Audio Analysis of Customer Calls for Predicting Purchase Intentions

A Novel Approach to E-Commerce Insights

By

Miao Yu

B.Eng., Tongji University, 2006

A Project Submitted in Partial Fulfillment of the
Requirements for the Degree of

Master of Engineering

In the Department of Electrical and Computer Engineering

Supervisory Committee

Dr. Kin Fun Li, Department of Electrical and Computer Engineering

Supervisor

Dr. Lin Cai, Department of Electrical and Computer Engineering

Departmental member

Abstract

Client audio recordings represent a valuable resource for many types of businesses. Utilizing these recordings to identify potential customers can help enhance purchase rates and reduce marketing costs, particularly with different kinds of machine learning methods that automatically label different groups, including positive, neutral, and negative buyers, instead of manual analysis. Though previous research has predominantly focused on text content analysis for this purpose, audio features, which effectively capture voice nuances such as tone, pitch, rhythm, and interaction patterns between interviewers and interviewees, may impact the model performance.

This project explored an innovative method. It firstly investigates the effectiveness of emotion detection through audio features, leveraging two datasets: the Toronto Emotional Speech Set (TESS) and the Surrey Audio-Visual Expressed Emotion Dataset (SAVEE). Furthermore, hierarchical clustering techniques are applied to explore the relationship between emotion-related audio features and customer categories using audio data provided by VINN Auto, an e-commerce firm. Next, Exploratory Data Analysis (EDA) is conducted to find the correlation between interaction-related audio features and customer categories, including positive, neutral, and negative buyers within the same dataset after labeling it. Using supervised learning, the results indicate that integrating audio features, including emotion-related and interaction pattern features, can affect the performance of models like Support Vector Machines (SVM), Decision Tree, and Extreme Gradient Boosting (XGBoosts), particularly when combined with traditional audio content-related features such as Term Frequency-Inverse Document Frequency (TF-IDF) scores while applying adjusted weight configuration for positive class. After these exploration, an ensemble method using a soft voting mechanism across these three models is developed to assess whether it can enhance the identification of potential purchasers.

The approach of combining emotion-related audio features, interaction pattern features, and content-based features like TF-IDF scores with tailored weight configurations highlights the value of collaborating audio features in customer identification tasks compared with only using content-based features like TF-IDF scores. It could be a robust strategy for improving classification outcomes for the relevant analysis in the future.

Table of Contents

List of Figures

List of Tables

Acronyms

TESS: Toronto Emotional Speech Set

SAVEE: Surrey Audio-Visual Expressed Emotion Dataset

EDA: Exploratory Data Analysis

SVM: Support Vector Machines

XGBoost: Extreme Gradient Boosting

ASR: Automated Speech Recognition

TF-IDF: Term Frequency-Inverse Document Frequency

FNR: False Negative Rate

CVSSP: Centre for Vision, Speech, and Signal Processing

RMS: Root Mean Square

ZCR: Zero Crossing Rate

F0 Mean: Fundamental Frequency Mean

CMNDF: the Cumulative mean normalized difference function

MFCC: Mel-Frequency Cepstral Coefficient

VAD: Voice Activity Detection

RNN: Recurrent Neural Networks

LSTM: Long Short-Term Memory

Acknowledgements

I am deeply indebted to my supervisor, Dr. Kin Fun Li, whose guidance, expertise, and unwavering support have been invaluable throughout the process of my research. His insightful advice and encouragement not only shaped the direction of my work but also inspired me to push the boundaries of my academic pursuits. I greatly appreciate the time and effort Professor Li invested in our discussions, offering thoughtful critiques and helping me refine my research proposal with care and precision.

In addition, I would like to express my sincere thanks to the members of Dr. Li's lab. Their encouragement and feedback made a significant difference in my research experience. The collaborative environment fostered by Dr. Li's leadership was both intellectually stimulating and supportive, and I am truly thankful to have been part of such a dynamic and inspiring group.

Chaper 1. Background and relevant works

1.1. Project background

With the surge of e-commerce platforms, consumers increasingly turn online to purchase large household goods, such as vehicles, drawn by the convenience of comparing information from multiple dealers in one place. In this context, companies like our industrial partner VINN Auto, operate platforms to offer additional services, such as tracking potential buyers through telephone interviews, facilitating transactions, and expanding sales channels for dealers. However, this business model leads to a notable rise in audio recordings and thus faces the challenging of identifying high-potential customers for follow-up. Traditional solutions for this problem include manual review processes, which are time-consuming and unsuccessful due to human error. Some need to implement basic keyword-based searches within the audio content, which often need more sophistication to identify key purchasing signals accurately.

As the volume of audio recordings grows, more advanced techniques, such as Automated Speech Recognition (ASR) systems combined with machine learning algorithms for audio analysis and intent detection, are becoming increasingly necessary. These approaches improve the efficiency of customer identification and provide insights that can help sales teams prioritize leads and tailor follow-up actions.

1.2. Project objective

The objective of this project is to conduct an exploratory investigation to estimate whether certain audio features related to speech emotion and interaction patterns can enhance several machine learning models' ability to identify potential buyers when combined with content-related features. The ultimate goal is to develop a novel approach for marketing outreach that leverages these combined features to more effectively and automatically target prospective purchasers.

1.3. Relevant works

Previous research on customer identification in business domain has predominantly focused on text content analysis, using techniques such as keyword extraction and phrases frequency analysis to gauge customer intent [1, 2, 3, 4], which leverage textual data from customer interactions but often fail to capture the full spectrum of emotional and interactional dynamics

present in spoken language because they are crucial for understanding deeper customer intent.

In contrast, recent studies have explored the use of audio features to enhance traditional text-based models. These features, including emotion-related characteristics such as tone, pitch, and rhythm, as well as interaction patterns between interviewers and customers, offer a richer and more nuanced representation of customer engagement [5, 6]. By capturing subtleties like frustration, excitement, or hesitation—cues often missed in text alone—audio features provide deeper insights into customer behavior. Research in emotion detection has demonstrated that incorporating these audio features into machine learning models can significantly enhance performance in tasks such as intent recognition and customer classification.

Several notable studies have explored the value of integrating speech emotion recognition into customer behavior models. For example, Mel-Frequency Cepstral Coefficients (MFCCs) [6, 7, 8] and Fundamental Frequency (F0) [9, 10] have been widely used for emotion classification, proving effective in distinguishing between different emotional states in speech. Additionally, interaction-based features, such as participants' gender [11, 12], have been shown to provide critical insights into customer engagement, further refining the ability to predict potential buyers.

Moreover, Automated Speech Recognition (ASR) systems have advanced to the point where accurate transcription of audio data allows for a hybrid approach, combining textual features with audio-derived emotion and interaction patterns. The integration of text content with these audio features [5, 13] has demonstrated substantial improvements in various classification models. These machine learning models benefit from a more holistic view of customer interactions, using audio to enhance traditional content-based analysis.

In summary, previous research has primarily centered on text content analysis [1, 2, 3, 4, 5, 13] or using different types of emotion-related and interaction-pattern features [6, 7, 8, 9, 10, 13], either individually or in combination with other feature groups in customer identification. However, no studies have yet explored the integration of twenty distinct audio features alongside content-related features, specifically TF-IDF scores, for this purpose. Therefore, this project aims to take this novel approach for further analysis.

Chaper 2. Methodology

2.1. Workflow overview

The project overall workflow consists of three main steps, showing in Figure 1 below:

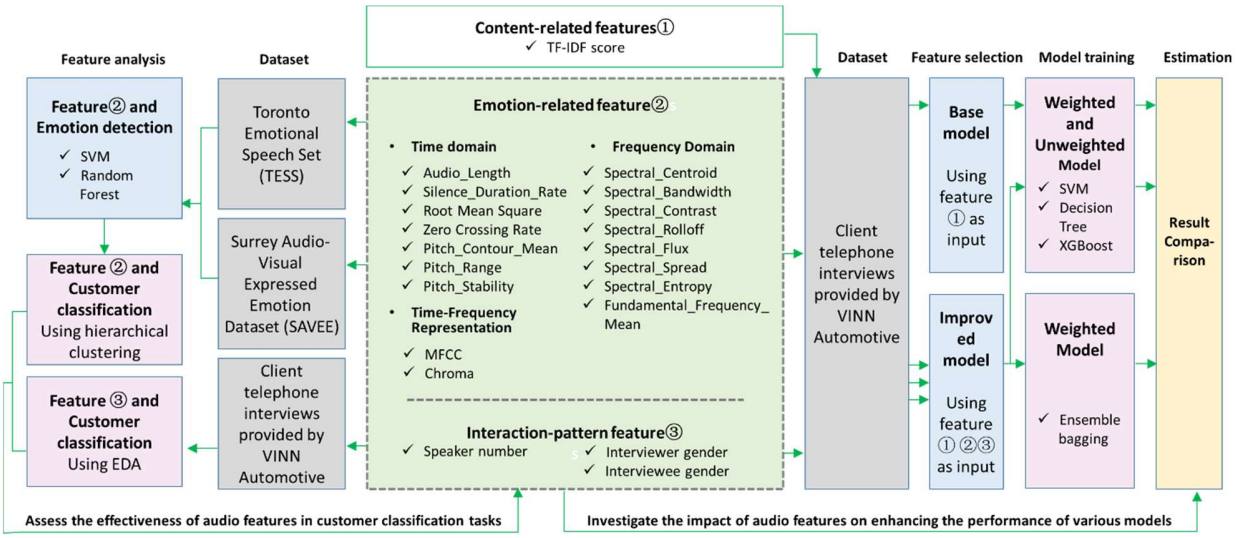

Figure 1 Overall workflow

Firstly, emotion-related audio feature validation are conducted using two datasets: the Toronto Emotional Speech Set (TESS) and the Surrey Audio-Visual Expressed Emotion Dataset (SAVEE) because of their wide use in audio emotion detection tasks and comparable characters such as using distinct characters such as recorder gender and accent. Seventeen distinct types of audio features are extracted from the WAV files of each dataset, covering the time domain, frequency domain, and time-frequency representations. These features are subsequently evaluated using SVM for its computational efficiency and Random Forest for its ability to fight over overfitting to evaluate their effectiveness in distinguishing various emotions. Following this, hierarchical clustering is employed to explore latent relationships between these emotion-related features and three different customer categories, utilizing telephone interview audio data provided by VINN Auto Company. In addition, EDA is applied to uncover potential correlations between interaction-pattern features and customer categories within the same dataset. These sub-workflows, as illustrated on the left side of Figure 1, are vital for evaluating the relevance of both emotion-related and interaction-pattern features in customer classification tasks.

Secondly, content-related features, particularly Term Frequency-Inverse Document Frequency (TF-IDF) scores are introduced as inputs for the base models' training. Classifiers such

as weighted and unweighted SVMs, Decision Trees, and XGBoosts are employed at this stage. Following this, the improved models incorporate additional inputs including emotion-related features and interaction-pattern features along with the TF-IDF scores. Meanwhile, class weight configuration is applied to check whether it could enhance the recall and mitigate the positive customer's False Negative Rate (FNR), followed by the development of an ensemble bagging technique, which is introduced to combine the outputs of the three weighted models including SVM, Decision Tree, and XGBoost, enhancing the robustness of the customer classification results. These steps are depicted in the right section of Figure 1.

Finally, the results of the base models which use only TF-IDF scores and improved models which use twenty types of emotion-related and interaction-pattern features alongside TF-IDF scores with different class weights are evaluated based on overall accuracy, positive recall, and positive FNR, especially in this project where different aspects of model performance provide nuanced insights that a single metric like F2 might not fully capture. As shown in the leftmost part of Figure 1, this step can help conclude the importance of emotion-related and interaction-pattern features implementation in customer classification.

2.1. Dataset

2.2.1. The Toronto Emotional Speech Set (TESS)

The Toronto Emotional Speech Set (TESS) is chosen to evaluate the effectiveness of seventeen types of audio features in emotion detection tasks in this project. Developed by researchers at the University of Toronto, Canada, it was released to support research in emotion detection, speech processing, and related areas.

TESS includes recordings from two female actors who vocalized 200 preselected target words, each expressed across seven emotional categories: Angry, Disgust, Fear, Happy, Sad, Surprise, and Neutral. Each word was spoken by both actors in all seven emotional tones, resulting in a total of 2,800 samples (400 samples per emotion) with speakers aged 22 and 64 years. TESS has become a widely used resource in emotion recognition projects for the research community, forming a basis for machine learning models focused on classifying emotional speech.

2.2.2. The Surrey Audio-Visual Expressed Emotion Dataset (SAVEE)

Similar to TESS, the Surrey Audio-Visual Expressed Emotion Dataset (SAVEE) is used to

further validate the ability of the seventeen types of audio features in emotion classification tasks. Developed at the University of Surrey, UK, in cooperation with the Centre for Vision, Speech, and Signal Processing (CVSSP), SAVEE was created for multimodal emotion recognition research. By capturing both audio and visual data to represent emotions conveyed through voice and facial movements, the dataset supports research aimed at interpreting emotional states in naturalistic contexts. It was chosen for this study because it is one of the most widely used datasets for audio emotion detection and complements TESS by incorporating variations in speaker gender and accent.

SAVEE consists of 480 audio-visual samples, with 120 neutral samples and 60 samples per emotion for six other categories: Angry, Disgust, Fear, Happy, Sad, and Surprise. The dataset got recordings from four male actors aged between 22 and 30, delivering predefined sentences that ensure emotional expressiveness. Although SAVEE includes both audio and visual data, only the audio portion is utilized in this project.

TESS and SAVEE are both used for audio emotion analysis in this project, which characters are summarized in Figure 2.

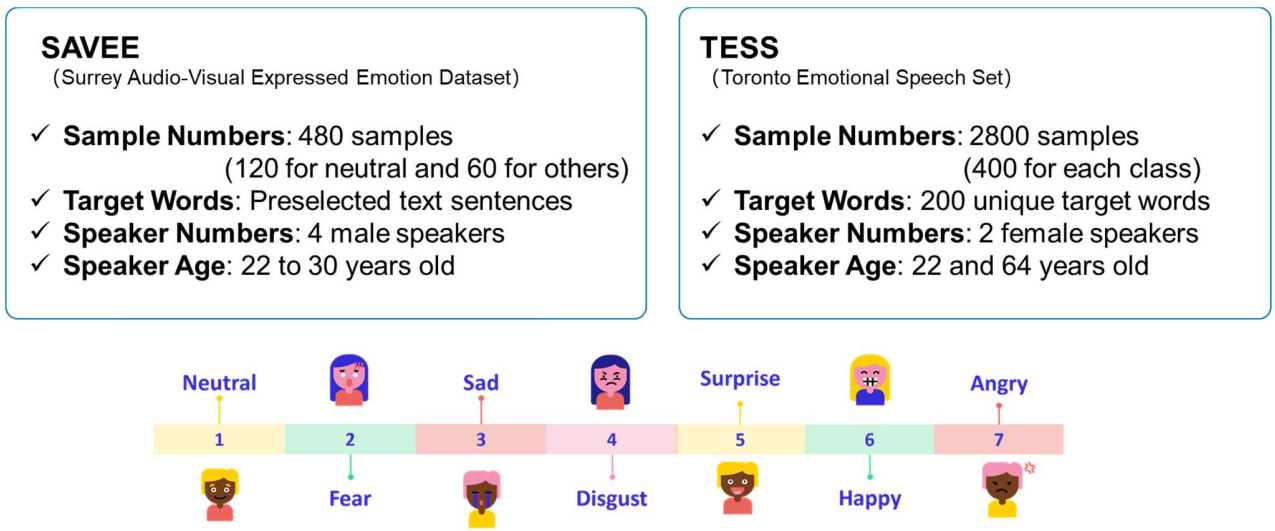

Figure 2 Audio emotion analysis dataset overview

2.2.3. The client telephone interview audios provided by VINN Auto

1) Raw data

The client telephone interview dataset provided by VINN Auto is utilized to analyze the latent relationship between customer categories and relevant audio characters including both seventeen

types of emotion-related features and three types of interaction-pattern features. After that, it is employed for model training and estimation with the aim of identifying high-potential buyers in this project. Developed as part of VINN Auto's services to facilitate transactions on e-commerce platforms for vehicles, this dataset is collected to support research in customer identification and sales elaboration.

The raw dataset consists of 1,809 audio recordings in mp3 format, collected from interactions between customer service representatives at VINN Auto (acting as interviewers) and 391 unique customers (interviewees) between October 1, 2022, and November 15, 2022. Each customer has between 1 to 19 recordings, which vary significantly in both file size and duration. The average length of the recordings is 260 seconds, with a median of 41 seconds. The shortest recording lasts 0 seconds, while the longest extends to 5,784 seconds due to a complex case. On average, the audio files are 11,237 KB in size, with a median size of 1,787 KB. The smallest file is 2 KB, while the largest reaches 249,127 KB. The significant standard deviation value in audio lengths is influenced by the extreme values, particularly the longer recordings, which increase the overall variability, as shown in Table 1. The data shows a strong linear correlation between the length and size of the audio files. On the other hand, the skewed distribution of audio lengths indicates that most audio recordings in the dataset are relatively short because some samples contain little interaction between the interviewers and interviewees. Meanwhile, some other recordings are much longer and larger, reflecting more complex interactions, as illustrated in Figure 3.

	size_KB	length_seconds
mean	11237	260
std	21927	509
min	2	0
25%	243	5
50%	1787	41
75%	12021	279
max	249127	5784

Table 1 Statistical analysis of the raw data

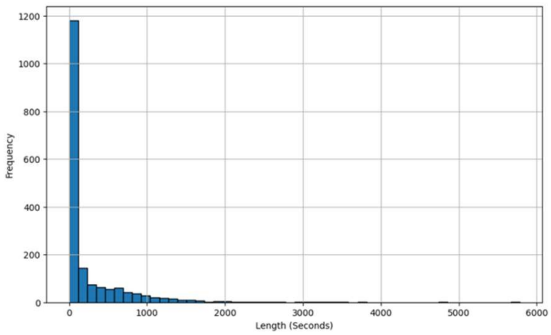

Figure 3 Length distribution of the raw data

2) Cleaned data

The raw dataset shows challenges in analyzing and extracting meaningful insights, because it includes some noise samples that contain little or no valuable information for further investigation.

To address these challenges, specific filtering strategies based on industry cognition are necessary to exclude unqualified samples, thereby enhancing its ability to conduct feature analysis and improving engagement with potential buyers in model training for the next step.

Initially, for the convenience of feature extraction later, all the mp3 audio files were transformed to wav format. After this, two criteria were employed for audio data cleaning. First, audio length greater than 10 seconds was used as a threshold, which resulted in the exclusion of 616 files. This step was useful because the audio, which was shorter than 10 seconds, contained too little information for meaningful analysis. Second, the number of speakers in the audio was required to be greater than 1, which excluded 624 files. This is because files with only 0 or 1 speaker often included scenarios where nobody was speaking or only one person speaking without a response. However, these two filters could not exclude those samples, including automatic responses from the mailbox with no direct interaction. They are left for further analysis because they can still be potential buyers.

Some samples met both exclusion criteria, and after applying these filters, a total of 1,114 audio files remained from the original set of 1,809. After filtering, the mean length of the cleaned audio data is now 418 seconds, with a median of 166 seconds, the shortest of 11 seconds, and the longest remaining 5,784 seconds, as displayed in Table 2. The histogram of audio lengths in Figure 4 shows a distribution where the majority of the recordings remain relatively short. At the same time, removing some samples with little or no meaningful information notably decreased the standard deviation and, therefore, could significantly improve the dataset's suitability for feature analysis and model training in the subsequent phases of the project.

	length_seconds
mean	418
std	596
min	11
25%	47
50%	166
75%	603
max	5784

Table 2 Statistical analysis of the cleaned data

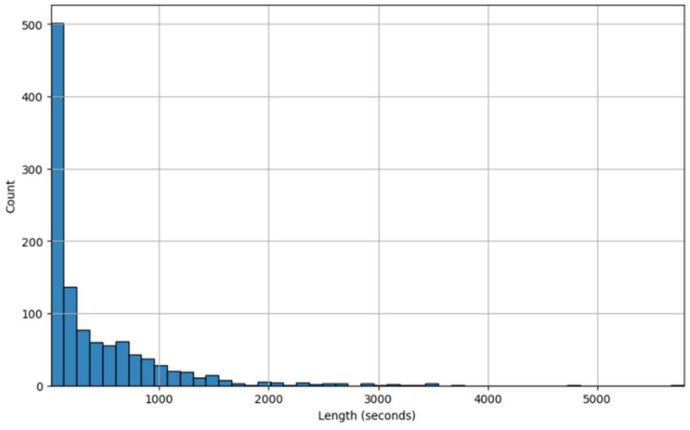

Figure 4 Length distribution of the cleaned data

3) Manual labeling

Although the VINN Auto platform has established an electronic profile for each client, including category labels indicating their purchasing trends, the clients uploaded the first and last names in the database without any form of validation. As a result, many of these names are either nicknames, abbreviations, or even left blank. This resulted in the challenge of matching the profiles accurately with the audio files, which often use the clients' real names as their identifiers. Additionally, even for the recordings that can be matched successfully, the platform's profiles only assign a certain status to each client, including six types (dead, deal, unsubscribed, sale, qualified, unqualified, unqualified). This creates further complexity, as the profile system contains too many subtypes while multiple audio files for a single customer may reflect different labels, capturing changes that occurred throughout the interview process. Due to these challenges of using the platform's original labels, a manual labeling process was necessary to create a clearer dataset, enabling correlation analysis between features and customer categories as well as accurate supervised learning in the next steps.

As shown in Table 3, manual labeling was divided into three classes: positive, negative, and neutral for further clarification, with each category guided by specific rules.

- Positive labels were utilized when the customer expressed a clear intent to engage in the buying process. This included actions such as wanting to book a test drive, making an appointment with a vehicle dealer, showing explicit interest in purchasing a car or asking to be called back later for further discussion.

- Negative labels were applied when the customer indicated disinterest or inability to proceed with a purchase. This included situations where the customer had no immediate plans or mentioned they might consider it in the future, lacked sufficient budget, had already purchased a car, or complained about the service.

- Neutral labels were assigned when there was insufficient information available for classification, such as when the interaction lacked meaningful content or when the call reached a different person other than the intended target.

By applying manual labeling with two people involved (one for labeling and one for checking using the same standard, as shown in Table 3 to mitigate bias), the majority of the audio files,

approximately 68.9% (768 files), fell into the positive category, and 23.1% (257 files) were labeled as neutral, and a smaller portion . In contrast, 78.0% (859 files), were labeled as negative, as shown in Figure 5. This results in a typically imbalanced dataset, which requires consideration during model training.

Label	Rules
positive	Want to book a test drive or make an appointment with a vehicle dealer
	Show explicit plan to purchase a car
	Show interest for a specific kind of car
	Ask to be called back later
negative	Have no plan right now or will consider it in the future
	Do not have enough budget
	Have already purchased a car
	Complain about the service
neutral	Lack of sufficient information
	Reach a different person other than target

Table 3 Label categories and labelling rules

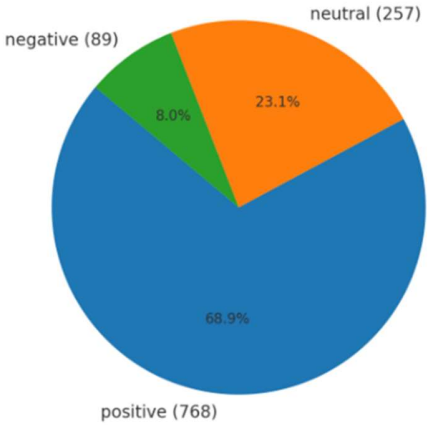

Figure 5 Label distribution of the cleaned data

Chaper 3. Feature selection and data mining

3.1. Audio emotion-related feature selection and validation

3.1.1. Audio splitting

Initially, three speaker-based spitting methods were explored to extract more relevant features from the interviewees' audio in the recordings provided by VINN Auto: audio splitting based on silent chunks [14], audio splitting based on speaker diarization [15] and the deep learning end-to-end models [16]. However, the first method, specifically audio splitting based on silent chunks, faced issues when dealing with overlapping speech, while the second method of splitting based on speaker diarizationor encountered inaccuracy for those samples with speakers of the same gender. Meanwhile, although deep learning end-to-end models were also employed, they proved too time-consuming due to the extensive tuning required for handling complex audio samples, especially the longer audio. Considering the limitation of our resources and the priority of the project's objective, audio splitting was deferred to future work, and the original recordings, including both interviewer and interviewee audio, were used for audio feature extraction in this project.

3.1.2. Feature selection

Audio or sound can be seen as a type of signal that exhibits characteristics in both the time and frequency domains, which are commonly represented by waveforms and frequency spectrums, as discussed in the work of O'Shaughnessy [17] and Rabiner and Schafer [18]. Meanwhile, to further simulate the human auditory system's sensitivity, particularly its higher sensitivity in lower frequencies, sound signals are often transformed into Mel-spectrogram, which adjusts the frequency scale to create a time-frequency representation that compresses higher frequencies and expands lower ones, providing features that more accurately reflect perceptually relevant characteristics. Such a transformation enables the representation to reflect how humans perceive signals while maintaining the temporal changes.

In this project, seven types of time domain features and eight types of frequency domain features, as well as two types of time-frequency representation features, are selected to evaluate their ability to enhance the identification of potential purchasers, particularly in tasks related to emotion detection, as displayed in Figure 6. The selection of these features, comprising a total of 46 values (7 for spectral contrast, 13 for MFCC and 12 for Chroma) is based on their widespread

use in various works.

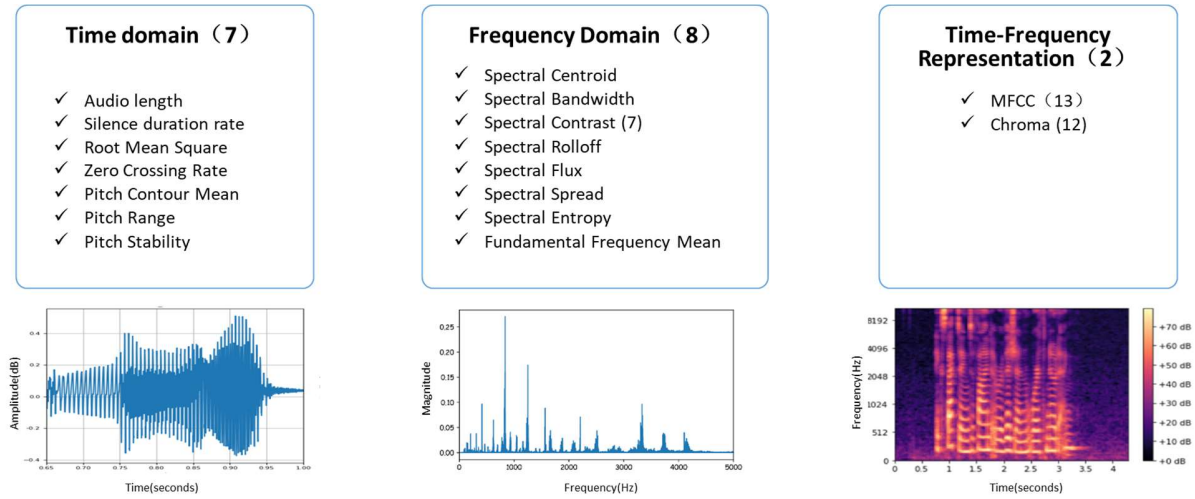

Figure 6 Summary of emotion-related audio feature selection

1) Time domain features

- Audio length

The audio length refers to the total duration of the audio file in seconds [7, 8, 10]. Longer or shorter interactions may indicate varying degrees of emotional engagement or stress levels, where prolonged interactions could reflect more complex emotional exchanges.

The audio length is computed as:

$$\text{Length} = \frac{N}{f_s}$$

, where N is the total number of samples and f_s is the sampling rate.

- Silence duration rate

The silence duration rate measures the proportion of silence within the audio signal, where silence is defined as audio segments below a certain amplitude threshold [6] (20 dB in this project). Longer pauses often signify discomfort or contemplation, while shorter pauses might indicate confidence or assertiveness.

The silence duration rate is computed as:

$$\text{Silence Duration Rate} = \frac{\sum_{i=1}^{N_s} T_s}{T}$$

, where *Ts* represents the duration of each silent segment,

Ns is the number of silent segments, and *T* is the total duration of the signal.

- Root Mean Square (RMS)

RMS measures the average power or intensity of the audio signal [19, 20], which is often linked to emotional arousal. Louder speech indicates excitement or anger while softer speech indicates calmness or sadness.

RMS is calculated as:

$\mathrm{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2}$, where x_i represents the amplitude of the audio signal at sample i, and N is the total number of samples.

- Zero Crossing Rate (ZCR)

ZCR measures the rate at which the signal crosses the zero amplitude line, providing insight into the roughness or noisiness of the signal [19, 20, 21]. ZCR is often higher for more energetic or aggressive speech, such as anger, and lower for calm or sad speech, making it a useful feature in emotion recognition [22].

ZCR is calculated as:

$\mathrm{ZCR} = \frac{1}{N-1}\sum_{i=1}^{N-1} 1_{[x_i \cdot x_{i+1} < 0]}$, where x_i represents the amplitude (or value) of the signal at the i-th sample (point in time) and x_{i+1} refers to the amplitude of the signal at the next sample $i+1$.1 is an indicator function that is 1 if the signal crosses zero between consecutive samples x_i and x_{i+1}, and 0 otherwise. N is the total number of samples.

- Pitch contour mean

Pitch contour mean measures the average pitch over time, representing how pitch (frequency) evolves throughout the audio signal. Higher pitch contour is typically associated with excitement or stress, while lower pitch contour might indicate calmness or sadness. Variations in pitch contours can be used to identify both the speaker's emotional valence and arousal [6, 21].

Pitch contour mean is calculated as:

$$\text{Pitch Contour Mean} = \frac{1}{N}\sum_{i=1}^{N} f_0(i)$$, where $f_0(i)$ is the fundamental frequency (pitch) at time step i, and N is the number of pitch values.

- Pitch range

Pitch range captures the difference between the highest and lowest pitch values, indicating the dynamic changes in pitch across time [19, 20]. A larger pitch range indicates more emotional expressivity, where emotions such as anger, fear, or surprise might show a higher variance in pitch compared to neutral or calm states.

Pitch range is calculated as:

$$\text{Pitch Range} = \max(f_0) - \min(f_0)$$, where max (f_0) and min (f_0) are the maximum and minimum pitch values, respectively.

- Pitch stability

Pitch stability assesses the consistency of the pitch throughout the audio, measuring how much the pitch fluctuates over time [21]. Emotional instability, such as nervousness or excitement, may result in fluctuating pitch values, whereas a stable pitch is often linked to calmness and control. Scherer [22] showed that varying pitch patterns are closely related to emotional states in speech.

Pitch stability is calculated as:

$$\text{Pitch Stability} = \sigma(f_0)$$, where σ (f_0) is the standard deviation of the pitch values.

2) Frequency domain features

- Spectral centroid

The spectral centroid indicates the "center of mass" of the spectrum and is associated with the brightness or sharpness of a sound. Emotional states like anger or excitement are often characterized by higher centroids due to the increased energy in higher frequencies, while sadness is typically associated with lower centroids [6, 7].

Spectral centroid is calculated as:

$$\text{Spectral Centroid} = \frac{\sum_{i=1}^{N} f_i X(f_i)}{\sum_{i=1}^{N} X(f_i)}$$, where f_i is the frequency of the i-th bin, and $X(f_i)$ is the magnitude of the spectrum at that frequency.

- Spectral bandwidth

Spectral bandwidth measures the spread of the spectrum, capturing the width of the frequency distribution. Sounds with wider bandwidths, such as those with richer harmonic content, are often associated with intense emotions, whereas narrower bandwidths may correspond to more subdued or monotonous emotional states [6, 19].

Spectral bandwidth is calculated as:

$$\text{Spectral Bandwidth} = \sqrt{\frac{\sum_{i=1}^{N} (f_i - C)^2 X(f_i)}{\sum_{i=1}^{N} X(f_i)}}$$, where f_i is the frequency of the i-th bin (specific frequency), $X(f_i)$ is the magnitude of the spectrum at that frequency, C is the average frequency of the spectrum, indicating the "center" of the spectral mass and N is the total number of frequency bins in the spectrum.

- Spectral Contrast

Spectral contrast measures the difference between spectral peaks and valleys across multiple frequency bands, representing the harmonic or inharmonic content of the signal [21]. By dividing the frequency spectrum into several bands (7 in this project), spectral contrast provides detailed insights into how different frequency ranges behave in the audio signal. High contrast is often linked to emotions such as surprise or fear, where the sound varies more dramatically in amplitude.

The k-th Spectral Contrast is computed as:

$$\text{Spectral Contrast}_k = \frac{\text{Peak Magnitude}_k}{\text{Valley Magnitude}_k}$$, where "Peak Magnitude$_k$" refers to the amplitude of the spectrum peaks in the k-th frequency band, and "Valley Magnitude$_k$" refers to the amplitude of the spectrum valleys in the k-th frequency band..

- Spectral Rolloff

Spectral rolloff describes the frequency below which a certain percentage (typically 85%) of

the total spectral energy lies. Higher rolloff values are linked to sharper, more energetic sounds, which may correlate with emotions like anger or excitement, while lower rolloff values can indicate softer, calmer sounds [6, 7].

Spectral rolloff is calculated as:

$$\text{Spectral Rolloff} = f_i \text{ such that } \sum_{i=1}^{k} X(f_i) = 0.85 \sum_{i=1}^{N} X(f_i)$$

, where f_i is the frequency of the i-th bin (specific frequency), $X(f_i)$ is the magnitude of the spectrum at that frequency and N is the total number of frequency bins in the spectrum.

- Spectral Flux

Spectral flux represents the rate of change in the power spectrum. Rapid changes in the spectral flux are often indicative of emotional arousal or volatility, as seen in emotions like fear, excitement, or surprise [6, 7].

Spectral flux is calculated as:

$$\text{Spectral Flux} = \sum_{i=1}^{N} (X(f_i, t) - X(f_i, t-1))^2$$

,where $X(f_i,t)$ is the magnitude of the spectrum at frequency f_i at time frame t, $X(f_i,t-1)$ is the magnitude of the spectrum at frequency f_i at the previous time frame $t-1$, and N is the total number of frequency bins in the spectrum.

- Spectral Spread

Spectral spread measures how widely frequencies are distributed around the spectral centroid [21]. Similar to bandwidth, higher spectral spread values indicate more complex or sharp sounds, often associated with heightened emotional states such as anger, excitement, or surprise.

Spectral spread is calculated as:

$$\text{Spectral Spread} = \frac{\sum_{i=1}^{N} (f_i - C)^2 X(f_i)}{\sum_{i=1}^{N} X(f_i)}$$

, where f_i is the frequency of the i-th bin (specific frequency), $X(f_i)$ is the magnitude of the spectrum at that frequency, C is the Spectral centroid and N is the total number of frequency bins in the spectrum.

- Spectral Entropy

Spectral entropy measures the complexity or unpredictability of a sound. Higher entropy is associated with chaotic or unpredictable sounds, which may correspond to emotions like anger, fear, or confusion, while lower entropy reflects more predictable, calm emotional states [7].

Spectral entropy is calculated as:

$$\text{Spectral Entropy} = -\sum_{i=1}^{N} p(f_i) \log p(f_i)$$

, where $p(f_i)$ is the normalized power of frequency bin f_i.

- Fundamental Frequency Mean (F_0 Mean)

F_0 is used to track instantaneous pitch at each moment in a signal, providing a time-varying function that shows how pitch changes throughout the signal. Higher F_0 mean values are often linked to increased emotional arousal, such as in fear, surprise, or excitement [6, 7, 8, 9]. In this project, YIN algorithm in librosa.pyin are employed to compute F_0.

F_0 is calculated as:

$$F_0 = \frac{1}{\tau_{\min}}$$

, where τ_{min} is the lag value that minimizes the cumulative mean normalized difference function (CMNDF) after thresholding.

To get the minimizer for CMNDF, there are three steps to follow:

Firstly, calculate the value for the difference function:

$$d(t,\tau) = \sum_{j=0}^{W-1} \left(x(t+j) - x(t+j+\tau)\right)^2$$

, where $x(t)$ is the amplitude of the audio at time t, τ is the time lag (used to measure how much the audio is shifted), W is the window size over which the difference function is computed (typically the length of a frame in the signal).

Then, compute the CMNDF, which ensures that the difference function decreases over time, making the pitch period more prominent compared to lower-lag values, using the following function:

$\mathrm{CMNDF}(\tau) = \frac{d(\tau)}{\frac{1}{\tau}\sum_{k=1}^{\tau} d(k)}$, where $d(\tau)$ is the difference function value at lag τ and $\frac{1}{T}\sum_{k=1}^{T} d(k)$ is the cumulative mean of the difference function up to lag τ.

Lastly, a threshold is applied to identify the minimum value of τ that corresponds to the pitch period with the function as below:

$\text{If } \mathrm{CMNDF}(\tau) < \text{Threshold, select } \tau \text{ as pitch period}$, where the threshold is 0.1 in this project.

3) Time-frequency representative

- Mel-Frequency Cepstral Coefficient (MFCC)

MFCCs capture the spectral envelope of a signal and are designed to approximate how the human ear perceives sound. Their ability to capture fine-grained information about both frequency and time makes them highly effective in identifying emotional nuances in speech. For example, higher-order MFCCs have been shown to correlate with emotional expressions such as anger, sadness, and excitement by reflecting subtle changes in tone and timbre. This makes MFCCs particularly valuable for emotion recognition tasks, where detecting these variations is essential [6, 7, 8, 9].

The *n*-th MFCC is computed as:

$MFCC_n = \sum_{k=1}^{K} \log X_k \cos\left[\frac{\pi n(k-0.5)}{K}\right]$, where X_k is the magnitude of the Fourier transform at bin *k*, *K* is the number of bins, and *n* represents the index of the MFCC coefficient (13 in this project).

- Chroma

Chroma features capture the harmonic content of the signal by mapping pitches into 12 distinct classes [21]. This is particularly useful for detecting tonal shifts, which may be associated with emotional modulation. For example, shifts in harmonic content can signal surprise or tension in speech.

The *k*-th Chroma is computed as:

$$\text{Chroma}_k = \frac{\sum_{f \in F_k} X(f)}{\sum_{f \in F} X(f)}$$

, where F_k is the set of frequencies corresponding to pitch class *k*, and *X(f)* is the magnitude of the spectrum at frequency *f*.

Feature type		Impact on emotion detection
Time domain	Audio length	It may indicate varying degrees of emotional engagement or stress levels
	Silence duration rate	Longer rates often signify discomfort or contemplation, while shorter ones might indicate confidence or assertiveness
	Root Mean Square	Louder RMS indicates excitement or anger while softer RMS indicates calmness or sadness
	Zero Crossing Rate	ZCR is often higher for more energetic or aggressive speech, such as anger, and lower for calm or sad speech
	Pitch Contour Mean	Higher pitch contour is typically associated with excitement or stress, while lower pitch contour might indicate calmness or sadness
	Pitch Range	Emotions such as anger, fear, or surprise might show a higher variance in pitch compared to neutral or calm states
	Pitch Stability	Emotional instability, such as nervousness or excitement, may result in fluctuating pitch values, whereas a stable pitch is often linked to calmness and control.
Frequency Domain	Spectral Centroid	Emotional states like anger or excitement are often characterized by higher centroids while sadness is typically associated with lower centroids
	Spectral Bandwidth	Sounds with wider bandwidths, such as those with richer harmonic content, are often associated with intense emotions, whereas narrower bandwidths may correspond to more subdued or monotonous emotional states
	Spectral Contrast	High contrast is often linked to emotions such as surprise or fear, where the sound varies more dramatically in amplitude
	Spectral Rolloff	Higher rolloff values are linked to sharper, more energetic sounds, which may correlate with emotions like anger or excitement, while lower rolloff values can indicate softer, calmer sounds
	Spectral Flux	Rapid changes in the spectral flux are often indicative of emotional arousal or volatility, as seen in emotions like fear, excitement, or surprise
	Spectral Spread	Higher spectral spread values indicate more complex or sharp sounds, often associated with heightened emotional states
	Spectral Entropy	Higher entropy is associated with chaotic or unpredictable sounds, which may correspond to emotions like anger, fear, or confusion, while lower entropy reflects more predictable, calm emotional states
	Fundamental Frequency Mean	Higher F0 mean values are often linked to increased emotional arousal, such as in fear, surprise, or excitement
Time-Frequency Representation	MFCC	Higher-order MFCCs have been shown to correlate with emotional expressions such as anger, sadness, and excitement by reflecting subtle changes in tone and timbre
	Chroma	Shifts in harmonic content can signal surprise or tension in speech

Table 4 Summary of the emotion-related features' influence on emotion detection

3.1.3. Feature validation

1) Correlation with emotion

To validate the effectiveness of the selected seventeen audio features in emotion detection, two datasets were employed: TESS (Toronto Emotional Speech Set) and SAVEE (Surrey Audio-Visual Expressed Emotion Dataset), both of which contain audio files in .wav format for feature extraction. After extracting the features, two classifiers—SVM and Random Forest—were applied to evaluate the emotion classification performance.

This research utilizes a different set of features as input for classification compared to the previous studies and thus makes direct result comparisons challenging. Focusing on itself, the overall classification accuracy across different datasets and classifiers is summarized in Table 5. The results show that the models generally perform better on the TESS dataset compared to

SAVEE, regardless of the input features. Additionally, while Principal Component Analysis (PCA) features lead to a slight drop in accuracy on TESS, its performance on SAVEE declines notably. Using original features after standardization, SVM achieved a high accuracy of 97.68% on TESS and 79.17% on SAVEE. Random Forest performed similarly, with 98.57% accuracy on TESS but lower on SAVEE, reaching only 57.29%. When PCA-transformed features were employed (retaining 90% of the variance after standardization), the performance remained robust on TESS, with SVM achieving 96.07% and Random Forest reaching 97.86%. However, accuracy on SAVEE declined sharply, with SVM dropping to 58.33% and Random Forest decreasing to 52.08%.

Accuracy			
Input	Classifier	TESS	SAVEE
Original feature	SVM	97.68%	79.17%
	Random forest	98.57%	57.29%
PCA feature	SVM	96.07%	58.33%
	Random forest	97.86%	52.08%

Table 5 Overall accuracy for each dataset using different features and classifiers

In terms of different emotion categories, both SVM and Random Forest demonstrate stronger capabilities in identifying emotions like neutral and sad across both datasets, with some fluctuations among other emotions. As illustrated in Table 6, SVM and Random Forest consistently present strong performance across all emotion types in the TESS dataset. Specifically, both classifiers achieved near-perfect F1-scores for emotions such as neutral, sad, fear, and angry, with minor variations observed in the other three categories—disgust, happiness, and surprise. However, as shown in Table 7, the performance was significantly lower on the SAVEE dataset, especially for more subtle emotions like disgust and surprise. Particularly, Random Forest struggled to classify some emotions accurately on SAVEE, with F1-scores as low as 0.19 and 0.25 for surprise using original features and angry using PCA-transformed features.

F1-score of TESS								
Input	Classifier	Emotion type						
		angry	disgust	fear	happy	neutral	sad	surprise
Original feature	SVM	0.99	0.97	0.99	0.95	1.00	1.00	0.94
	Random forest	0.99	0.97	0.99	0.99	0.99	0.99	0.97
PCA feature	SVM	0.97	0.97	0.98	0.91	1.00	1,00	0.89
	Random forest	0.98	0.99	0.99	0.96	1.0	1.00	0.94

Table 6 F1 score for TESS dataset using different features and classifiers

F1-score of SAVEE								
Input	Classifier	Emotion type						
		angry	disgust	fear	happy	neutral	sad	surprise
Original feature	SVM	0.63	0.62	0.61	0.67	0.90	0.73	0.77
	Random forest	0.50	0.58	0.42	0.56	0.78	0.63	0.19
PCA feature	SVM	0.53	0.50	0.50	0.42	0.79	0.63	0.46
	Random forest	0.25	0.29	0.55	0.48	0.69	0.67	0.37

Table 7 F1 score for SAVEE dataset using different features and classifiers

Overall, the results demonstrate that the seventeen audio features are validated for emotion detection tasks and can, therefore, be classified as emotion-related audio features, although for some classifiers, performance remains suboptimal. Specifically, the TESS dataset with standardized structure, female speakers, clear emotion samples, and larger sample size, performed better in emotion classification tasks compared to the more complex SAVEE dataset, which includes male voices, British accents, fewer samples, and varied sentence structures, which likely contributed to the reduced accuracy on SAVEE. On the other hand, while PCA-transformed features maintained strong performance on TESS, they led to a notable decline in accuracy on SAVEE, likely due to the loss of certain latent cues during feature reduction. Meanwhile, the Random Forest classifier encountered difficulties in identifying certain emotions in SAVEE, such as surprise. This difficulty may arise from its reliance on multiple Decision Trees, which require a sufficient number of representative samples to effectively capture all potential patterns. This observation suggests that introducing alternative classifiers may improve the result, especially for the datasets with fewer and more diverse samples.

Moving forward, this project will focus on using the original audio features for feature analysis and model training, rather than PCA-transformed features, and will explore classifiers beyond Random Forest, considering the limited sample size in the client telephone interview audios provided by VINN Auto, which contains only 1,114 qualified cases after filtering and having similarities with SAVEE.

2) Correlation with customer categories

In this stage, seventeen emotion-related audio features were extracted from the cleaned client telephone interview data provided by VINN Auto. After standardization, which involves transforming each feature to have a mean of 0 and a standard deviation of 1, hierarchical clustering was performed using two approaches: agglomerative clustering (Bottom-Up) and divisive clustering (Top-Down). The silhouette score was then used to determine the optimal number of

clusters. The dataset was then labeled based on these optimal clusters. By identifying the emotion-related feature clusters that contributed most significantly to the manually labeled customer categories, further analysis of the relationship between the feature clusters and customer categories was facilitated.

- Hierarchical agglomerative clustering

In the exploration of hierarchical agglomerative clustering, Ward's method was applied to group samples based on emotion-related features, which follows a bottom-up strategy, where each data point initially forms its own cluster. Then Clusters are iteratively merged in pairs by minimizing the increase of within-cluster variance, referring to the sum of squared differences. This process continues until all data points are combined into a single cluster. It produces a group of nested clusters that effectively capture the underlying structure within the data.

The within-cluster variance is computed as:

$$W(C) = \sum_{i \in C} \|x_i - \mu_C\|^2$$

, where C is a cluster, x_i is a point in cluster μ_C is the mean (centroid) of the cluster and $\| x_i - \mu_C \|^2$ is the squared Euclidean distance between the point x_i and the centroid μ_C.

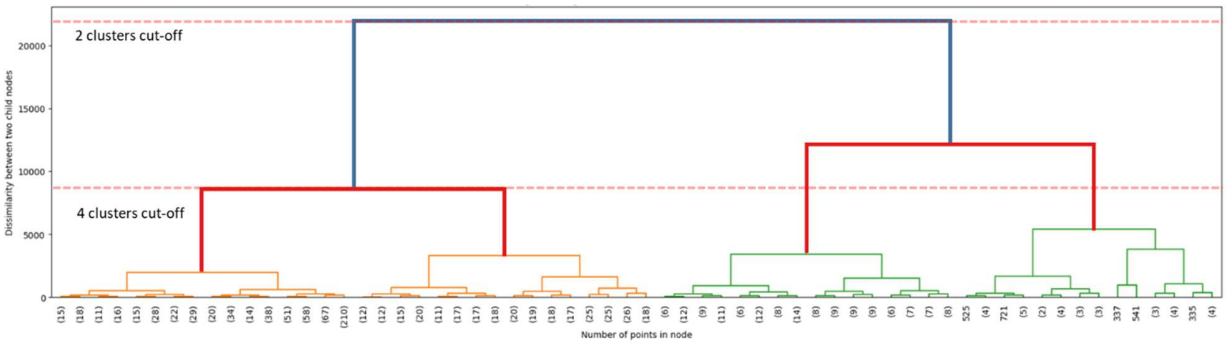

Figure 7 Agglomerative clustering dendrogram based on Ward's method with Euclidean distance

As shown in Figure 7, the dendrogram visualization highlights the longest vertical distances (thick blue and red lines) occur at the height of around 22000 and 9000, which suggest that the dataset can be divided into two or four main clusters with significant separation, achieved by minimizing within-cluster variance increases. The x-axis shows the sample numbers included in each class after truncation. The total sum of the x-axis is larger than the sample size because some data points are counted multiple times due to the specific hierarchical structure.

To further decide this cluster number, the calculation of silhouette score is introduced at this stage for each cluster number to support a clearer visual analysis. The silhouette score measures the similarity of an object to its own cluster (cohesion) compared to other clusters (separation), with values ranging from -1 to 1. A value close to 1 indicates that this point is well-clustered, with high cohesion within its own cluster and good separation from other clusters, while a value close to 0 indicates that the point lies very close to the decision boundary between two clusters. A negative value suggests that the point has been misclassified and is closer to a different cluster than the one it is assigned to.

For each data point iii, the silhouette score S(i) is calculated as:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

, where a(i) is the average distance from point i to all other points in the same cluster (intra-cluster distance or cohesion), b(i) is the average distance from point i to all points in the nearest cluster (inter-cluster distance or separation), and max(a(i), b(i)) ensures the score is normalized between -1 and 1.

Once the silhouette score for each point in the cluster is calculated, the silhouette score for the entire cluster is the average of the silhouette scores for all points in that cluster as below:

$$S(C) = \frac{1}{|C|} \sum_{i \in C} S(i)$$

, where $|C|$ is the number of points in cluster C, and S(i) is the silhouette score of each point in that cluster.

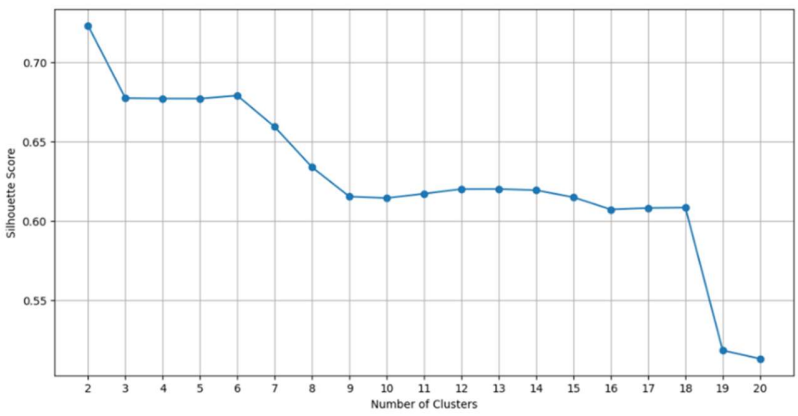

Figure 8 Agglomerative clustering silhouette scores

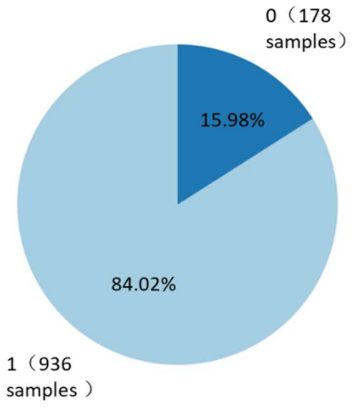

Figure 9 Agglomerative clustering label count and proportion

The overall silhouette score is the mean of the silhouette scores of all points across all clusters. It indicate a peak at 2 clusters, with the a relatively high value of 0.7232, which suggests that the samples are well-separated and cohesively grouped. As a result, this 2 cluster is selected for labeling all samples subsequently. Figure 9 further illustrates the distribution of the labels according to the given 2 clusters, with class 1 containing the 84.02% of the data(936 samples), and class 0 consisting of 15.98% (178 samples).

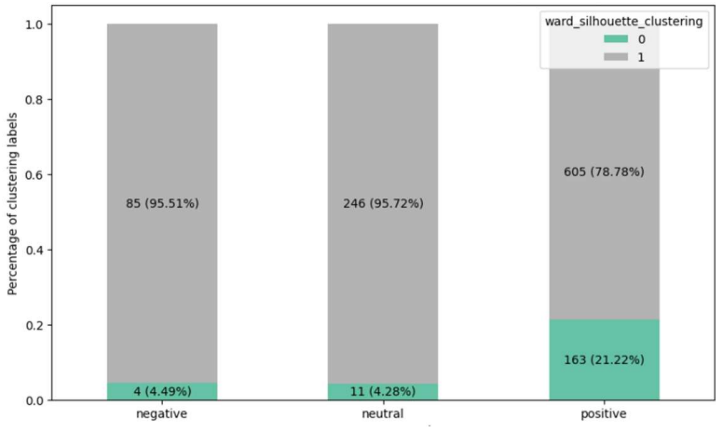

Figure 10 Relationship between agglomerative clustering labels and customer categories

In the next step, the relationship between clustering labels and customer categories was analyzed, as demonstrated in Figure 10. The stacked bar chart reveals that for the "negative" and "neutral" customer categories, class 1 (gray) overwhelmingly dominates, with 95.51% and 95.72% of the samples, respectively. Conversely, in the "positive" category, a larger portion of samples is found in class 0 (green), accounting for 21.22% of the total, while class 1 still holds the majority at 78.78%. This distribution suggests a potential correlation between customer categories and the classes formed by the agglomerative clustering method, indicating that positive customer category are more likely to be classified into class 0 compared to those with the type of negative or neutral.

- Hierarchical divisive clustering

In the exploration of hierarchical divisive clustering, instead of starting with each data point as its own cluster and merging them step by step, it starts with all data points in one cluster and recursively splits them into smaller clusters, following a top-down strategy. This approach aims to maximize dissimilarity within the cluster at each step. At each iteration, the cluster with the highest internal variance or dissimilarity is selected for splitting. Points within this cluster are divided into two subclusters based on their distance from the most dissimilar point, forming a binary split. This process continues recursively, producing a hierarchy of progressively finer clusters. The result is a nested structure of clusters that provides insight into the natural divisions within the data, similar

to the bottom-up approach, but starting from the most general grouping and refining toward more specific divisions.

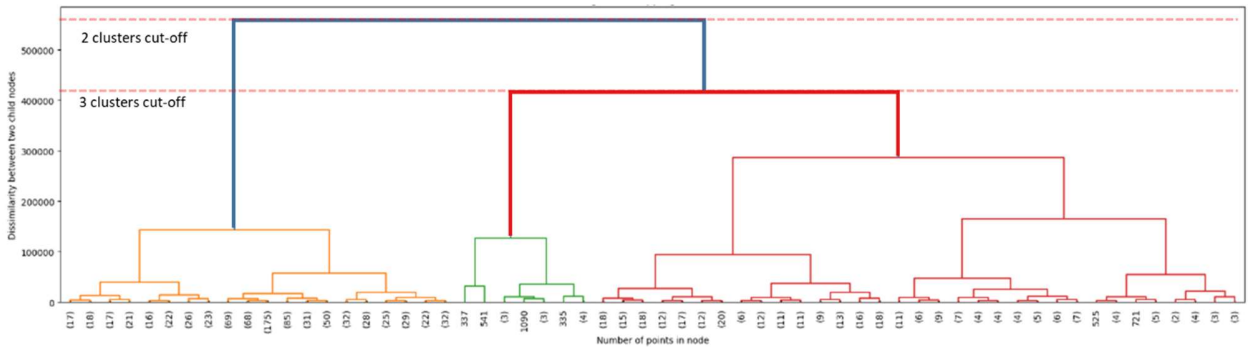

Figure 11 Divisive clustering dendrogram based on Ward's method with Euclidean distance

The dendrogram visualization in Figure 11 shows the most significant distances (thick blue and red lines) occur at the height of around 530000 and 410000, indicating that the dataset can be grouped into two or three main clusters with substantial separation, which reflects notable within-cluster variance differences between each group.

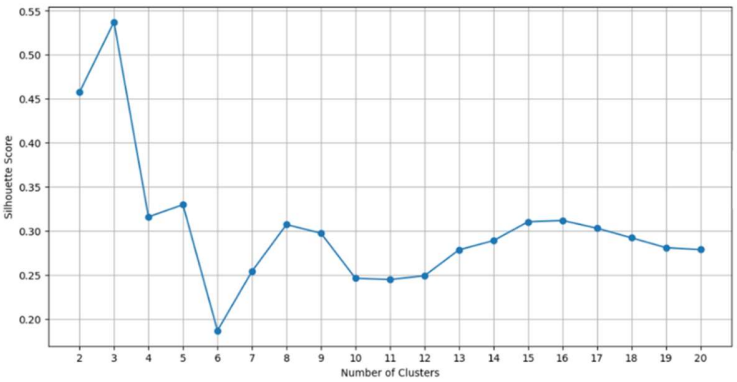

Figure 12 Divisive clustering silhouette scores

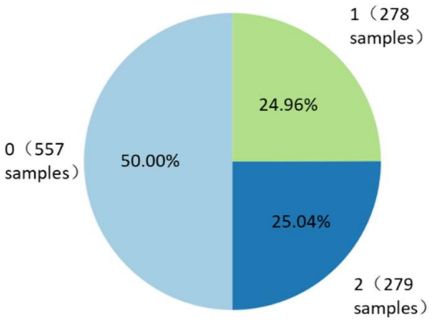

Figure 13 Divisive clustering label count and proportion

The silhouette scores for different numbers of clusters, as displayed in Figure 12, reveal a peak at 3 clusters, with a silhouette score of approximately 0.52. This relatively moderate score indicates that the samples are pretty well-separated, although not as cohesively grouped as in the case of 2 clusters in agglomerative clustering. As the number of clusters increases beyond 3, the silhouette scores steadily decline, indicating diminishing clustering quality. This suggests that while splitting the data into more than 3 clusters captures additional granularity, it does so at the expense of cohesion and separation within the clusters. In Figure 13, the distribution of the selected clusters (3 clusters) is visualized using a pie chart. Cluster 0 represents 50% of the dataset (557

samples), Cluster 1 accounts for 24.96% (278 samples), and Cluster 2 makes up 25.04% (279 samples). The relatively even distribution among clusters reflects a reasonable balance in the way the samples have been partitioned.

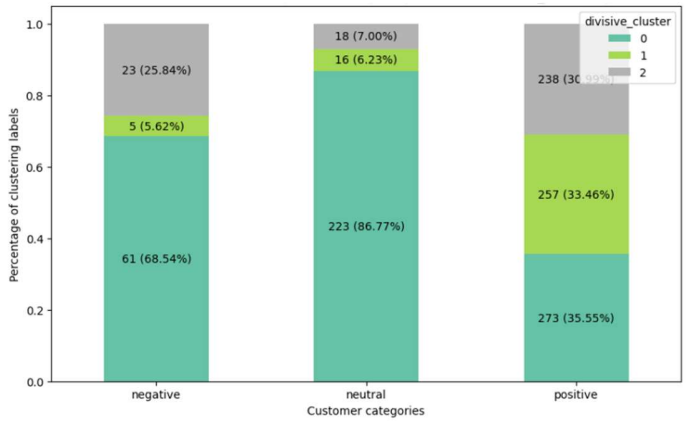

Figure 14 Relationship between agglomerative clustering labels and customer categories

The stacked bar chart in Figure 14 shows the distribution of divisive clustering labels across customer categories. For the "negative" and "neutral" categories, Cluster 2 (green) is dominant, containing 68.54% and 86.77% of the samples, respectively, while Cluster 0 (gray) and Cluster 1 (yellow) account for smaller proportions. In contrast, the "positive" category is more equally distributed across the three clusters, with Cluster 2 containing 35.55%, Cluster 1 having 33.46%, and Cluster 0 representing 30.99%. This suggests that negative and neutral customers cluster together in Cluster 2, while positive customers are more evenly spread across all clusters.

The two clustering approaches, hierarchical agglomerative and divisive clustering, show distinct patterns in how emotion-related features relate to customer categories. The difference between the two methods can be attributed to their clustering strategies. Agglomerative clustering, which merges smaller clusters, emphasizes clearer separations based on customer emotion, particularly for "positive" customers, while divisive clustering, which splits larger clusters, captures more subtle variations within the "positive" group, leading to a more balanced distribution. Despite their differences, both methods show that emotion-related features have a latent relationship with customer categories, indicating their potential use as input for identifying potential purchasers in model training.

3.2. Audio interaction pattern feature selection and data mining

3.2.1. Speaker number analysis

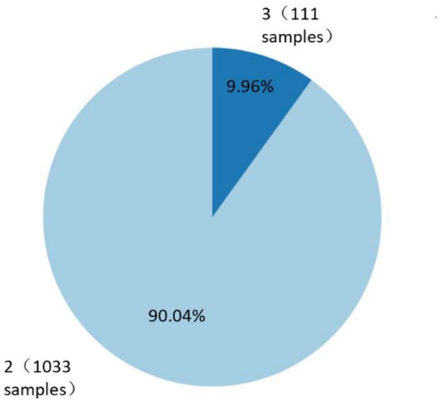

Figure 15 Speaker number distribution

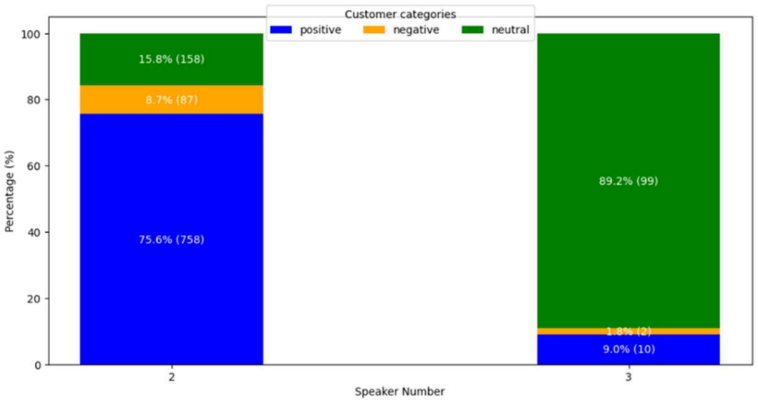

Figure 16 Relationship between speaker number and customer categories

Speaker numbers capture the complexity of interactions, such as whether a conversation involves only one interviewee and one interviewer (one-on-one conversations) or includes multiple customers, which could influence a customer's emotional state and purchasing behavior. For example, one-on-one conversations may foster a more personal, direct interaction, whereas multiple speakers could introduce varied tones, interruptions, or collaborative decisions, affecting the flow of the conversation and the likelihood of conversion. Studies, such as Stolcke et al.'s work on speech act classification, show that features like speaker numbers can improve prediction in decision-making dialogues, making it a relevant feature for customer identification tasks in e-commerce [23].

In the cleaned dataset of the Client Telephone Interview audios provided by VINN Auto, the speaker number distribution, as displayed in Figure 15, shows that the majority of the interviews involve two speakers, which accounts for approximately 90.04% (1003 samples) of the data, while three-speaker interactions make up only 9.96%(111 samples). This suggests that most interactions are between the customer and one representative, making two-person dialogues the dominant form of communication in these interviews.

By analyzing the relationship between speaker numbers and customer categories, the results displayed in Figure 16 show that two-speaker interactions are predominantly associated with positive customer categories, making up 75.6% of these interactions, followed by neutral and negative categories at 12.7% and 8.0%, respectively. This suggests that direct, one-on-one conversations facilitate smoother communication, increasing the likelihood of positive outcomes for potential purchasers. For three-speaker interactions, however, the context becomes more

nuanced. The distribution shifts, with neutral customers representing the largest proportion at 46.7%, followed by positive customers at 41.4%, and negative customers accounting for 11.9%. Closer examination reveals that some three-speaker interactions involve automated responses from mailbox systems, such as the interviewer's introduction, a pre-recorded message from the interviewee indicating unavailability, and the system's automated voice. These interactions, lacking real-time dialogue, likely account for the higher prevalence of neutral customers, as they are less likely to impact purchasing decisions or yield immediate outcomes. On the other hand, in cases involving two interviewee and one interviewer, positive interactions outnumber negative ones, highlighting a greater likelihood of favorable outcomes in these scenarios.

While speaker number can help differentiate personal interactions across three scenarios, it have limitations in identifying potential purchasers on its own. In this context, combining with other features such as content-related attributes alongside speaker numbers, may serve as valuable indicators for identifying potential purchasers, thereby enhancing model performance in customer identification.

3.2.2. Participants' gender analysis

Participants' gender is another important factor, as research has shown that gender dynamics between interviewers and respondents can affect outcomes, such as response retention and survey length. These gender-based differences can play a significant role in how information is processed and decisions are made during customer interactions. For instance, studies like those by Huddy et al. and Himelein demonstrate that gender matching between interviewers and respondents can impact the quality of responses, as people may communicate differently based on gender pairings. This dynamic could potentially be leveraged in predictive models for identifying customer categories, where nuances in gender-based communication styles might correlate with purchasing behaviors [11, 12].

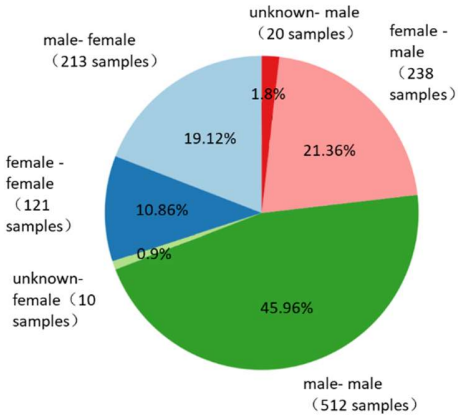

Figure 17 Interviewer-interviewee distribution

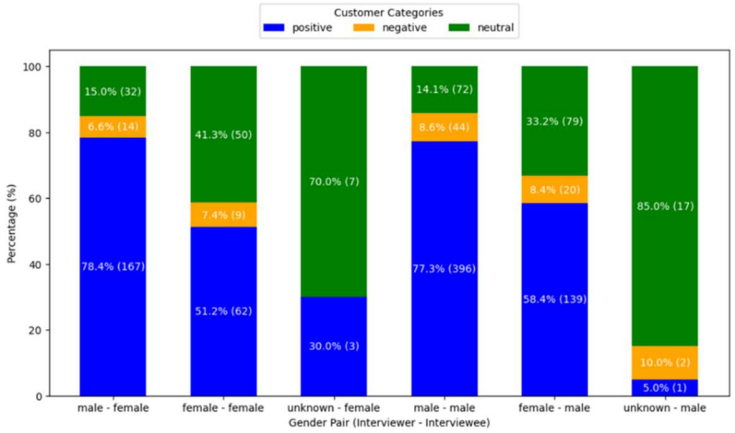

Figure 18 Relationship between participants' gender and customer categories

In the cleaned dataset of the Client Telephone Interview audios provided by VINN Auto, to simplify the analysis, for those cases involving two customers, only the gender of the target interviewee was recorded. This led to the generation of six distinct interaction pairs between interviewers and interviewees. The gender distribution of participants, as shown in Figure 17, reveals that male-male pairs (both interviewer and interviewee being male) constitute the majority, making up nearly 46% of the total interactions. This is followed by female-male pairs (interviewer being female and interviewee being male) and male-female pairs (interviewer being male and interviewee being female), contributing 21.36% and 19.12%, respectively. The remaining interactions, including female-female (interviewer and interviewee being female), unknown-female (unknown interviewer and interviewee being female), and unknown-male (unknown interviewer and interviewee being male) pairs, each account for smaller portions.

When analyzing the relationship between participants' gender and customer categories, the results, as illustrated in Figure 18, reveal several key insights. Male-to-female interactions exhibit a higher proportion of positive customer categories, accounting for nearly 78.4% of the total, while neutral and negative categories are comparatively smaller. Conversely, in female-to-female interactions, the distribution shifts, with a substantial portion of neutral customers (41.3%) compared to male-to-female conversations. This pattern may suggest that gender dynamics between interviewers and interviewees influence the nature of the conversation, potentially leading to differences in customer response. The results also show that unknown-to-female and female-to-male interactions tend to lean towards neutral and negative categories. Interestingly, in male-to-male interactions, the majority of customers fall into the neutral category, indicating that these

conversations may be more formal or structured, which could influence customer engagement. This analysis highlights that gender dynamics in telephone interviews can influence customer classification, offering valuable insights into customer engagement strategies for the company.

Therefore, incorporating participants' gender as a feature in customer classification models could improve the accuracy of identifying potential customers, as it captures nuances in communication dynamics that may affect customer engagement and decision-making. These findings highlight the potential contribution of gender dynamics in enhancing customer identification models in e-commerce and other customer-focused industries.

3.3. Audio content-related feature selection and data mining

3.3.1. Audio transcription

Audio recordings provided by VINN Auto were transcribed using the Clipto platform, after comparing its performance against other popular transcription services such as Google Cloud API, IBM Watson, and Microsoft Azure Speech to Text. Clipto was selected due to its advantages in several key areas critical for handling large-scale audio data efficiently.

Firstly, Clipto demonstrated superior flexibility in processing file length and number limits compared to other platforms, which often impose stricter constraints on file durations or the number of files that can be processed. This made it a more suitable choice for long, continuous samples like those provided by VINN Auto. Additionally, Clipto offered a simpler and more intuitive interface compared with Google Cloud APIs, making the transcription process easier to conduct with less setup complexity.

In terms of transcription speed, Clipto consistently outperformed its competitors by delivering faster results, even when handling long recordings, which is crucial for applications that demand quick turnaround times. Additionally, Clipto's accuracy in transcribing specific terminology and capturing nuanced speech patterns with multiple speakers was found to be comparable to, and in some cases superior to, other platforms like Google Cloud APIs or IBM Watson, which occasionally struggled with domain-specific vocabulary. However, its current service pattern have limitations in real-time applications.

3.3.2. TF-IDF score

Audio content-related features, such as Bag of Words (BoW), Part of Speech (POS) Tags, N-grams (bigrams, trigrams), and Term Frequency-Inverse Document Frequency (TF-IDF) scores, are instrumental in customer classification tasks because they capture the core semantic content conveyed during interviews or conversations. These features allow a precise and direct representation of the spoken content, helping machine learning models interpret critical cues from customer responses.

In this project, TF-IDF is employed as the primary content-based feature for baseline analysis. TF-IDF is widely recognized in text mining and information retrieval for its ability to evaluate the importance of words in a document relative to a collection of documents [24]. When applied to transcriptions of customer audio recordings, TF-IDF excels at identifying key terms and phrases that can indicate potential customer behaviors, interests, or emotional states By emphasizing the most relevant content within a conversation, this method enables models to detect important patterns linked to purchasing decisions or customer satisfaction [1].

Furthermore, the smoothed version of TF-IDF is employed in this project to avoid the issue where terms that appear in all documents would receive an IDF score of 0. Without smoothing, these terms could be incorrectly treated as irrelevant. Smoothing ensures that even common words maintain a meaningful weight while still de-emphasizing them relative to rare words, enhancing the overall model's robustness by retaining potentially important information across documents.

The smooth TF-IDF is calculated as:

$\text{TF-IDF}(t, d_i) = \text{TF}(t, d_i) \times \text{IDF}(t)$, where TF(t, d_i) refers to the count of how often a particular word t appears in the i-th document of a document collection d, and IDF(t) measures how common or rare a word t is across all documents in a corpus.

TF(t, d_i) and *IDF(t)* are computed as below, respectively:

$$\text{TF}(t, d_i) = \frac{\text{Number of occurrences of term } t \text{ in document } d_i}{\text{Total words in document } d_i}$$, where t is the word that are evaluated and d_i is the i-th document of a document collection d.

$$IDF(t) = \log\left(\frac{N+1}{DF(t)+1}\right) + 1$$, where N is the total number of documents and *DF(t)* is the number of documents containing the word t.

Additionally, two key configurations were applied when extracting TF-IDF scores in this project. First, a maximum word limit of 5000 was set by selecting the top 5,000 words with the highest TF-IDF scores to ensure computational efficiency and mitigate overfitting when dealing with high-dimensional text data. Second, English stop words and customer names were excluded. Numbers were also removed, despite containing some potentially valuable information, such as vehicle prices or purchase budgets. However, removing them was necessary as numbers often represent unrelated details like age, dates, or distances, which could introduce noise. These steps ensure that TF-IDF scores reflect only meaningful terms in customer interactions, filtering out words that typically do not offer insights into customer behavior or intent. By removing these distractions, the extracted features more accurately represent the core content, aligning the analysis with the underlying meaning of the text. Several studies have explored similar strategies to enhance text mining and information retrieval. For example, Qiu and Frei [25] highlight the significance of removing stop words and limiting the number of features in query expansion to focus on the most salient aspects of the text, which has been shown to improve classification performance. In another study, Ramos [26] explores how filtering out irrelevant or noisy terms, such as names and numbers, can significantly enhance the accuracy of models in customer-related applications by preventing these non-informative words from skewing the analysis.

3.3.3. Data mining

By computing the average TF-IDF score for each word across different customer categories, the top 30 words with the highest values for each label (positive, negative, and neutral) are selected and visualized in Figures 18, 19, and 20, respectively.

Figure 19 Word cloud for positive

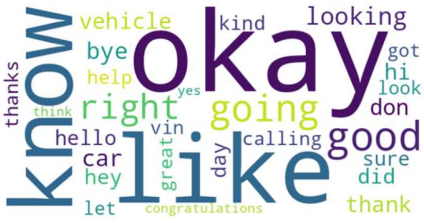

Figure 20 Word cloud for negative

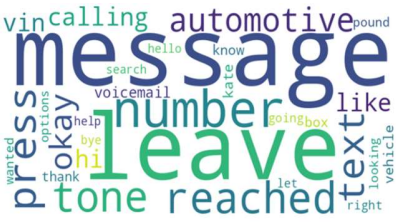

Figure 21 Word cloud for neural

The results underscore distinct patterns in word usage across customer categories. In Figure 19 (positive category), frequent terms like "okay", "like", "know" and "right" suggest a general tendency towards affirmations and agreement in customer responses. Figure 20, which represents the negative category, still highlights similar words with the positive category such as "okay",

"like" and "know," but in a context that potentially reflects disagreement or hesitation. Examining the key words unique to the positive and negative categories, terms like “dealership” and “drive” suggest a stronger inclination toward test drives, classifying them as positive indicators. Conversely, words such as “congratulations” and “thank” imply that the customer has likely already purchased a car and is unlikely to consider buying another in the near future, classifying them as negative indicators. In Figure 21, words such as "message" and "leave" dominate, aligning with neutral or automated system responses, indicating no direct reply from the customers, often associated with voicemail or message systems.

These observations suggest that TF-IDF scores are particularly effective for distinguishing neutral category by capturing key content words that characterize these responses. However, the method appears less powerful in identifying negative and positive categories because of the diversity of personal choices involved. In such cases, context and sentiment nuances may play a more significant role. For this dataset, TF-IDF scores are strong in isolating positive interactions but may require further enhancement or supplementary features, such as sentiment analysis or contextual interpretation, to better capture negative or nuanced responses.

3.4. Summary of feature selection and data mining

Conducting the relevant analysis depicted in this chapter, three kinds of features, including emotion-related features, interaction-pattern features, and content-related features, with a total of twenty-one types, are selected as input for further model training in the next steps.

Emotion-related features, which consist of time-domain features, frequency domain features, and time-frequency representatives, have proved their potential for identifying prospective purchasers by employing hierarchical agglomerative and divisive clustering. Agglomerative clustering emphasizes clear separations, especially for "positive" customers, while divisive clustering captures subtle variations within this group. Both approaches highlight a latent relationship between emotion-related features and customer categories,

Interaction-pattern features have certain limitations, such as being unable to clearly identify different types of customers on their own. However, they reflect the customers’ behavior style and decision-making pattern, which indicates their latent use when collaborating with other features.

Content-related features, specifically the TF-IDF score, can effectively distinguish positive

and neutral customers, while it is weaker at identifying negative categories.

Although these three subsets of features may contribute to customer classification uniquely, their combination might be more powerful in capturing customer nuances across different labels.

Chaper 4. Experiment

4.1. Model selection and training

4.1.1. Model selection

Several classifiers are widely used in customer identification, each with distinct strengths. For example, Chaudhuri et al. investigates customer purchase behavior on e-commerce platforms utilizing Random Forest, selected for its ability to manage large, imbalanced datasets [27]. Lerch aimed to advance audio content analysis in signal processing and music informatics by employing k-Nearest Neighbors (k-NN) due to its non-parametric nature and straightforward implementation, making it suitable for smaller datasets with clear clusters [28]. Selmy et al. aimed to enhance big data analytics in customer identification, by employing Neural Networks (NN), recognized for their capacity to capture complex, high-dimensional relationships within audio features [29].

Two key aspects need to be considered in this project for model selection. First is the nature of the audio dataset provided by VINN Auto, which contains a limited number of samples for model training and testing, along with an imbalanced label distribution.This poses challenges such as overfitting and difficulty in controlling the false positive rate (FPR). The second aspect is the limitation of computational resources, which constrains the choice of models that can be effectively implemented.

Given these considerations, SVM, Decision Tree, and XGBoost are selected for their suitability in handling small datasets and their flexibility in adjusting class weights to address imbalance. Additionally, these models have a manageable computational cost, making them feasible within the project's resource constraints. Moreover, an ensemble mothed is developed with the combination of the three models, enhancing the robustness of the final result.

1) SVM

SVMs are highly effective for handling small datasets with a large number of features, making them ideal for high-dimensional data. They excel at finding the optimal decision boundary between classes, even when the data is imbalanced, by adjusting class weights or implementing cost sensitive learning. This ensures that minority and majority classes are appropriately weighted, enhancing model performance in situations where certain categories are prioritized.

The objective function for SVMs involves maximizing the margin between classes while minimizing misclassification errors. For a linear SVM which is chosen for this project, this can be expressed as a convex optimization problem:

$$\min_{\mathbf{w},b,\xi}\left(\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i\right)$$

, with constraints $y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i$ and $\xi_i \geqslant 0$, where w is the weight vector (parameters of the hyperplane), b is the bias term (offset of the hyperplane), $\xi_i \geq 0$ are slack variables that allow for misclassification, C is a regularization parameter controlling the trade-off between maximizing the margin and minimizing the classification error (misclassification). $y_i \in \{-1,+1\}$ is the true class label for sample i, x_i is the feature vector for sample i.

2) Decision Tree

Decision Trees are advantageous for their simplicity and interpretability. They handle high-dimensional data by selecting the most relevant features through a process of recursive splitting, reducing complexity. Additionally, Decision Trees can be adjusted to account for class imbalances with the same method as SVM, ensuring that the model gives more attention to certain categories, which is crucial for capturing nuanced patterns in the data.

Decision Trees optimize a splitting criterion, typically either Gini Impurity or Entropy for classification tasks. The goal is to split the data such that each node becomes as pure as possible, meaning that each leaf node should contain instances of mostly one class.

Gini Impurity is computed as:

$$Gini(t) = 1 - \sum_{i=1}^{k} p_i^2$$

, where p_i is the probability of class i at node t, and k is the number of classes.

Entropy (used in information gain) is computed as:

$$Entropy(t) = -\sum_{i=1}^{k} p_i \log(p_i)$$

, where p_i is the proportion of class i instances at node t.

In this project, gird search and cross validation are conducted to tune the model, finding which of the two methods along with other hyperparameters could result in a better performance.

3) XGBoost

XGBoost is a powerful algorithm designed for efficiency and scalability. It performs well with high-dimensional data, using regularization techniques to prevent overfitting. XGBoost's ability to handle imbalanced datasets through class weighting and its strength in modeling complex interactions between features make it particularly effective for capturing detailed patterns in the data, even when certain classes require more focus than others.

XGBoost uses an objective function based on gradient boosting, where it minimizes the following regularized loss function:

$$\text{Objective} = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{t=1}^{k} \Omega(f_t)$$

, where n is the number of training examples, $L(y_i, \widehat{yi})$ is the loss function measuring prediction error for each data point, k is the number of trees, $\Omega(f_t)$ is the regularization term for each tree.

The loss function for multi-class classification is the negative log likelihood of the true class labels, which is formulated as:

$$L(y, \hat{y}) = -\sum_{i=1}^{n} \sum_{k=1}^{K} y_{ik} \log(\hat{p}_{ik})$$

, where y_{ik} is a binary indicator (0 or 1) that indicates whether class k is the true label for observation i, $\hat{p}_{ik}$ is the predicted probability for class k on observation i, and K is the number of classes, and n is the total number of samples.

The regularization term is defined as:

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

, where T is the number of leaves in the decision tree, w_j is the weight associated with leaf j, γ is the regularization parameter that controls the penalty on the number of leaves T, λ is the L2 regularization parameter, controlling the magnitude of the weights w_j.

4) Bagging ensemble

Using a Bagging Ensemble offers several advantages in this project. By resampling data through bootstrapping and combining predictions from multiple models, Bagging helps reduces variance and makes predictions more stable. It also mitigates overfitting by averaging predictions across models, making it more robust and generalizable, especially with the complex and high-dimensional audio features used. Meanwhile, Bagging is effective in capturing non-linear relationships, making it well-suited for the multiclass classification task in this project.

Furthermore, a soft voting mechanism is employed in this project, where multiple classifiers contribute to the final prediction based on the predicted class probabilities rather than just the class labels, as in hard voting. The final predicted class is the one with the highest averaged probability as below:

$$\hat{p}_c = \frac{1}{M} \sum_{i=1}^{M} p_{i,c}$$

, where $\hat{p}_c$ is the final predicted probability for class c, M is the total number of classifiers, $p_{i,c}$ is the predicted probability of class c by the i-th classifier.

This method proves advantageous because it considers the confidence level of each classifier's prediction, possibly leading to more stable results. Unlike hard voting, where each classifier's vote is weighted equally, soft voting leverages the varying degrees of certainty from each model, enhancing the overall performance.

4.1.2. Model training

1) Hyperparameter tuning and cross-validation

In this project, both Decision Tree and XGBoost classifiers were tuned using grid search and a 5-fold cross-validation to optimize their hyperparameters for better performance.

For Decision Tree, the grid setup included parameters such as max_depth with values of [None, 15, 25], min_samples_split with values of [2, 10], min_samples_leaf with values of [1, 5], and the criterion with options ['gini', 'entropy'], as shown in Table 7. The range for each hyperparameter was established based on common values used in practice, empirical testing, and the specific characteristics of the dataset. After tuning with different inputs, the chosen hyperparameters were found to be max_depth=15, min_samples_split=10, and

min_samples_leaf=1 for the content-related input, and min_samples_split=2, min_samples_leaf=5 for the cases with the combination of content-related features, emotion-related feature and interaction-pattern features as input, both using the 'gini' criterion, as summarized in Table 8.

Decision tree	
hyperparameters	values
max_depth	None, 15, 25
min_samples_split	2, 10
min_samples_leaf	1, 5
criterion	gini, entropy

Table 8 Grids setup for Decision Tree

Decision tree		
	content	content+audio
max_depth	15	15
min_samples_split	10	2
min_samples_leaf	1	5
criterion	gini	gini

Table 9 Chosen hyperparameters for Decision Tree

For XGBoost, the grid setup included parameters such as n_estimators with values [50, 100, 200], max_depth with values [3, 8], learning_rate with values [0.01, 0.1, 0.2], subsample with values [0.5, 0.85], and colsample_bytree with values [0.5, 0.85], as shown in Table 9. After tuning, the best hyperparameters for the content-related input were n_estimators=50, max_depth=8, learning_rate=0.2, subsample=0.85, and colsample_bytree=0.5. For using the combination of content-related features, emotion-related feature and interaction-pattern features as input, the chosen hyperparameters were n_estimators=200, max_depth=8, learning_rate=0.01, subsample=0.85, and colsample_bytree=0.85, as summarized in Table 10.

XGBoost	
hyperparameters	values
n_estimators	50, 100, 200
max_depth	3, 8
learning_rate	0.01, 0.1, 0.2
subsample	0.5, 0.85
colsample_bytree	0.5, 0.85

Table 10 Grids setup for XGBoost

XGBoost		
	content	content+audio
n_estimators	50	200
max_depth	8	8
learning_rate	0.2	0.01
subsample	0.85	0.85
colsample_bytree	0.5	0.85

Table 11 Chosen hyperparameters for XGBoost

2) Weight configuration and cross-validation

In the context of e-commerce platforms, especially for companies like VINN Auto, identifying potential purchasers is critical to driving business success. The positive class, which represents potential buyers, holds the highest value for sales teams because these are the leads most likely to convert into actual sales. Given the nature of the business, where follow-up actions can be resource-intensive, it is crucial to correctly identify and prioritize potential purchasers for outreach. Missing a potential purchaser (false negative) directly translates to missed sales opportunities, which can significantly affect revenue generation. For such reason, while accuracy is important, ensuring that the model captures as many potential purchasers as possible—without misclassifying them—is even more crucial.

To address this, class weight configuration plays a key role in this project. By assigning lower weights to the neutral and negative classes, we can mitigate the model's tendency to favor the non-targeted class and reduce false positives. However, the model must still remain sensitive to correctly identifying positive cases. This balance is achieved by carefully tuning the class weights, ensuring that potential purchasers are detected accurately while maintaining a low false negative rate. Through this approach, the model is optimized not just for overall performance but for business-critical outcomes, ensuring that high-potential leads are captured effectively and sales efforts are maximized.

During the training process, a modified class weight configuration of [10:2:1] (positive: neutral: negative), which assigned a higher weight to the positive class, was applied across all classifiers, except for the ensemble method, to establish a comparison group against the default class weights. This setup aimed to assess the impact of these weight adjustments on model performance. Further testing with alternative class weight configurations, including [11:3:1], [12:3:1], [15:1:1], and [10:1:1], showed no significant changes in the results across all three classifiers. Therefore, the predefined class weights of [10:2:1] were used in the ensemble method, rather than applying grid search to tune class weights for the voting classifiers.

4.2. Result, evaluation, and comparison

This project employs several evaluation metrics, including accuracy, positive recall, and positive FNR (False Negative Rate), to provide a comprehensive evaluation of model performance. Accuracy measures the overall correctness of the predictions, giving a broad indication of the model's general effectiveness. However, in customer identification tasks, it is critical to ensure that potential purchasers are not misclassified as uninterested. This is where positive recall become essential. Positive accuracy focuses specifically on the model's correctness when identifying potential customers, while positive recall ensures the model's ability to correctly detect all potential purchasers. Additionally, introducing positive FNR helps highlight instances where actual potential customers are incorrectly classified as non-interested. Balancing these metrics is crucial for ensuring that the model captures as many valuable leads as possible without missing opportunities, thereby maximizing customer conversion rates for e-commerce platforms. The F2 score is not used because it emphasizes a balance between precision and recall, focusing on recall. However, in customer identification, our priority is to maximize positive recall and minimize

positive FNR to ensure all potential purchasers are identified without missing valuable leads. Using positive recall and FNR directly provides clearer, more targeted insights for this purpose.

		Unweighted(77:23:8)			Weighted(10:2:1)		
Input features	Estimater	svm	decision tree	xgboost	svm	decision tree	xgboost
content	accuracy	79.82%	73.54%	84.30%	81.61%	76.23%	82.96%
	positive recall	92%	84%	99%	99%	93%	99%
	positive FNR	8%	16%	1%	1%	7%	1%
Audio + content	accuracy	82.06%	82.06%	87.89%	82.06%	80.27%	86.55%
	positive recall	89%	88%	99%	97%	98%	99%
	positive FNR	11%	3%	1%	3%	2%	1%

Table 12 Classification results for SVM, Decision Tree and XGBoost

The results across different input features and classifiers with varying weight configurations are shown in Table 11.

In the unweighted configuration, performance improved when both types of features (audio and content) were included, leading to an increase in accuracy, despite minor changes in positive recall and FNR.

- Specifically, for SVM, accuracy improved from 79.82% to 82.06% when audio features were included, although there was a slight trade-off: positive recall dropped from 92% to 89%, and the FNR increased from 8% to 11%. Despite this, the inclusion of audio features still provided overall benefits in performance, particularly in enhancing accuracy and distinguishing other two customer categories.

- Meanwhile, Decision Tree showed a substantial boost in performance when both audio and content features were included. The overall accuracy increased to 82.06%, and positive recall rose to 88%, compared to 73.54% and 84%, respectively, when using only content features. Notably, the False Negative Rate (FNR) dropped from 16% to 3%. The substantial improvement suggests that audio features help the model capture more nuanced patterns in customer interactions, leading to fewer misclassifications, especially in the positive category.

- Similarly, XGBoost demonstrated improved accuracy, rising from 84.30% with content-only data to 87.89% when audio features were added. Furthermore, the model maintained

nearly perfect scores for positive recall (99%) and FNR (1%), showing its strong ability to classify positive cases correctly. This shows XGBoost's superior stability and effectiveness across different feature sets.

In the weighted configuration, the overall performance further improved when using both audio and content-related features compared to content features alone. Positive recall and FNR either showed improvements or remained consistent compared to the unweighted configuration across all classifiers. However, the accuracy metrics presented varied results depending on the classifier.

- In the case of SVM, accuracy increases from 79.82% (unweighted, content-only) to 81.61% (weighted, content-only) and remains unchanged at 82.06% when both features are used in both configurations. However, there is a notable improvement in positive recall when both features are used, rising from 89% (unweighted) to 97% (weighted), while the FNR decreases from 11% to 3%. This indicates that the SVM model benefits more from the inclusion of audio features in the weighted configuration, especially in improving recall performance, although the impact on accuracy remains minimal.

- When using Decision Tree as a classifier, the model exhibits a slight increase in accuracy when using content-related features alone, improving from 73.54% in the unweighted configuration to 76.23% in the weighted configuration. However, when both audio and content features are incorporated, the accuracy decreases slightly from 82.06% (unweighted) to 80.27% (weighted). Positive recall improves significantly in both cases, rising from 84% to 93% (content-only) and from 88% to 98% (both features), while the FNR shows improvement, dropping from 16% to 7% (content-only) and from 3% to 2% (both features). This suggests that while overall accuracy slightly drops when using both features in the weighted configuration, the model's ability to correctly identify positive cases improves, especially when both features are included.

- For XGBoost, the weighted configuration results in a minor accuracy drop when using only content-related features, from 84.30% in the unweighted configuration to 82.96%. However, accuracy remains consistent when both features are used, with a marginal change from 87.89% (unweighted) to 86.55% (weighted). Positive recall remains perfect at 99%, and FNR stays consistently low at 1% in all configurations, indicating that

XGBoost is highly stable and performs well regardless of the feature set or weighting applied.

For the ensemble method using soft voting with a predefined class weight of [10:2:1] (positive: negative: neutral) and both audio and content features, the results in Table 9 show an accuracy of 83.41%, a positive recall of 98%, and a positive FNR of 1.95%. The bagging ensemble outperforms SVM (82.06%) and Decision Tree (80.27%) in accuracy by 1.35% and 3.14%, respectively, but lags behind XGBoost (86.55%) by 3.14%. In terms of positive recall, the ensemble is on par with Decision Tree (98%) and slightly behind XGBoost (99%). For positive FNR, it performs similarly to XGBoost (1%) and better than SVM (3%) and Decision Tree (2%).

Estimater	Value
accuracy	83.41%
positive recall	98%
positive FNR	1.95%

Table 13 Classification result for Bagging ensemble

The ensemble method performs better than SVM and Decision Tree because bagging reduces overfitting and combines the strengths of multiple models. However, it falls short of XGBoost, which excels in capturing complex feature interactions through gradient boosting. While the ensemble improves accuracy and FNR over SVM and Decision Tree, it lacks XGBoost's ability to fine-tune predictions, resulting in slightly lower performance in positive recall and overall accuracy.

Overall, the results indicate that incorporating audio features, particularly in the weighted configuration, consistently enhances recall and reduces the FNR, although its effect on accuracy varies across classifiers. This finding reinforces the notion that audio features, especially those related to emotion and interaction patterns, significantly contribute to improving the predictive power of customer classification models. On the other hand, the use of weight configuration appears to be an effective strategy for further boosting recall and reducing FNR, which is particularly important for identifying potential targets in this context. Moreover, among the classifiers, XGBoost stands out as the most reliable and consistent, delivering excellent performance regardless of the input features or weighting configurations used.

Meanwhile, this research uses a different set of features for classification than previous studies, making direct comparisons challenging. However, similar studies show that models like XGBoost and ensemble methods perform well in various classification contexts, especially with

customer-related tasks. For instance, Gan [30] and Wang et al [31] demonstrate XGBoost's effectiveness in e-commerce segmentation and purchase behavior prediction, compared with Random forest and Logistic Regression respectively. Additionally, Alojail and Bhatia [32] highlight the value of ensemble methods in customer behavior prediction, specifically through combining bagging techniques, showcasing its flexibility and robustness in capturing a diverse range of customer behaviors.

4.3. Discussion

There are several important considerations when evaluating the broader applicability and limitations of this approach, particularly concerning its generalization across different datasets and the computational cost of data processing, and the trade-offs involved in optimizing model complexity and performance.

- Generalization and Domain Adaptability

While the method achieved strong results recursively on the TESS, SAVEE, and VINN Automotive datasets, the question of generalization to other datasets or industries is crucial. The TESS and SAVEE datasets are controlled and emotion-labeled, which might not fully represent the noisy, real-world audio from actual customer interactions. Different industries may also exhibit distinct interaction patterns, which could limit the direct application of the model to new domains without additional tuning. Future work should explore domain adaptation techniques such as transfer learning to fine-tune models for different datasets and customer behavior contexts. This would enhance the model's robustness and ensure better generalization across various e-commerce sectors.

- Computational Cost and Scalability Challenges

The computational demands of this approach are another significant consideration, especially when processing large volumes of audio data and exploring different type of model. On the one hand, extracting detailed audio features, such as emotional cues and interaction patterns, is both time-consuming and resource-intensive. This high computational cost may limit the approach's scalability, particularly for small and micro businesses in e-commerce that may have limited resources for advanced data processing. On the other hand, exploring various machine learning models, especially those with high complexity such as ensemble or deep learning models, can be

resource-intensive. Training and fine-tuning these models on large datasets require substantial computational resources, which may not be feasible for businesses with budget constraints. Furthermore, complex models tend to have longer training times, which may hinder experimentation and iterative improvements, limiting the agility of the development process.

Chaper 5. Conclusion and future work

5.1. Conclusion

This project demonstrates the significant value of incorporating audio features, particularly those related to emotions and interaction patterns, in improving customer identification models for e-commerce platforms. By analyzing two prominent datasets—TESS and SAVEE—along with proprietary audio data from VINN Auto, the study showcases the effectiveness of leveraging audio features in combination with traditional content-based features to enhance model performance. Through the application of supervised learning techniques such as Support Vector Machines (SVMs), Decision Trees, XGBoost and Bagging ensemble, this approach provides a robust strategy for enhancing customer classification tasks.

The use of a weighted configuration further enhanced positive recall and reduced FNR across all classifiers, confirming that weighting is a viable method for boosting model sensitivity to potential purchasers.

Among the models, XGBoost consistently emerged as the most reliable and stable performer, delivering robust results regardless of feature set or weighting configuration. This consistency, along with its superior recall and low FNR, positions XGBoost as the most effective model for customer identification tasks.

In conclusion, the findings underscore the importance of integrating emotion-related and interaction pattern features into customer classification models to improve predictive performance. By enhancing the ability to identify potential customers, this approach offers e-commerce platforms a powerful tool for optimizing marketing strategies, increasing purchase rates, and minimizing missed opportunities, thereby contributing to more efficient and targeted customer engagement efforts.

5.2. Future work

Future efforts could focus on optimizing the flow of the project including various aspects such as audio cleaning, feature selection, feature extraction, model selection, training optimization to further enhance model performance.

- Audio preprocess and cleaning

In this project, the limited number of audio samples presents a significant constraint for the subsequent steps, similar to the challenges faced with the SAVEE dataset. It is essential to collect more qualified data or perform data fusion by integrating other similar datasets for further analysis.

Additionally, due to the current filtering criteria, samples with three speakers that include automated responses from a mailbox without direct interaction were not excluded, as identifying them requires considerable manual inspection. As a result, they have been left for further analysis. While potential targets may be present in these samples, their inclusion may introduce noise, particularly within the neutral customer category. Future work could explore new criteria to better handle this scenario.

- Feature selection

To further enhance audio feature selection, incorporating interaction-pattern features could be a promising direction. Elements such as turn-taking, pauses, and speaker dominance may improve customer intent prediction. For instance, frequent exchanges between speakers and shorter pauses may indicate higher engagement, while longer speech durations by the customer could suggest stronger purchasing interest. These features, which capture the dynamic flow of conversations, provide deeper insights beyond traditional emotion detection methods. By reflecting the nuances of customer interactions more effectively, this approach may enhance the model's accuracy in identifying potential buyers, leading to more targeted and informed predictions.

- Feature extraction

An effective next step for improving the model's performance in terms of feature extraction could involve separating the interviewer's and interviewee's audio during the preprocessing stage, prior to feature extraction. In many customer interaction recordings, the interviewer's speech may introduce noise or irrelevant information that could affect the accuracy of the emotion detection and interaction pattern analysis. By isolating the interviewee's audio, the model can focus on extracting relevant features such as tone, pitch, and emotional cues directly related to the customer, which may lead to better identification of potential purchasers. Advanced speech separation techniques like speaker diarization or Voice Activity Detection (VAD) can be applied to accomplish this. This refined feature extraction may also reveal subtle patterns that were

previously masked by the overlapping dialogues.

- Model selection

Although the method employed in this project proved effective, the time required for audio feature extraction and the complexity of feature selection may present challenges in other applications, especially those with long and complex audio recordings. To address these issues, a promising direction for future exploration is automating feature extraction through deep learning techniques, particularly those that capture temporal and sequential patterns. Customer interactions often involve dynamic changes in tone, pauses, and conversational flow, which can provide valuable insights into intent. Models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are well-suited to capturing these temporal relationships. By analyzing the evolution of emotion and interaction patterns throughout the conversation, these models may enhance the prediction of purchasing intent, leading to more accurate customer identification.

- Training optimization

Instead of using a class weight configuration, which was employed in this study to balance class imbalance, future work could explore cost-sensitive learning methods to optimize the training process. This approach assigns higher penalties for misclassifying critical classes, such as potential purchasers, thus better addressing the imbalance problem. Cost-sensitive learning integrates the misclassification cost into the model's learning process, ensuring that the model focuses on minimizing high-cost errors (e.g., false negatives for the positive class). This can be more flexible and precise than static class weights, leading to a more tailored optimization of recall and FNR across all classifiers.

Reference

[1] U. Rahardja, T. Hariguna, and W. M. Baihaqi, "Opinion Mining on E-Commerce Data Using Sentiment Analysis and K-Medoid Clustering," in 2019 Twelfth International Conference on Ubi-Media Computing (Ubi-Media), Bali, Indonesia, 2019, pp. 168–170.

[2] Y. Xiong, N. Wei, K. Qiao, Z. Li, and Z. Li, "Exploring Consumption Intent in Live E-Commerce Barrage: A Text Feature-Based Approach Using BERT-BiLSTM Model," IEEE Access, vol. 12, pp. 69288–69298, 2024.

[3] M. Choudhary and P. K. Choudhary, "Sentiment Analysis of Text Reviewing Algorithm using Data Mining," in 2018 International Conference on Smart Systems and Inventive Technology (ICSSIT), Tirunelveli, India, 2018, pp. 532–538.

[4] H. Srivastava, S. Sunil, K. Shantha Kumari, and P. Kanmani, "Multi-modal Sentiment Analysis Using Text and Audio for Customer Support Centers," in Proceedings of ICACTCE'23 — The International Conference on Advances in Communication Technology and Computer Engineering, C. Iwendi, Z. Boulouard, and N. Kryvinska, Eds., Cham: Springer, 2023, vol. 735, Lecture Notes in Networks and Systems.

[5] W. Xu, X. Zhang, R. Chen, and Z. Yang, "How do you say it matters? A multimodal analytics framework for product return prediction in live streaming e-commerce," Decision Support Systems, vol. 172, p. 113984, 2023.

[6] L. F. Parra-Gallego and J. R. Orozco-Arroyave, "Classification of emotions and evaluation of customer satisfaction from speech in real world acoustic environments," Digital Signal Processing, vol. 120, p. 103286, 2022.

[7] Z. K. Abdul and A. K. Al-Talabani, “Mel Frequency Cepstral Coefficient and its Applications: A Review,” IEEE Access, vol. 10, pp. 122136–122158, 2022.

[8] D. Ververidis and C. Kotropoulos, “Emotional Speech Classification Using Gaussian Mixture Models and the Sequential Floating Forward Selection Algorithm,” in 2005 IEEE International Conference on Multimedia and Expo, IEEE, 2005, pp. 1500–1503.

[9] C. H. Wu, J. F. Yeh, and Z. J. Chuang, "Emotion perception and recognition from speech," in Affective Information Processing, 2009, pp. 93–110.

[10] P. Gangamohan, V. K. Mittal, and B. Yegnanarayana, "Relative importance of different components

of speech contributing to perception of emotion," in Speech Prosody 2012, 2012.

[11] K. Himelein, "Interviewer Effects in Subjective Survey Questions: Evidence From Timor-Leste," International Journal of Public Opinion Research, vol. 28, no. 4, pp. 511–533, 2016.

[12] L. Huddy, J. Billig, L. Bracciodieta, L. Hoeffler, P. J. Moynihan, and P. Pugliani, "The effect of interviewer gender on the survey response," Political Behavior, vol. 19, no. 3, pp. 197–220, 1997.

[13] J. Cho, R. Pappagari, P. Kulkarni, J. Villalba, Y. Carmiel, and N. Dehak, "Deep neural networks for emotion recognition combining audio and transcripts," arXiv preprint arXiv:1911.00432, 2019.

[14] O. Kenai, S. Ouamour, M. Guerti, and N. Asbai, "A new architecture based VAD for speaker diarization/detection systems," International Journal of Speech Technology, vol. 22, pp. 827–840, 2019.

[15] J. Foote, “Automatic audio segmentation using a measure of audio novelty,” in 2000 IEEE International Conference on Multimedia and Expo, Proceedings Vols I-III, New York: IEEE, 2000, pp. 452–455, vol. 1.

[16] Y. Fujita, N. Kanda, S. Horiguchi, Y. Xue, K. Nagamatsu, and S. Watanabe, "End-to-end neural speaker diarization with self-attention," in 2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pp. 296–303, Dec. 2019.

[17] D. O'Shaughnessy, Speech Communications: Human and Machine, New York: IEEE Press, 2000, pp. 367–433.

[18] L. R. Rabiner, Digital Processing of Speech Signals, Pearson Education India, 1978.

[19] B. Gold, N. Morgan, and D. Ellis, Speech and Audio Signal Processing: Processing and Perception of Speech and Music, 2nd ed., Wiley, 2011.

[20] F. Eyben, Real-Time Speech and Music Classification by Large Audio Feature Space Extraction, 1st ed., Cham: Springer International Publishing AG, 2016.

[21] G. Peeters, "A Large Set of Audio Features for Sound Description (Similarity and Classification) in the Cuidado Project," Ircam Technical Report, 2004.

[22] K. R. Scherer, "Vocal Affect Expression: A Review and a Model for Future Research," Psychological Bulletin, vol. 99, no. 2, pp. 143–165, 1986.

[23] A. Stolcke, K. Ries, N. Coccaro, E. Shriberg, R. Bates, D. Jurafsky, P. Taylor, R. Martin, C. Ess-

Dykema, and M. Meteer, "Dialogue act modeling for automatic tagging and recognition of conversational speech," Computational Linguistics, vol. 26, no. 3, pp. 339–373, 2000.

[24] C. D. Manning, P. Raghavan, and H. Schütze, Introduction to Information Retrieval, Cambridge: Cambridge University Press, 2008.

[25] Y. Qiu and H. P. Frei, "Concept-based query expansion," in Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '93), 1993, pp. 160–169.

[26] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries," in Proceedings of the First Instructional Conference on Machine Learning, 2003, pp. 133–142.

[27] N. Chaudhuri, G. Gupta, V. Vamsi, and I. Bose, "On the platform but will they buy? Predicting customers' purchase behavior using deep learning," Decision Support Systems, vol. 149, p. 113622, 2021. Available: https://doi.org/10.1016/j.dss.2021.

[28] A. Lerch, An Introduction to Audio Content Analysis: Applications in Signal Processing and Music Informatics, Wiley, 2012.

[29] H. A. Selmy, H. K. Mohamed, and W. Medhat, "Big data analytics deep learning techniques and applications: A survey," Information Systems (Oxford), vol. 120, p. 102318, 2023. Available: https://doi.org/10.1016/j.is.2023.

[30] L. Gan, "XGBoost-Based E-Commerce Customer Loss Prediction," Computational Intelligence and Neuroscience, vol. 2022, no. 1, p. 1858300, 2022.

[31] W. Wang, W. Xiong, J. Wang, L. Tao, S. Li, Y. Yi, X. Zou, and C. Li, "A User Purchase Behavior Prediction Method Based on XGBoost," Electronics, vol. 12, no. 9, p. 2047, 2023. Available: https://doi.org/10.3390/electronics12092047.

[32] M. Alojail and S. Bhatia, "A novel technique for behavioral analytics using ensemble learning algorithms in E-commerce," IEEE Access, vol. 8, pp. 150072–150080, 2020.