance WER metric will be used, which is calculated as the number of substitutions (S), deletions (D), and insertions (I) relative to the total number (N) of words in the reference transcript. Lower WER values indicate better performance:

$$WER = \frac{S + D + I}{N}$$

Chapter 4

Metodology

This chapter presents a methodology for SSC for mixed Ukrainian-Russian audio data. It addresses a research gap in this area and outlines the problem setting. The research goal is to develop an effective end-to-end SSC system for the multi-lingual scenario given limited amount of data. The chapter details the approach to the solution, starting with a text-only baseline model using LLMs and ASR for transcription (Section 4.2.1). It then describes acoustic models using hand-crafted features and Wav2Vec2.0 architecture (Section 4.2.1). Finally, it explores bi-modal fusion techniques to combine text and audio components (Section 4.2.2).

4.1 Research Gap and Problem Formulation

4.1.1 Research Gap

Our search for related works for Sentiment or Emotion Classification tasks in audio for the Ukrainian or mix of Ukrainian and Russian languages yielded very limited results. Only work of Mrozek and Danylov, 2021 relates to the aim of this work, but it is focused on emotion classification on utterance level, while our work consider sentiment classification for the whole conversation.

Also, there is room for improvement in terms of text-only and bi-modal fusion models. In previous works that implemented the fusion of audio and text components (Wu, Zhang, and Woodland, 2021, Atmaja, Shirai, and Akagi, 2019) , word embeddings like GloVe or language representation extracted from pre-trained Large Language Models (LLMs) such as BERT were used. In our work, for text component extraction, we suggest using models specifically trained on sentiment classification or emotion recognition tasks.

4.1.2 The Problem Setting

4.1.3 Research Goal

To fill the gap in research on SSC for the Ukrainian language and mixed Ukrainian-Russian datasets, we aim to investigate this topic and contribute valuable insights to the field by evaluating the effectiveness of recent SOTA methods like the fusion of text component with audio and using pre-trained self-supervised transformer models. We will develop and evaluate an end-to-end SSC system specifically designed for the multi-lingual scenario of Ukrainian, Russian, and their mix, focusing on performance under limited training data. The evaluation will be conducted on a real-world bilingual dataset, with separate results reported for the Ukrainian portion to aid future research. The resulting model will be used for conflict detection based on the dominant sentiment of the conversation.

By evaluating recent SOTA methods on our bi-lingual dataset, we hope to establish a foundation for future development of conflict detection tools that are applicable to real-world scenarios like mental health audio chatbots or customer service analysis. While the proprietary nature of our dataset limits public data sharing, we plan to open-source the resulting training and inference pipelines, potentially benefiting researchers and the Ukrainian industry working on similar tasks.

4.2 Research Setting and Approach to Solution

4.2.1 Approach to Solution

Our approach to the solution for the SSC task will leverage the advancements of the SER field discussed in 2 and the recent rapid evolution of the NLP field, particularly LLM field.

Given the significant imbalance in the emotion classes within the original dataset, we have merged emotions on a positive-negative scale and will concentrate on binary classification. Our primary objective, due to the importance of not missing conflicts in call center scenarios, will be to maximize the F1 score for conflicts.

With the absence of an SSC benchmark for Ukrainian or a mix of Russian and Ukrainian, our initial step will be to develop a text-only model. This model will serve as a baseline for comparing audio and bi-modal models. We intend to leverage existing LLMs trained for zero-shot or sentiment classification, and utilize existing ASR models to extract transcriptions or translations into English from the audio for use in the text model.

Next we will start working with audio component. We will evaluate handcrafted features as well as Wav2Vec2.0 encodings using classical Machine Learning (ML) methods such as Logistic Regression. The evaluation will be conducted at both the conversation and utterance levels. To enhance the Wav2Vec2.0 results, we will fine-tune the model on our dataset. Finally, we will integrate the text and audio components.

Text-only baseline approach

As demonstrated Wu, Zhang, and Woodland, 2021 and Atmaja, Shirai, and Akagi, 2019, text modality may itself provide a sufficient result in the SSC tasks. The field of Generative AI has seen rapid advancements in recent years, with the introduction of increasingly powerful LLM architectures on a regular basis. Notably, some of these models are pretrained for Ukrainian and Russian languages, making them suitable for various tasks, including SSC.

LLM selection. To develop a text-only baseline, we selected 3 LLMs from the HuggingFace repository based on the target task and suitability for our dataset:

- MoritzLaurer/mDeBERTa-v3-base-mnli-xnli (Laurer et al., 2024) DeBERTa model (He, Gao, and Chen, 2021), initially pre-trained on multilingual CC100 dataset, which includes 14G of Ukrainian data, and then tuned on XNLI dataset (Conneau et al., 2018) for Natural Language Inference (NLI) task.

FIGURE 4.1: Diagram of a baseline

- SamLowe/roberta-base-go_emotions [1] is a RoBERTa base model fine-tuned for multi-label classification task on the English emotion dataset GoEmotions (Demszky et al., 2020) with 28 emotion categories.
- cardiffnlp/twitter-roberta-base-sentiment-latest (Loureiro et al., 2022), also RoBERTa model with base architecture tuned for sentiment classification task on manually collected Twitter sentiment data.

The first mDeBERTa model can be used with either original data, transcription and translation, while the latter two must be evaluated using English translation obtained through ASR system.

ASR model. Our initial dataset contains transcription for conversations in both Ukrainian and Russian manually provided by annotators. However, our system will work with raw audio in a real-world setup. Because of that, we need to select the model that may work with non-ideal text transcription obtained automatically. Additionally, as reported in Wu, Zhang, and Woodland, 2021, there might be a significant drop in model performance when the bi-modal model was trained on the original text and evaluated on ASR transcription, in comparison with results when the model was trained and evaluated on ASR transcription: 63.47% vs. 71.9% UA.

To obtain transcription from the audio, we will use an ASR model, such as Whisper (Radford et al., 2023), which was trained on multi-lingual data, including the Ukrainian language.

Pipeline. As a result of this initial step, we will obtain an end-to-end pipeline consisting of an ASR model followed by an LLM model. The pipeline will take audio as input, and the output will be conflict classification (see Figure **??**).

At the experimental stage we will select LLM and ASR model with highest performance to use in the baseline. Further experiments with SSC acoustic models will be compared with the baseline. We will explore whether end-to-end audio or bi-modal models can enhance the performance of the text-only model, particularly for the Ukrainian language. We expect to confirm the results of Wu, Zhang, and Woodland, 2021; Morais et al., 2022; Atmaja, Sasou, and Akagi, 2022 on our bilingual dataset, that the acoustic model alone results in lower accuracy than the text baseline model.

Acoustic model

Hand-crafted features To have a starting point for the acoustic component, we will use hand-crafted feature with Logistic Regression classifier, since it is less prone to overfitting, than more complex models like neural networks, when training on a small data set, as ours (90 audio files). We will leverage an open-source openSMILE (Eyben, Wöllmer, and Schuller, 2010) feature extractor to obtain audio features for

[1] https://huggingface.co/SamLowe/roberta-base-go_emotions

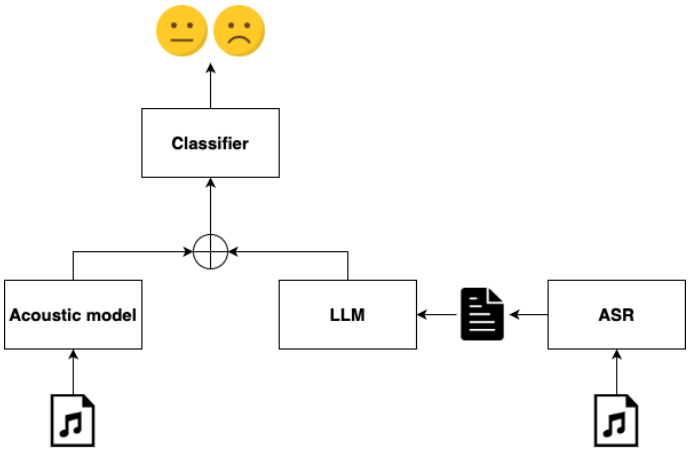


FIGURE 4.2: Diagram of an early fusion pipeline

this experiment. The software is able to extract Low-Level Descriptors (LLD), such as signal energy, pitch, FFT spectrum, Mel spectrum, and Spectral. Then, it calculates functional features (statistical, polynomial regression coefficients, and transformations) from the LLD. We use the ComParE 2016 feature set (Schuller et al., 2016), which provides 6373 functional paralinguistic features.

Wav2Vec2.0 architectures. As suggested by Wagner et al., 2023, pre-trained in self-supervised way models based on Wav2Vec2.0 architecture may perform less on the specific dataset but are more generalized. To evaluate if the Wav2Vec2.0 model will show better performance compared to ComParE features, we will evaluate the XLS-R model (Mohamed et al., 2022, Figure 2.2), based on Wav2Vec2.0 architecture with a different number of parameters: 300 million, 1 billion, 2 billion.

Pipeline. The pipeline for acoustic model contain only possible audio preprocessing step (not discussed in this work) following by the audio model itself. In our case it can be Logistic Regression classifier or tuned Wav2Vec2.0 on SSC task.

4.2.2 Bi-modal fusion

As proposed by Wu, Zhang, and Woodland, 2021 and Atmaja, Shirai, and Akagi, 2019, combining audio and text components may enhance model performance. Since some of the negative conversations contain text with neutral or positive sentiment but are spoken with angry intonation, adding an acoustic component to the text model should improve classification results.

For the text component we will use hidden states obtained from LLMs suggested for baseline. For acoustic component we will use either ComParE feature set as well as audio encodings extracted from Wav2Vec2.0 architecture.

Pipeline. Our work explored two fusion methods: early fusion and late fusion. For early fusion (Figure 4.2), we will concatenate text encodings with audio encodings and passed a new vector to the classifier. For late fusion (Figure 4.3, we will adopt the ensemble method with majority voting on the resulting probabilities.

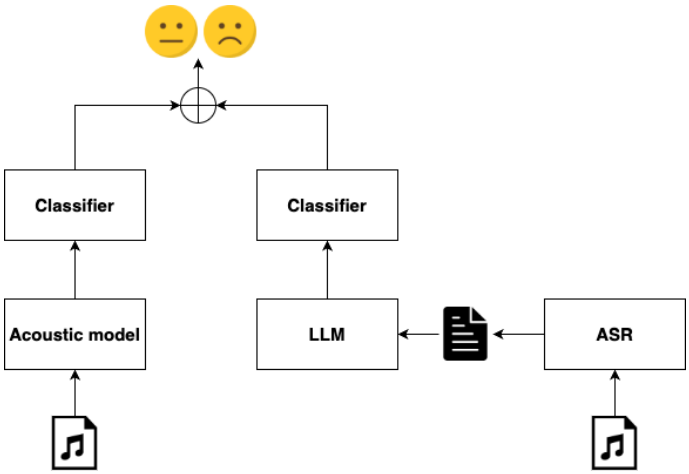


FIGURE 4.3: Diagram of a late fusion pipeline

4.3 Conclusion

The proposed methodology aims to fill a gap in research on SSC for the Ukrainian language and mixed Ukrainian-Russian datasets. It leverages advancements in the SER and NLP fields, utilizing LLMs and recent techniques like Wav2Vec2.0 as well as audio ComParE features.

Chapter 5

Experiments

This chapter discusses experiments on sentiment classification using both textual and acoustic data. We started our experiments by creating a baseline, which will be further used for comparison. For the baseline, the ASR model is chosen experimentally in Section 5.1.1, and Zero-Shot classification LLMs were evaluated in Section 5.1.2. Then we used a Linear Regression classifier to evaluate text encodings (Section 5.1.3), audio hand-crafted (Section 5.2.1) features, and audio representations extracted from Wav2Vec2.0 model (Section 5.2.2). We combined audio and text components together in Section 5.3. Obtained results are summarized in Section 5.5.

5.1 Text model

5.1.1 ASR model evaluation for transcription extraction

To select the best ASR model, we evaluated Whisper ASR models and models trained by the Ukrainian Speech Recognition community [1]. We used HuggingFace [2] model hub, along with the transformer library, to load models.

Given that both the Whisper and Wav2Vec2 models were trained on audio with a sampling rate of 16kHz, it was necessary to resample the original audio. Furthermore, for the Whisper model, the audio needed to be split into chunks, with a maximum duration of 30 seconds. These preprocessing steps were automatically executed using the 'automatic-speech-recognition' pipeline for Whisper.

According to the obtained results in Table 5.1, Whisper large-v2 and large-v3 models showed the best results for both Russian and Ukrainian language. We decided to use the transcription from the v2 model, as the v3 model has a higher tendency to hallucinate, particularly in low-resourced languages. This is due to the v3 model's training dataset, which includes 4 million hours of audio that was pseudolabeled using the Whisper large-v2 model. In addition to transcription, we collected translations into English for further experiments.

In the original annotation, the text was split into separate utterances by speaker. We attempted to utilize the Whisper feature to return word timestamps and combine them into similar sentences. However, this experiment failed as the received utterance timestamps were not aligned with the original ones. In order to obtain similar annotation, a speaker detection tool should have been used to map the speaker to the received timestamp. Consequently, we have chosen to postpone this experiment for future iterations.

[1] https://github.com/egorsmkv/speech-recognition-uk
[2] https://huggingface.co/

Model (HF)	RU	UK
openai/whisper-large-v2_wer	0.40	**0.39**
openai/whisper-large-v3_wer	**0.38**	0.41
arampacha/wav2vec2-xls-r-1b-uk_wer	0.89	0.42
openai/whisper-medium_wer	0.44	0.50
Yehor/w2v-bert-2.0-uk_wer	0.92	0.51
openai/whisper-large_wer	0.42	0.52
Yehor/wav2vec2-xls-r-1b-uk-with-lm_wer	0.91	0.53
Yehor/wav2vec2-xls-r-300m-uk-with-small-lm_wer	0.92	0.58
robinhad/wav2vec2-xls-r-300m-uk_wer	0.91	0.58

TABLE 5.1: WER for Ukrainian and Russian texts. Lower WER indicates better performance.

Labels	ACC	F1
"щастя" "нейтральне" "злість"	**81.11**	**0.76**
"позитив" "негатив" "нейтральне"	78.89	0.72
"позитив" "негатив"	76.67	0.70
"конфлікт" "нейтральне"	64.44	0.67

TABLE 5.2: Labels tuning on annotated text for MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model

5.1.2 Zero-Shot classifiers

In alignment with our original dataset, we combined labels associated with positiveness and neutrality together to evaluate a binary classification task. In our experiments, first we conducted zero-shot classification using various labels in Ukrainian on the original data annotations with the MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model (Table 5.2). We then translated some of these labels into Russian and English and repeated the evaluation (Table 5.3). Finally, we used ASR transcription and English translation for the final evaluation (Table 5.4).

As we may observe in the obtained results, models trained specifically for sentiment classification tasks perform better on noisy data produced by ASR and translation than the zero-shot classification model evaluated on original texts. Furthermore, there are two interesting observations related to the zero-shot classification model:

- The model performs slightly better for Ukrainian and English labels when using 3 labels (a more fine-grained sentiment scale) instead of 2.
- Despite being fine-tuned on XNLI, which includes Russian, the model's performance significantly deteriorates when using Russian labels compared to

Labels	ACC	F1
"щастя" "нейтральне" "злість"	**81.11**	**0.76**
"happy" "angry" "neutral"	70.00	0.54
"счастье" "неитральное" "злость"	62.22	0.26
"positive" "negative"	75.56	0.74
"позитив" "негатив"	76.67	0.70

TABLE 5.3: UK/RU/ENG labels for annotated text for MoritzLaurer/mDeBERTa-v3-base-mnli-xnli model

Model	Text	Labels	ACC	F1
cardiffnlp /twitter-roberta-base-sentiment-latest	transl		**84.44**	**0.81**
SamLowe /roberta-base-go_emotions	transl		80.00	0.77
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	transl	"positive" "negative" "neutral"	75.56	0.72
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	transl	"positive" "negative"	74.44	0.72
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	transc	"позитив" "негатив" "нейтральне"	80.00	0.71
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	transc	"щастя" "нейтральне" "злість"	75.56	0.69

TABLE 5.4: LLM classifiers evaluation on ASR transcription and translation

Model	Text	ACC	F1
cardiffnlp /twitter-roberta-base-sentiment-latest	transl	**84.44**	**0.85**
SamLowe /roberta-base-go_emotions	transl	82.22	0.81
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	transc	71.11	0.72
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	origin	70.00	0.67
MoritzLaurer /mDeBERTa-v3-base-mnli-xnli	transl	65.56	0.62

TABLE 5.5: Logistic Regression classifier results (5-fold CV) using last hidden state of LLMs

Ukrainian or English (Table 5.3).

5.1.3 Classifier training

To further assess the potential of pre-trained models for our task, we trained a logistic regression classifier on the last hidden states extracted from the models evaluated in Section 5.1.2. We performed 5-fold cross-validation and report the F1 score and accuracy. The results are available in Table 5.5. As expected, models, tuned specifically for emotion or sentiment classification performs better, than general features obtained from the model trained for NLI task. Even more, even with training, NLI model performs worse in training case, than zero-shot classification using experimentally selected labels (see Table 5.3)

As expected, the hidden states of the models that are specifically trained for emotion or sentiment classification encode more valuable information for our task than the general features obtained from the model that was trained for the NLI task. Furthermore, the LR classifier trained on domain data using NLI model features performs worse than zero-shot classification using experimentally selected labels (see Table 5.3).

Audio level	Aggregation	ACC	F1
utterance	propagation-clients-only	76.19	**0.77**
utterance	propagation-all	74.12	0.76
conversation	-	75.56	0.75
utterance	max pooling	67.06	0.64
utterance	min pooling	64.71	0.62
utterance	-	**82.94**	0.60
utterance	avg pooling	57.65	0.50

TABLE 5.6: Cross validation results for Logistic regression training using opensmile ComParE features

5.2 Acoustic model

Our aim is to evaluate acoustic only model. We will experiments with both on conversation-level (using audio file of the whole conversation, with average duration 50 sec) and on utterance-level (using parts of the conversation said by one speaker and annotated by hand, with average duration 7 sec). We will use handcrafted audio features and experiment with XLS-R Wav2Vec2.0 model. For utterance-level we will mainly use 2 aggregation strategies along with speaker split: propagation-all, where we consider sentiment of the conversation is negative if at least one utterance is negative, propagation-clients-only, where we train only on client utterances and apply the same negative sentiment propagation.

5.2.1 ComParE audio features

We used an open-source openSMILE (Eyben, Wöllmer, and Schuller, 2010) feature extractor to obtain audio features for this experiment. The ComParE feature set provides 6373 functional paralinguistic features. The experiments executed with these features are similar to those executed for the text component: training of LR and SVC models with 5-fold cross-validation. We extracted audio features for both the whole conversation and each utterance in the conversation. To classify the sentiment of the conversation based on the utterances, we tried multiple aggregation strategies: mean, average and max pooling of audio features of the utterances in the conversation before training; label propagation - if at least one utterance in the conversation has negative sentiment the whole conversation considered to have negative sentiment; label propagation, but the model is trained only on utterances said by the customer, since 97% of the replicas that support person says are neutral. Also, we run one experiment to evaluate how well the model may learn to classify each utterance. Results of the experiments are available in Table 5.6.

In the results we may observe, that pooling strategy of individual utterances together shows worse results, than extracting features for the whole conversation. Training model on client utterances, while support utterances are ignored, performs on 2% better in terms of F1 score, than classifying the whole conversation. Our experiment on classifying each individual utterance shows overfitting signs and bias towards no-conflict class: we obtained 82% accuracy with only 0.6 F1 score.

Feature selection

ComParE feature set contains 6373 different paralinguistic functional features. In order to assess the features with the most significant impact on the SSC task for

N params	Feature level	ACC	F1
2b	hidden states	**68.89**	**0.67**
2b	CNN features	62.22	0.62
1b	hidden states	67.78	**0.67**
1b	CNN features	62.22	0.61
300m	CNN features	61.11	0.61
300m	hidden states	57.78	0.57
300m	hidden states ranom init	47.78	0.46
300m	CNN features random init	41.11	0.40

TABLE 5.7: Wav2Vec2.0 XLS-R Facebook model embeddings evaluation: conversation-level

our dataset, simplify the model, and potentially enhance accuracy, we run logistic regression cross-validation with L1 regularization (max_iter=100, C=1).

As a result, the feature set was reduced by a factor of 10 to 643 features, which contains statistics for: magnitude of a Fast Fourier Transform (FFT) function (fftMag); relative spectral transform and filtering applied to auditory spectrum (audSpec_Rfilt, audspecRasta); MFCC (mfcc); local jitter and shimmer (jitterLocal, shimmerLocal); fundamental frequency (F0) (F0final); logarithmic harmonics-to-noise ratio (logHNR).

After retraining LR classifier with cross-validation on the conversation level we obtained 77% accuracy and 0.76 F1 score, which improved results by 1.5% accuracy and 0.01 F1 score compared to the original ComParE feature set results.

5.2.2 Wav2Vec2 embeddings

We want to evaluate if pre-trained on multimodal data Wav2Vec2 embeddings will perform well on our data.

No tuned

First we are going to evaluate how well pre-trained in self-supervices way models will work for SSC task for specifically mix of Ukrainian and Russian languages. For this experiment we used XLS-R model based on Wav2Vec2.0 architecture trained to learn cross-lingual speech representation with 3 sets of parameters: 300m, 1b and 2b. We extracted hidden state from the last transformer layer as well as audio features from CNN feature extraction layer for the whole conversation and individual utterances. Also, we initialize 300m model with random weights in order to evaluate if pre-training really makes difference specifically for our out-of-domain dataset with original sampling rate 8kHz up-sampled to 16kHz. Than as usual we trained LR with 5-fold cross validation.

The evaluation results of the LR classifier using conversation-level data can be found in Table 5.7. As expected, randomly initialized weights led to poor results. The low-level audio features extracted from the feature extraction CNN layers showed a decrease in performance by approximately 5% for each model, in contrast to the high-level language representations extracted from the final transformer layer of the model (hidden states). Additionally, we observed only a minimal improvement between a model with 1 billion parameters and a model with 2 billion parameters.

Performance of LR classifier trained on utterance-level data (all or only client utterances) are similar to ones obtained from

N params	Feature level	Aggregation	ACC	F1
1b	CNN features	propagation-clients-only	**67.86**	**0.70**
1b	hidden states	propagation-all	67.06	0.69
1b	hidden states	propagation-clients-only	66.67	0.67
1b	CNN features	propagation-all	62.35	0.67
300m	CNN features	propagation-clients-only	61.90	0.66
300m	hidden states	propagation-all	60.00	0.65
300m	CNN features	propagation-all	58.82	0.62
300m	hidden states	propagation-clients-only	59.52	0.61

TABLE 5.8: Wav2Vec2.0 XLS-R Facebook model embeddings evaluation: utterance-level

Features	Fusion	ACC	F1	F1 (audio)
twitter-roberta translation (T)	-	84.44	**0.85**	-
T+ComParE	early	**85.56**	**0.85**	**0.75**
T+ComParE	ensemble	78.89	0.79	**0.75**
T+Wav2Vec-1b-CNN	early	84.44	0.84	0.61
T+Wav2Vec-1b-CNN	ensemble	82.22	0.82	0.61
T+Wav2Vec-1b-HS	early	82.22	0.82	0.67
T+Wav2Vec-300m-EF	ensemble	78.89	0.79	0.67
T+Wav2Vec-300m-EF	early	83.33	0.83	0.61
T+Wav2Vec-300m-HS	early	82.22	0.82	0.61

TABLE 5.9: Evaluation of the fusion of 'cardiffnlp /twitter-roberta-base-sentiment-latest' hidden states (using ASR English translation) with audio features. The original score for the text component is in the 1st row, and the original F1 score for the audio component is in the last column

For hidden states of model with 300m parameters, classifier performed significantly worse, than model trained on hand-crafted ComParE features.

5.3 Bi-modal fusion

To combine audio and text component we conducted experiments involving early (Atmaja, Shirai, and Akagi, 2019) and late fusion (Sahu, 2019) with LR cross-validation. Two models were utilized to extract hidden states of the text component: MoritzLaurer/mDeBERTa-v3-base-mnli-xnl with ASR transcription (see Table 5.10), and cardiffnlp/twitter-roberta-base-sentiment-latest with ASR-generated translation (see Table 5.9). ComParE full functional feature set was used for the acoustic component as well as hidden states (HS) and extracted features (EF) from CNN layers of Wav2Vec2.0 1b and 300m architectures.

In the obtaining results for English translation using 'cardiffnlp /twitter-roberta-base-sentiment-latest' encodings, we obtained slightly better results with an early fusion of text component and ComParE feature set.

For the 'MoritzLaurer/mDeBERTa-v3-base-mnli-xnl' text encodings of ASR transcription (Ukrainian and Russian text), ComParE feature set alone, without fusion, provides the best results.

Features	Fusion	ACC	F1	F1 (audio)
mDeBERTa transcript (T)	-	71.11	0.72	-
T+ComParE	early	**74.44**	**0.73**	**0.75**
T+ComParE	ensemble	70.00	0.69	**0.75**
T+Wav2Vec-1b-CNN	early	70.00	0.71	0.61
T+Wav2Vec-1b-CNN	ensemble	71.11	0.72	0.61
T+Wav2Vec-1b-HS	early	72.22	0.71	0.67
T+Wav2Vec-1b-HS	ensemble	72.22	**0.73**	0.67
T+Wav2Vec-300m-EF	early	71.11	0.72	0.61
T+Wav2Vec-300m-EF	ensemble	70.00	0.70	0.61
T+Wav2Vec-300m-HS	early	68.89	0.68	0.57

TABLE 5.10: Evaluation of the fusion of 'MoritzLaurer /mDeBERTa-v3-base-mnli-xnl' hidden states (using ASR transcription) with audio features. The original score for the text component is in the 1st row, and the original F1 score for the audio component is in the last column

In the case of 'MoritzLaurer/mDeBERTa-v3-base-mnli-xnl' text encodings of ASR transcription (Ukrainian and Russian), the ComParE feature set alone, without fusion, provides the best results.

Additionally, early fusion consistently outperformed the ensemble method in the majority of experiments.

5.4 Execution time evaluation

To thoroughly evaluate the SSC system, it is crucial to consider the resources required to run it in addition to the F1 score and accuracy. We measured the accuracy for each component of the audio-only, acoustic-only, and bi-modal fusion systems. For Wav2Vec2 features we decided to take 1B model, since the difference with 2B model is only 1%, but the model has 2 times less parameters. The results can be found in Table 5.11. The experiments were conducted using Google Colab. For the CPU, we used a "High-RAM" setup with an Intel(R) Xeon(R) CPU @ 2.20GHz (4 cores, 2 threads per core) and 52GB of RAM. For the GPU, we utilized an "L4" setup with 22.5GB of GPU VRAM and 52GB of RAM.

It's interesting to note that ASR models take the longest time to transcribe or translate audio files. On the other hand, text feature extraction is nearly instantaneous. Surprisingly, translating the audio into a different language takes longer than transcribing it into the original language. Additionally, ComParE feature extraction is much faster on a CPU than extracting Wav2Vec2.0 last hidden state. However, the time difference is not as significant when using a GPU.

5.5 Discussion

In this section, we conducted multiple experiments involving ASR systems, LLM classifiers, acoustic hand-crafted features, and Wav2Vec2.0 encodings. Table 5.12 shows the best results for each pipeline. The highest accuracy and F1 for text-only model are obtained using a text-only LLM classifier trained specifically for sentiment tasks using an English dataset. Adding an audio component with ComParE features may enhance this result slightly. For the mix of Ukrainian and Russian languages, the audio model itself with ComParE paralinguistic features provides the best result

Task	Type	Features	Model	N params	CPU time (s)	GPU time (s)
Feature extraction	Audio	ComParE	-	N/a	2	2
Feature extraction	Audio	HS	facebook/wav2vec2-xls-r-1b	1 B	100	3
Feature extraction	Text	HS (translation)	cardiffnlp /twitter-roberta-base-sentiment-latest	119 M	0.2	0.03
Feature extraction	Text	HS (transcript)	MoritzLaurer/mDeBERTa-v3-base-mnli-xnli	265 M	0.4	0.17
ASR	Transcr	-	openai/whisper-large-v2	1.5 B	120	8.5
ASR	Transl	-	openai/whisper-large-v2	1.5 B	100	7.5
Prediction	-	-	Logistic Regression	-	0.4	-

TABLE 5.11: System components execution time for 1 audio file with duration 50 seconds

Modality	Model	ACC	F1	CPU time (s)	GPU time (s)
T (trasl)	cardiffnlp/twitter-roberta..	84.44	**0.85**	100.6	7.93
T (transc)	MoritzLaurer/mDeBERTa..	71.11	0.72	120.8	9.07
A	ComParE	75.56	0.75	**2.4**	**2.4**
A	facebook/wav2vec2-xls-r-1b	67.78	0.67	100.4	3.4
T+A	ComParE+twitter-roberta	**85.56**	**0.85**	102.6	9.93
T+A	ComParE+mDeBERTa	74.44	0.73	120.8	11.07
T+A	wav2vec2-1b+twitter-roberta	84.44	0.84	200.6	10.93
T+A	wav2vec2-1b+mDeBERTa	72.22	0.73	220.8	12.07

TABLE 5.12: Models comparison in terms on accuracy, F1 score and execution time

(see Table 5.10, but the accuracy is 10% lower, compared to the text-only English model. We also compared the execution time for each system by summing up the time required to execute each component (for instance, for bi-modal fusion it will be: ASR + text feature extraction + audio feature extraction + LR). From this perspective, when only the CPU is available, the acoustic-only model using ComParE features significantly outperforms all other models in terms of execution time. This difference is not as significant when using the GPU, but it remains the lowest.

Chapter 6

Conclusions

6.1 Results

In this study, we aimed to address the research gap on SSC task for Ukrainian and mix of Ukrainian and Russian languages. Despite limited data, we achieved 85.5% accuracy through early fusion of text features from the 'cardiffnlp/twitter-roberta-base-sentiment-latest' model (using Whisper large-v2 translations) and ComParE audio features. While an acoustic-only model may yield slightly lower accuracy, it offers advantages in speed, memory efficiency, and a simpler pipeline with fewer potential error accumulated sources. This work demonstrates the feasibility of sentiment classification for under-resourced languages and highlights the trade-offs between different modeling approaches.

The aim initially set by Stream Telecom company was to develop an end-to-end conflict detection system. We made a first step towards this goal by considering the negative sentiment of the conversation as a conflict situation. Our evaluation and inference pipelines are published to GitHub repository [1], while the dataset itself is proprietary and remains private.

6.2 Future work

- **Increasing dataset**. In this work we mainly used Logistic Regression classifiers to avoid overfitting for more complex DNN-based networks. To further improve our results and create a more generalized model, we need to explore such techniques as data augmentation and use pseudo-labeling on the raw dataset.
- **Training more complex models**. After increasing the dataset size, we may increase the generalization of our system by training more complex models, such as LSTMs, or fine-tuning Wav2Vec2.0 models.
- **Time analysis on utterance-level**. In this work, we classified the whole conversation audio files or performed label propagation at the utterance level (which didn't significantly improve the results). Additionally, we may analyze and train the model to recognize patterns of sentiment change in the conversation, which may indicate conflict situations.
- **Analysis in VAD dimension** In this work we used only Valence (positive-negative) dimension, but adding Arousal and Dominance components may possible improve conflict detection system.

[1] https://github.com/lp-ucu/ssc_ukr_ru

Bibliography

Akçay, Mehmet Berkehan and Kaya Oğuz (2020). "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers". In: *Speech Communication* 116, pp. 56–76. DOI: 10.1016/j.specom.2019.12.001.

Atmaja, Bagus Tris, Akira Sasou, and Masato Akagi (2022). "Survey on bimodal speech emotion recognition from acoustic and linguistic information fusion". In: *Speech Communication* 140, pp. 11–28. DOI: 10.1016/j.specom.2022.03.002.

Atmaja, Bagus Tris, Kiyoaki Shirai, and Masato Akagi (2019). "Speech emotion recognition using speech feature and word embedding". In: *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, pp. 519–523. DOI: 10.1109/APSIPAASC47483.2019.9023098.

Badshah, Abdul Malik et al. (2017). "Speech emotion recognition from spectrograms with deep convolutional neural network". In: *2017 international conference on platform technology and service (PlatCon)*. IEEE, pp. 1–5. DOI: 10.1109/PlatCon.2017.7883728.

Baevski, Alexei et al. (2020). "wav2vec 2.0: A framework for self-supervised learning of speech representations". In: vol. 33, pp. 12449–12460. DOI: 10.48550/arXiv.2006.11477.

Bălan, Oana et al. (2019). "Emotion classification based on biophysical signals and machine learning techniques". In: *Symmetry* 12.1, p. 21.

Burkhardt, Felix et al. (2005). "A database of German emotional speech." In: *Interspeech*. Vol. 5, pp. 1517–1520. DOI: 10.21437/Interspeech.2005-446.

Busso, Carlos et al. (2008). "IEMOCAP: Interactive emotional dyadic motion capture database". In: *Language resources and evaluation* 42, pp. 335–359. DOI: 10.1007/s10579-008-9076-6.

Conneau, Alexis et al. (2018). "XNLI: Evaluating cross-lingual sentence representations". In: *arXiv preprint arXiv:1809.05053*.

Demszky, Dorottya et al. (2020). "GoEmotions: A dataset of fine-grained emotions". In: *arXiv preprint arXiv:2005.00547*.

Eyben, Florian, Martin Wöllmer, and Björn Schuller (2010). "Opensmile: the munich versatile and fast open-source audio feature extractor". In: *Proceedings of the 18th ACM international conference on Multimedia*, pp. 1459–1462.

He, Pengcheng, Jianfeng Gao, and Weizhu Chen (2021). "Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing". In: *arXiv preprint arXiv:2111.09543*. DOI: 10.48550/arXiv.2111.09543.

Hsu, Wei-Ning et al. (2021). "Hubert: Self-supervised speech representation learning by masked prediction of hidden units". In: *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29, pp. 3451–3460. DOI: 10.1109/TASLP.2021.3122291.

Ioannides, George et al. (2023). "Towards paralinguistic-only speech representations for end-to-end speech emotion recognition". In: *Proc. INTERSPEECH 2023*, pp. 1853–1857. DOI: 10.21437/Interspeech.2023-497.

Iosifov, Ievgen et al. (2022). "Transferability evaluation of speech emotion recognition between different languages". In: *International Conference on Computer Science, Engineering and Education Applications*. Springer, pp. 413–426.

Laurer, Moritz et al. (2024). "Less annotating, more classifying: Addressing the data scarcity issue of supervised machine learning with deep transfer learning and BERT-NLI". In: *Political Analysis* 32.1, pp. 84–100.

Li, Yuanchao et al. (n.d.). "ASR and Emotional Speech: A Word-Level Investigation of the Mutual Impact of Speech and Emotion Recognition". In: (), pp. 1449–1453. DOI: 10.21437/Interspeech.2023-2078.

Lotfian, Reza and Carlos Busso (2017). "Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings". In: *IEEE Transactions on Affective Computing* 10.4, pp. 471–483. DOI: 10.1109/TAFFC.2017.2736999.

Loureiro, Daniel et al. (2022). "TimeLMs: Diachronic language models from Twitter". In: *arXiv preprint arXiv:2202.03829*.

Lu, Z. et al. (2020). "Speech sentiment analysis via pre-trained features from end-to-end asr models". In: *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7149–7153. DOI: 10.1109/ICASSP40776.2020.9052937.

Metallinou, Angeliki, Sungbok Lee, and Shrikanth Narayanan (2008). "Audio-visual emotion recognition using gaussian mixture models for face and voice". In: *2008 Tenth IEEE international symposium on multimedia*. IEEE, pp. 250–257.

Mohamed, Abdelrahman et al. (2022). "Self-supervised speech representation learning: A review". In: *IEEE Journal of Selected Topics in Signal Processing* 16.6, pp. 1179–1210.

Morais, Edmilson et al. (2022). "Speech emotion recognition using self-supervised features". In: *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6922–6926. DOI: 10.1109/ICASSP43922.2022.9747870.

Mrozek and Danylov (2021). "Speaker recognition and sentiment analysis". MA thesis. Igor Sikorsky Kyiv Polytechnic Institute. URL: https://ela.kpi.ua/bitstream/123456789/46505/1/Mrozek_magistr.pdf.

Radford, Alec et al. (2023). "Robust speech recognition via large-scale weak supervision". In: *International Conference on Machine Learning*. PMLR, pp. 28492–28518. DOI: 10.48550/arXiv.2212.04356.

Ringeval, Fabien et al. (2013). "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions". In: *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)*. IEEE, pp. 1–8. DOI: 10.1109/FG.2013.6553805.

Sahu, Gaurav (2019). "Multimodal Speech Emotion Recognition and Ambiguity Resolution". In: *arXiv preprint arXiv:1904.06022*.

Schuller, Björn et al. (2016). "The interspeech 2016 computational paralinguistics challenge: Deception, sincerity & native language". In: *17TH Annual Conference of the International Speech Communication Association (Interspeech 2016), Vols 1-5*. Vol. 8. ISCA, pp. 2001–2005.

Trigeorgis, George et al. (2016). "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network". In: *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5200–5204. DOI: 10.1109/ICASSP.2016.7472669.

Tzirakis, Panagiotis, Jiehao Zhang, and Bjorn W Schuller (2018). "End-to-end speech emotion recognition using deep neural networks". In: *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, pp. 5089–5093. DOI: 10.1109/ICASSP.2018.8462677.

Vaswani, Ashish et al. (2017). "Attention is all you need". In: vol. 30. DOI: 10.48550/arXiv.1706.03762.

Wagner, Johannes et al. (2023). "Dawn of the transformer era in speech emotion recognition: closing the valence gap". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence*. DOI: 10.1109/TPAMI.2023.3263585.

Wang, Yingzhi, Abdelmoumene Boumadane, and Abdelwahab Heba (2021). "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding". In: *arXiv preprint arXiv:2111.02735*. DOI: 10.48550/arXiv.2111.02735.

Wu, Wen, Chao Zhang, and Philip C Woodland (2021). "Emotion recognition by fusing time synchronous and time asynchronous representations". In: *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 6269–6273. DOI: 10.1109/ICASSP39728.2021.9414880.

Zadeh, AmirAli Bagher et al. (2018). "Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246. DOI: 10.18653/v1/P18-1208.