AN ENHANCED MULTIMODAL NEGATIVE FEEDBACK DETECTION FRAMEWORK WITH TARGET RETRIEVAL IN THAI SPOKEN AUDIO

Pantid Chantangphol[1]*,Sattaya Singkul*[1]*, Thanawat Lodkaew*[1]*,*
Nattasit Maharattanamalai[1], Atthakorn Petchsod[1], Theerat Sakdejayont[1],Tawunrat Chalothorn[1]

[1]Kasikorn Labs Kasikorn Business–Technology Group, Thailand

ABSTRACT

This research addresses the challenge of effectively identifying negative feedback in spoken audio within the context of voluminous and complex user-generated content. The study introduces an integrated audio analytics framework designed to enhance processing speed and accuracy. The framework combines Query-by-Example Spoken Term Detection (QbE-STD), Speaker Diarization (SD), and Automatic Speech Recognition (ASR) with text-based feedback (sentiment, toxicity and sarcasm detection). By employing QbE-STD, the system facilitates targeted retrieval of specific terms, thus optimizing processing duration. Additionally, the application of transfer learning techniques to under-resourced languages, such as Thai, demonstrates significant improvements in the accuracy of both ASR and text-based feedback analysis. This research paves the way for future studies in large-scale analysis of audio-based negative feedback. It also highlights the potential for deploying efficient audio analytics in various fields, including content moderation and decision support systems.

Index Terms— Multimodal Negative Feedback Detection, Target retrieval, Transfer learning, Audio analysis, Thai

1. INTRODUCTION

In the digital age, the vast array of user-generated content, including videos, comments, and reviews, serves as a critical resource for understanding public sentiment. Such content is invaluable for decision-making in business, government, and individual contexts. However, a primary challenge lies in efficiently processing and analyzing this extensive data within constrained time limits, considering its volume and diversity. Particularly challenging is the interpretation of negative feedback, a task complicated by these constraints.

In various subfields related to audio and text processing for negative feedback retrieval, including QbE-STD, SD, ASR, and Natural Language Processing (NLP) techniques for sentiment analysis, toxicity, and sarcasm analysis. We note a research gap in under-resource languages such as Thai in QbE-STD, despite advancements in deep learning frameworks for feature extraction [1]. In SD, recent methods utilize deep neural networks [2] for managing overlapped speech, but often neglect integration with other components for comprehensive emotion analysis. ASR advancements, such as the Whisper model [3], require extensive training data, which we address through transfer learning for Thai. Sentiment analysis models, such as BERT [4] and WangchanBERTa [5], have shown promise, as have techniques in toxic content detection and sarcasm identification, where transfer learning [6, 7] is commonly applied. However, emotion analysis in spoken content, especially in Thai [8], remains underexplored, with limited integrated use of SD, ASR, and text-based sentiment analysis. Our research aims to bridge these gaps by creating a framework for efficient and precise emotion analysis in Thai.

To address the challenge of identifying negative feedback in voluminous user-generated spoken content, we propose an integrated audio analytics approach. This framework combines QbE-STD, SD, and ASR to efficiently pinpoint negative comments in videos, reducing analysis time. QbE-STD aids in efficient retrieval and analysis of relevant spoken content, a necessity in lengthy and diverse audio sources.

Our research focuses on developing a framework for assessing negative feedback in under-resourced languages like Thai, using transfer learning techniques. We aim to demonstrate the functionality of our framework and its applications in decision-making processes. Thus, our contributions are:

1. We propose an Integrated Audio Analytics Framework designed to identify negative feedback more efficiently and quickly. This framework combines QbE-STD, SD, and ASR to enhance performance. It is capable of extracting and analyzing speech information from videos, facilitating precise retrieval of spoken content, converting dialogue into text, and evaluating the presence of toxicity, sentiment and sarcasm.

2. We present employing transfer learning techniques to enhance the framework's efficiency in resource-scarce languages, particularly Thai. This method overcomes the issue of limited language-specific data, significantly improving the accuracy of negative feedback analysis and individual performance in Thai spoken content.

The structure of the paper is as follows: Section 2 details our proposed system. Implementation aspects are discussed in

Section 3, and results are presented in Section 4. The discussion and future work are covered in Section 5, and the paper concludes with Section 6.

2. METHODOLOGY

The proposed method efficiently handles negative feedback and comments associated with specific terms in spoken audio. It achieves this by integrating QbE-STD, SD, and ASR with text-based analyses of sentiment, toxicity, and sarcasm. Our research emphasizes rapid data processing and analysis of audio content, particularly focusing on terms of interest. This approach not only saves time but also reduces the computational resources required. To outline the proposed framework, the process is initiated by extracting terms from the audio using QbE-STD. Subsequently, the audio is segmented based on speaker identity through SD, and each utterance is transcribed using ASR. These transcriptions are then analyzed for sentiment, sarcasm, and toxicity, as illustrated in Figure 1.

2.1. Query-by-Example Spoken Term Detection

QbE-STD is focused on identifying specific terms within audio data. In our study, we use a pre-trained Acoustic Word Embeddings (AWEs) network [9] to encode a Thai corpus. To improve word discrimination, we implement Deep word discrimination loss (DWD loss) [9]. Additionally, we utilize Word embedding basis [9], a sliding window technique for acoustic word matching in QbE-STD tasks. During fine-tuning, acoustic embeddings are generated following specific guidelines [9]. To evaluate the effectiveness of these AWEs, we use the same-different word discrimination task,as described by [9]. We measure the cosine similarity between each pair of embeddings, comparing it against a threshold of 0.5 to determine if they represent the same word.

2.2. Speaker Diarization

SD is a crucial component in our framework for identifying 'who spoke when'. The outputs of SD, which are speech segments with corresponding speaker identities, are fed into an ASR model for transcription. This process amalgamates into a text-based conversational dialogue extracted from audio recordings. Developing an SD model involves addressing three distinct challenges: voice activity detection, speaker change detection, and overlapped speech detection, which complicate efforts to enhance model performance. To address these challenges, we adopt an end-to-end neural speaker diarization approach [10], enabling our SD model to incorporate these sub-tasks. This approach is particularly effective in handling overlapped speech regions with resegment to consider only non-overlap speech regions, distinguishing it from other methods. Its efficacy is especially notable in analyzing Thai YouTube videos, where overlapped speech is common.

2.3. Automatic Speech Recognition

ASR is designed to transcribe spoken words into text. In our study, we enhance ASR for resource utilization, which is adapted after QbE-STD and SD are processed. This integration is aimed at identifying and focusing on non-overlapping speaker utterances. By doing so, we aim to optimize the efficiency. Besides, to handle under-resource, we employ a transfer learning method using pre-trained Whisper models [3, 11] for Thai language ASR. These models are initially trained on various publicly available sources and then fine-tuned using a dedicated Thai dataset, yielding remarkable transcription accuracy on benchmark datasets. For fine-tuning in Thai, we utilize the Large Thai Multi-domain Multi-environment (LTMM) dataset[1] for model training in conjunction with Thai benchmark dataset. Our ASR model is fine-tuned Whisper [11] with the LTMM dataset using cross-entropy function for training sequence-to-sequence systems on classification tasks. In this setting, the system is trained to accurately classify the target text token from a pre-defined vocabulary of text tokens, as followed from this script[2]. Models trained on the LTMM dataset demonstrate improved generalization capabilities.

2.4. Natural Language Processing

In this study, we implement three text-based analytical methods following ASR processing: sentiment analysis, toxic classification, and sarcasm detection. Addressing the unique challenges presented by the Thai language, which is notably difficult for text preprocessing due to its lack of clear word boundaries [12], requires innovative solutions for effective feedback extraction. To tackle this, our methodology leverages a sophisticated transfer learning approach centered on the Elementary Discourse Unit (EDU) [13] for initial word tokenization. This technique is specifically designed to improve the accuracy of Thai text tokenization, where traditional methods are inadequate due to the complexity of the script. After tokenization, a refined analysis process is applied to these tokenized units to extract nuanced text-based feedback. This method not only addresses the linguistic challenges of Thai but also contributes to improving text analysis for under-resourced language, thereby ensuring the adaptability and efficacy of the system in deriving meaningful insights from Thai spoken content.

2.4.1. Sentiment Analysis

Thai sentiment analysis, classifying text into sentiment categories. We utilize advancements in pre-trained models, such as WangchanBERTa [5], and adopt EDU [13] for word tokenization during text preprocessing. Moreover, we fine-tune WangchanBERTa using a multi-layer perceptron (MLP) based on our dataset, specifically for text classification. This process

[1] github.com/JoesSattes/Large-Thai-Multi-domain-Multi-environment.git
[2] huggingface.co/blog/fine-tune-whisper

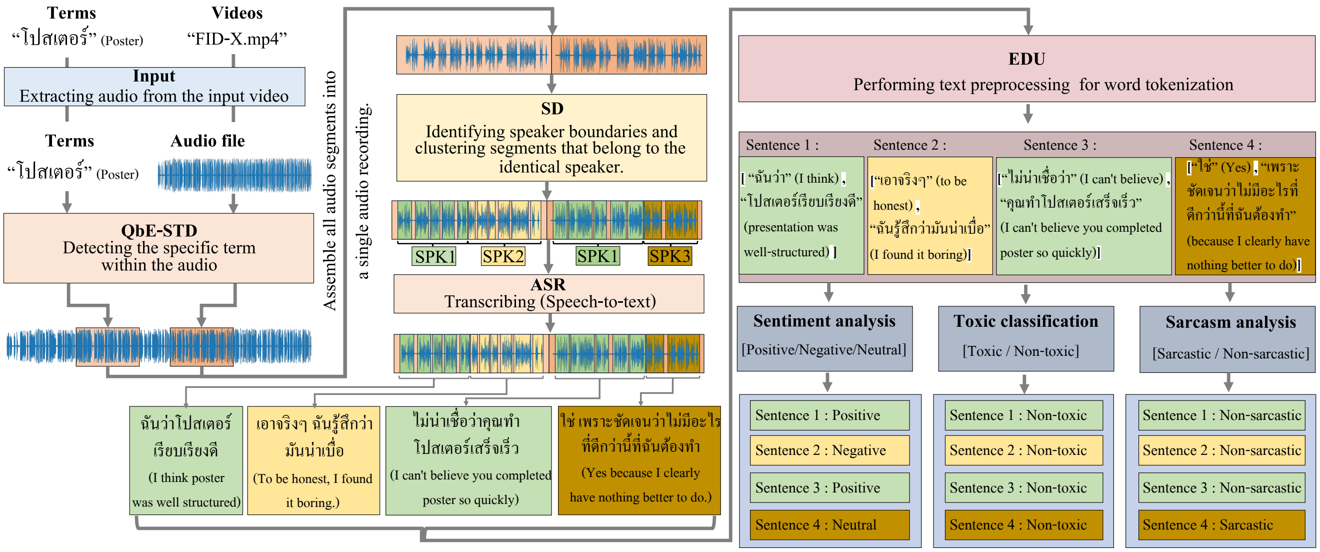
Terms
“โปสเตอร์” (Poster)
Videos
“FID-X.mp4”
Input
Extracting audio from the input video
Terms
“โปสเตอร์” (Poster)
Audio file
QbE-STD
Detecting the specific term within the audio
Assemble all audio segments into a single audio recording.
SD
Identifying speaker boundaries and clustering segments that belong to the identical speaker.
SPK1
SPK2
SPK1
SPK3
ASR
Transcribing (Speech-to-text)
ฉันว่าโปสเตอร์เรียบเรียงดี (I think poster was well structured)
เอาจริงๆ ฉันรู้สึกว่ามันน่าเบื่อ (To be honest, I found it boring.)
ไม่น่าเชื่อว่าคุณทำโปสเตอร์เสร็จเร็ว (I can't believe you completed poster so quickly)
ใช่ เพราะชัดเจนว่าไม่มีอะไรที่ดีกว่านี้ที่ฉันต้องทำ (Yes because I clearly have nothing better to do.)
EDU
Performing text preprocessing for word tokenization
Sentence 1 :
Sentence 2 :
Sentence 3 :
Sentence 4 :
[“ฉันว่า” (I think) , “โปสเตอร์เรียบเรียงดี” (presentation was well-structured)]
[“เอาจริงๆ” (to be honest) , “ฉันรู้สึกว่ามันน่าเบื่อ” (I found it boring)]
[“ไม่น่าเชื่อว่า” (I can't believe) , “คุณทำโปสเตอร์เสร็จเร็ว” (I can't believe you completed poster so quickly)]
[“ใช่” (Yes) , “เพราะชัดเจนว่าไม่มีอะไรที่ดีกว่านี้ที่ฉันต้องทำ” (because I clearly have nothing better to do)]
Sentiment analysis
[Positive/Negative/Neutral]
Toxic classification
[Toxic / Non-toxic]
Sarcasm analysis
[Sarcastic / Non-sarcastic]
Sentence 1 : Positive
Sentence 2 : Negative
Sentence 3 : Positive
Sentence 4 : Neutral
Sentence 1 : Non-toxic
Sentence 2 : Non-toxic
Sentence 3 : Non-toxic
Sentence 4 : Non-toxic
Sentence 1 : Non-sarcastic
Sentence 2 : Non-sarcastic
Sentence 3 : Non-sarcastic
Sentence 4 : Sarcastic

Fig. 1. The proposed negative feedback detection pipeline in Thai spoken content

is detailed in Section 3, and follows the script available here [3]. This approach aims to enhance the performance of the model in accurately classifying sentiment in Thai text.

2.4.2. *Toxic Classification*

We employ a transfer learning approach using the pre-trained WangchanBERTa model [5] for Thai toxic content classification. Our study explores two fine-tuning methods: normal and comparative. In the normal method, text representation is handled through a MLP. This involves point-wise MLP, implementing dropout at a rate of 0.1, and incorporating an MLP classification head. The comparative method, on the other hand, uses KimCNN [14] with contrastive learning for improved model robustness and extend comparative view between toxic and non-toxic.

2.4.3. *Sarcasm Classification*

Sarcasm Classification, which identifies instances of sarcasm in both spoken and written language, is a challenging task in NLP. Similar to our toxic content classification approach, our method utilizes a transfer learning technique with the pre-trained WangchanBERTa model [5]. This model is specifically fine-tuned for detecting sarcasm within the Thai language context, aiming to effectively discern sarcasm in various forms of communication.

3. EXPERIMENTAL SETUP

3.1. Experimental setting

In the realm of QbE-STD, our research employs a variety of English and Thai datasets, such as Buckeye [15], Librispeech [16], TIMIT [17], Command Voice 12 (English and Thai) [18] and Gowajee [19]. We initially trained a monolingual AWEs model using the Thai Common Voice dataset. This is followed by an assessment of word embedding models, enhanced through transfer learning from the pre-trained XLSR-53 model, with a specific focus on the Gowajee dataset. We use fixed-dimensional acoustic embedding matching for evaluation, comparing unseen Thai spoken queries from the 2017 and 2019 editions of Gowajee against test utterances from its 2018 and 2020 editions. For datasets without existing word timestamps, we utilize The Montreal Forced Aligners for accurate timestamping, adhering to evaluation criteria that target words longer than 0.5 seconds.

Speaker Diarization models, which identify non-speech segments, overlapping speech, and individual speakers, require extensively annotated audio recordings. The high cost and time investment of data collection is mitigated by using a spoken dialogue generator (SDG) [20] to create synthetic dialogue audios and annotations. This method, proven effective in training SD models, is applied using Thai data from the VoxLingua107 dataset. Our approach involves generating 1,000 dialogues, divided into an 80:20 training and test set ratio, to optimize SD model training through experimentation with various SDG parameters.

For ASR, we utilize the LTMM dataset[1], consisting of over 381 hours of speech recordings, to train a Thai ASR model. This dataset is crucial for training and evaluation, and we implement data preparation, as followed in this script[1], including

[3] huggingface.co/docs/transformers/main/en/tasks/sequence_classification

data cleansing, audio normalization and segmentation.

For Toxic Classification, we use the Thai tweet toxic dataset[3] contains 2160 sentences (828 sentences for non-toxic and 1332 sentences for toxic), with a division of 80% training, 10% validation, and 10% testing, employing a five-fold cross-validation methodology. Preprocessing includes removing irrelevant content and tokenization.

In Sentiment Analysis, we focus on financial banking content from the private Social dashboard dataset that aggregates Thai sentiment data from various sources (e.g. Twitter and Wisesight), comprising Thai user-generated content. Sentences are annotated by linguistic experts into positive, neutral, and negative categories, following preprocessing steps similar to those in our toxic classification experiment.

Sarcasm Classification is conducted using a specially curated Thai sarcasm dataset. Due to the scarcity of labelled Thai sarcasm data, Twitter hashtags are used to compile a dataset containing 50,000 instances each of sarcasm and non-sarcasm. In contrast to English, where the hashtag "#sarcasm" clearly signals sarcastic content and its absence indicates sincerity, Thai lacks a definitive marker to distinguish sarcasm. To bridge this gap, we propose introducing specific hashtags to help distinguish between sarcasm and sincerity in Thai. These include one hashtag for sarcasm ("#ประชด") and three for sincerity ("#ไม่ประชด", "#จริง," and "#พูดจริง"), facilitating a clear distinction between sarcastic and non-sarcastic statements. During data preprocessing, elements like user mentions, URL links, and the retweet label are eliminated. The dataset is processed in a manner consistent with our toxic classification experiment. Additionally, features such as emojis and hashtags are extracted to improve model interpretability.

Finally, the audio analytics system is implemented to compare the performance of two configurations: one with Spoken Term Detection and the other without it. This pipeline includes several components: QbE-STD, Speaker Diarization (SD), ASR, Sentiment Analysis, Sarcasm Classification, and Toxicity Classification. The comparison emphasizes processing time and overall efficiency, highlighting the differences between systems employing QbE-STD and those that do not. For this analysis, user-generated videos from YouTube of various lengths are utilized. The specific videos selected for this study can be accessed through the following YouTube links: J1MAZ9hAv2Q, PuQzw9RHg7c, CWl_we4fkEQ, 5ZUwsYN3y8E and bfSuFQObjcQ.

3.2. Evaluation

Evaluation metrics include Mean Average Precision (MAP) and Precision at 5 (P@5) for QbE-STD, Diarization Error Rate (DER) for SD, Word Error Rate (WER) for ASR, macro and micro F1-Scores for Sentiment Analysis, and accuracy, F1-score, precision, and recall for Toxic and Sarcasm Classifications. Processing time, measured in seconds, is the key metric for end-to-end evaluation.

Table 1. The performance of audio analytic system (%)

Audio analytic system	Aud. (sec.)	Proc. (sec.)	QbE-STD mAP	SD DER	ASR WER	Sent. F1-score	Tox. F1-score	Sarc. F1-score
w QbE-STD	3014.2	**48.4**	**83.7**	**10.4**	**22.9**	**93.5**	91.4	86.2
w/o QbE-STD		331.1	-	10.6	23.2	92.4	**92.1**	**86.5**

Table 2. The performance of word discrimination and QbE-STD (%)

Model	word dis.	QbE-STD			
		selected word		unseen word	
	mAP	mAP	P@5	mAP	P@5
A2E-Net with DWD loss	**81.48**	**72.23**	**76.9**	**58.21**	**65.18**
A2E-Net with softmax loss	77.29	66.98	75.88	52.12	64.22
XLR-53 with DWD loss	65.17	56.21	59.70	43.41	41.40
XLR-53 with softmax loss	62.83	51.06	58.59	38.72	32.27
XLSR-53	62.19	50.45	56.91	21.34	28.99

3.3. Data Preprocessing

Audio data undergoes resampling and conversion to mono channels for QbE-STD, SD, and ASR models. Text data is filtered for special characters and tokenized using EDU tokenization for sentiment analysis, toxic classification, and sarcasm classification.

4. EXPERIMENTAL RESULT

4.1. The performance of audio analytic system

The table presents a comparative analysis of the performance of audio analytics systems with and without QbE-STD. It details both the average audio duration (Aud.) and the average processing duration (Proc.) in seconds for the selected files. The inclusion of QbE-STD significantly reduces the processing time of the Audio Analytics System and marginally improves the WER and DER. Additionally, The system with QbE-STD excels in sentiment analysis (Sent.), showing a higher F1-score compared to the version without QbE-STD, which performs marginally better in toxicity classification (Tox.) and sarcasm classification (Sarc.).

4.2. The individual performance

The performance of QbE-STD : Table 2 shows the performance of our A2E-Net model in the QbE-STD task. Achieving an 81.48% mAP on command voice data, the effectiveness of the model is evident, particularly with the DWD loss application. On Gowajee dataset, a slight performance decrease for unseen words suggests potential areas for optimization.

The performance of SD : Evaluating different configurations of SD models, where the maximum number of utterances per speaker is set to 5, Table 3 indicates a DER of around

[3] https://huggingface.co/datasets/thai_toxicity_tweet

Table 3. The performance of speaker diarization

Setting	Max dialogue length (sec.)	Max no. of spk	DER (%)
1	2	3	15.79
2	2	2	**10.34**
3	3	3	13.18

13.10%. Notably, the number of speakers substantially affects a DER, while overlapping speech duration has a lesser impact.

Table 4. The WER of Thai ASR performance (%)

Approach	Model	Micro Average	Macro Average
Baseline	AIResearch's Wav2Vec2 (XLSR) [21]	39.53	38.88
Baseline	VISTEC's Wav2Vec2 (XLSR) with deepcut + LM [22, 23]	29.89	30.97
Baseline	VISTEC's Wav2Vec2 (XLSR) with newmm + LM [22, 24]	30.22	32.40
Baseline	OpenAI's Whisper-Small [25, 11]	120.16	98.23
Baseline	OpenAI's Whisper-Medium [25, 26]	86.34	65.42
Baseline	Thonburian Whisper [27, 28]	55.33	42.18
Fine-tuning	Facebook's Wav2Vec2-XLS-R-300M [29]	37.64	37.97
Fine-tuning	Facebook's Wav2Vec2-XLS-R-300M (LM) [29]	28.45	31.37
Fine-tuning	Whisper-Small	**23.06**	**19.12**

The performance of ASR : The WER for different Thai ASR models on the LTMM dataset is detailed in Table 4, with the Whisper-Small model demonstrating the lowest WER, underscoring its proficiency in Thai language transcription.

Table 5. The performance of sentiment classification (%)

Pretraining	Macro F1-Score	Micro F1-Score
XLM-Roberta-base	71.74	88.72
WangchanBERTa	**75.92**	**91.98**

The performance of sentiment classification : As indicated in Table 5, our WangchanBERTa model excelled in sentiment analysis, outperforming XLM-Roberta-base in both macro and micro F1-Scores.

The performance of toxic classification : Table 7 details the superior performance of WangchanBERTa in toxic classification, surpassing XLM-Roberta-base in all metrics.

The performance of sarcasm classification : Our WangchanBERTa model again proved superior, as shown in Table 7. It outperformed XLM-Roberta-base across all metrics, especially in comparative fine-tuning.

5. DISCUSSION AND LIMITATIONS

This research introduces the framework for identifying negative feedback within Thai spoken audio, leveraging an integration of QbE-STD, Speaker Diarization, ASR, and advanced NLP techniques. Our approach markedly enhances the efficiency and accuracy of processing complex, user-generated content by targeting specific spoken terms, a methodology especially beneficial in under-resourced languages like Thai. The incorporation of transfer learning has markedly enhanced the performance of the ASR and NLP components, as shown in Table 4, 5,6 and 7. The QbE-STD not only substantially reduced processing time but also achieved minor enhancements in WER, DER and F1-scores, proving its effectiveness in accurately categorizing text by sentiment. These advancements underscore the potential of the framework across various applications, from content moderation to decision support systems. The study encountered limitations, including decreased QbE-STD efficiency with unfamiliar words and increased DER in specific scenarios. These challenges underscore the necessity of further studies in future work to enhance the adaptability, generalization and computational efficiency of the framework.

Table 6. The performance of toxic classification model (%)

Fine-tuning	Pretraining	Accuracy	F1-Score	Precision	Recall
Comparative [14]	XLM-Roberta-base	82.5	82.5	82.5	82.5
Comparative [14]	WangchanBERTa	82.1	82.0	82.7	82.1
Normal	mBERT	83.21	83.21	83.21	83.21
Normal	XLM-Roberta-base	85.82	85.82	85.82	85.82
Normal	WangchanBERTa	**91.05**	**91.04**	**91.19**	**91.05**

Table 7. The performance of sarcasm classification model (%)

Fine-tuning	Pretraining	Accuracy	F1-Score	Precision	Recall
Comparative [14]	XLM-Roberta-base	83.13	83.12	83.14	83.13
Comparative [14]	WangchanBERTa	**86.01**	**86.01**	**86.02**	**86.01**
Normal	XLM-Roberta-base	81.47	81.47	81.48	81.48
Normal	WangchanBERTa	82.90	82.89	83.02	82.93

6. CONCLUSION

This study proposed an innovative framework to enhance the identification of negative feedback in Thai spoken audio, leveraging integration of QbE-STD, Speaker Diarization, ASR, and advanced NLP techniques, including sentiment analysis, toxic classification, and sarcasm detection. This approach significantly improves feedback retrieval in terms of the speed and accuracy.Notably, the utilization of QbE-STD for efficient term retrieval reduced processing time by 15% and increased the F1 score for sentiment analysis. Moreover, incorporating transfer learning boosted ASR performance, achieving WER of 19%, and enhanced text-based feedback analysis, yielding F1 scores of 76% for sentiment analysis, 91% for toxic classification, and 86% for sarcasm detection. These advancements are significant for under-resourced languages with ambiguous word boundaries, such as Thai, making substantial contributions to audio and text analytics.

7. REFERENCES

[1] Dhananjay Ram, Lesly Miculicich, and Hervé Bourlard, "Neural network based end-to-end query by example spoken term detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 1416–1427, 2019.

[2] Hadjer Bounazou, Nassim Asbai, and Sihem Zitouni, "Speaker diarization in overlapped speech," *2022 19th*

International Multi-Conference on Systems, Signals & Devices, pp. 890–895, 2022.

[3] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” *ArXiv*, vol. abs/2212.04356, 2022.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *ArXiv*, vol. abs/1810.04805, 2019.

[5] Lalita Lowphansirikul, Charin Polpanumas, Nawat Jantrakulchai, and Sarana Nutanong, “Wangchanberta: Pretraining transformer-based thai language models,” *ArXiv*, vol. abs/2101.09635, 2021.

[6] Edoardo Savini and Cornelia Caragea, “Intermediate-task transfer learning with bert for sarcasm detection,” *Mathematics*, 2022.

[7] Vrunda N. Sukhadia and Srinivasan Umesh, “Domain adaptation of low-resource target-domain models using well-trained asr conformer models,” *2022 IEEE Spoken Language Technology Workshop*, pp. 295–301, 2022.

[8] Thititorn Seneewong Na Ayutthaya and Kitsuchart Pasupa, “Thai sentiment analysis via bidirectional lstm-cnn model with embedding vectors and sentic features,” *2018 International Joint Symposium on Artificial Intelligence and Natural Language Processing*, pp. 1–6, 2018.

[9] Pantid Chantangphol, Theerat Sakdejayont, and Tawunrat Chalothorn, “Enhancing word discrimination and matching in query-by-example spoken term detection with acoustic word embeddings,” in *Proceedings of the 6th International Conference on Natural Language and Speech Processing*. 2023, pp. 293–302, Association for Computational Linguistics.

[10] Hervé Bredin and Antoine Laurent, “End-to-end speaker segmentation for overlap-aware resegmentation,” in *Proc. Interspeech 2021*, 2021.

[11] Radford, Alec and Kim, Jong Wook and Xu, Tao and Brockman, Greg and McLeavey, Christine and Sutskever, Ilya, “whisper-small,” 2022.

[12] Sattaya Singkul, Borirat Khampingyot, Nattasit Maharattamalai, Supawat Taerungruang, and Tawunrat Chalothorn, “Parsing thai social data: A new challenge for thai nlp,” in *2019 14th International Joint Symposium on Artificial Intelligence and Natural Language Processing*. IEEE, 2019, pp. 1–7.

[13] Chanatip Saetia, Supawat Taerungruang, and Tawunrat Chalothorn, “Combining thai edus: Principle and implementation,” in *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, 2020, pp. 270–278.

[14] Reem Abdel-Salam, “reamtchka at SemEval-2022 task 6: Investigating the effect of different loss functions for sarcasm detection for unbalanced datasets,” in *Proceedings of the 16th International Workshop on Semantic Evaluation*, Seattle, United States, July 2022, pp. 896–906, Association for Computational Linguistics.

[15] Mark A. Pitt, Keith Johnson, Elizabeth Hume, Scott F. Kiesling, and William D. Raymond, “The buckeye corpus of conversational speech: labeling conventions and a test of transcriber reliability,” *Speech Commun.*, vol. 45, pp. 89–95, 2005.

[16] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” *2015 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 5206–5210, 2015.

[17] Carla Lopes and Fernando Perdigão, “Timit acoustic-phonetic continuous speech corpus,” 2012.

[18] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber, “Common voice: A massively-multilingual speech corpus,” in *International Conference on Language Resources and Evaluation*, 2019.

[19] Ekapol Chuangsuwanich, Atiwong Suchato, Korrawe Karunratanakul, Burin Naowarat, Chompakorn CChaichot, Penpicha Sangsa-nga, Thunyathon Anutarases, Nitchakran Chaipojjana, and Yuatyong Chaichana, “Gowajee Corpus,” Tech. Rep., Chulalongkorn University, Faculty of Engineering, Computer Engineering Department, 12 2020.

[20] Chunlei Zhang, Jiatong Shi, Chao Weng, Meng Yu, and Dong Yu, “Towards end-to-end speaker diarization with generalized neural speaker clustering,” in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2022, pp. 8372–8376.

[21] VISTEC-depa AI Research Institute of Thailand, “wav2vec2-large-xlsr-53-th,” 2023.

[22] Wannaphong Phatthiyaphaibun, Chompakorn Chaksangchaichot, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong, “Thai wav2vec2. 0 with commonvoice v8,” *arXiv preprint arXiv:2208.04799*, 2022.

[23] Wannaphong Phatthiyaphaibun, “wav2vec2-large-xlsr-53-th with deepcut,” 2022.

[24] Wannaphong Phatthiyaphaibun, “wav2vec2-large-xlsr-53-th with newmm,” 2022.

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*. PMLR, 2023, pp. 28492–28518.

[26] Radford, Alec and Kim, Jong Wook and Xu, Tao and Brockman, Greg and McLeavey, Christine and Sutskever, Ilya, “whisper-medium,” 2022.

[27] Atirut Boribalburephan, Zaw Htet Aung, Knot Pipatsrisawat, and Titipat Achakulvisut, “Thonburian whisper:

A fine-tuned whisper model for thai automatic speech recognition," 2022.

[28] Biomedical and Data Lab, Mahidol University, "whisper-th-medium-combined," 2022.

[29] Arun Babu and Changhan Wang and Andros Tjandra and Kushal Lakhotia and Qiantong Xu and Naman Goyal and Kritika Singh and Patrick von Platen and Yatharth Saraf and Juan Pino and Alexei Baevski and Alexis Conneau and Michael Auli, "wav2vec2-xls-r-300m," 2021.