

Enhance Mono-modal Sentiment Classification With Federated Cross-modal Transfer

Xueyang Wu[†] Di Jiang[‡] Yuanfeng Song[‡] Qian Xu[‡] Qiang Yang[†]

[†] Department of CSE, HKUST, Hong Kong, China {xwuba, qyang}@cse.ust.hk

[‡] AIGroup, WeBank Co. Ltd., Shenzhen, China {dijiang, yfsong, qianxu}@webank.com

Abstract

Sentiment analysis is a complex process that involves multiple modalities, which can provide more accurate and informative results than using a single modality. Although existing multimodal approaches have shown to be superior to mono-modal sentiment classification, they are not always practical in real-world scenarios where only mono-modal input is available, or where multimodal data is limited due to data scarcity or privacy concerns. To address this issue, we propose a novel approach that enhances mono-modal sentiment classification through federated transfer learning. Specifically, we focus on a practical industrial problem where text and speech data are owned by different affiliations, and we aim to bridge these modalities by sharing a cross-modal feature generator and phone classifier. Our proposed framework also incorporates differential privacy techniques to ensure privacy-preserving cross-modal transfer. Our experimental results on real-world spoken language sentiment classification corpora demonstrate the effectiveness of our proposed framework. We show that our approach can significantly improve the accuracy of mono-modal sentiment classification, even when only a limited amount of data is available.

1 Introduction

Sentiment classification has recently attracted significant interest due to its ability to automatically recognize the polarity of human emotional states or attitudes expressed in spoken or written language, which is crucial for improving the user experience in human-machine interaction. Communication among humans involves various modalities, including textual and acoustic content, facial expressions, and body gestures. However, single modality fails to capture sentimental information entirely and leads to inaccurate classification. To address this issue, researchers have proposed multi-modal sentiment or emotion classification approaches that utilize features from different modalities to enhance classification accuracy [1, 2]. Text, speech, and vision are three critical modalities for sentiment classification [3, 4], and previous research has proposed various fusion strategies to better utilize multiple modalities [5, 6]. Among these modalities, some researchers highlight the importance of text and speech modalities [7].

Despite the advantages of multi-modal sentiment classification, two obstacles hinder the application of traditional multi-modal sentiment classification in industry: *data scarcity* and *user privacy*. Collecting multi-modal

Copyright 2004 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

data is challenging in real-world applications, and mono-modal speech and text data are often the only available options in call centers or text-based conversation systems. Furthermore, annotating parallel multi-modal data is costly and laborious. Moreover, user privacy has become an increasing concern, and many authorities have enacted regulations to protect data and model privacy, such as the EU General Data Protection Regulation (GDPR) [8] in 2018, followed by the US and China.

To address these challenges, we propose a novel federated machine learning paradigm for sentiment classification. Since mono-modal sentimental corpora are more accessible and cost-effective, it is worth enhancing mono-modal sentiment classification if we can leverage multiple corpora with different modalities, leveraging the complementarity of different modalities from multiple sources. To address user privacy concerns when using multiple corpora, we introduce federated learning [9], a collaborative machine learning paradigm where multiple parties jointly learn a global model without exposing their private local data. Our work focuses on scenarios where each party in the federation owns a single-modal sentimental corpus and aims to enhance their sentiment classification performance collaboratively. Further details on related work are provided in Section 2.

Our proposed framework enables multiple institutes with different modality corpora to collaborate and transfer their knowledge learned from their data in a privacy-preserving manner. The cross-modal transfer is based on the phone sequence, which can be obtained from both text and speech modalities, embedding both phonetic cues and lexical information. Our method takes mono-modal input but utilizes cross-modal transfer to expand the number of modalities. Using the phone sequence as the intermedium between speech and text modality has three benefits: 1) the phone sequence is easy to obtain for both textual and audio input without modifying the existing model significantly; 2) as an intermedium between speech and text, it preserves semantic meaning and speech characteristics; 3) it is straightforward to align the audio pieces and words using the phone sequence at the utterance level. The alignment allows us to enrich the raw text input with speech features learned from the audio modality and embed the acoustic features with semantic meanings learned from the text. During collaborative training, speech-side and text-side institutions join as a federation, only exchanging intermediate model parameters protected by the differential-privacy mechanism instead of sharing raw data or bare model parameters.

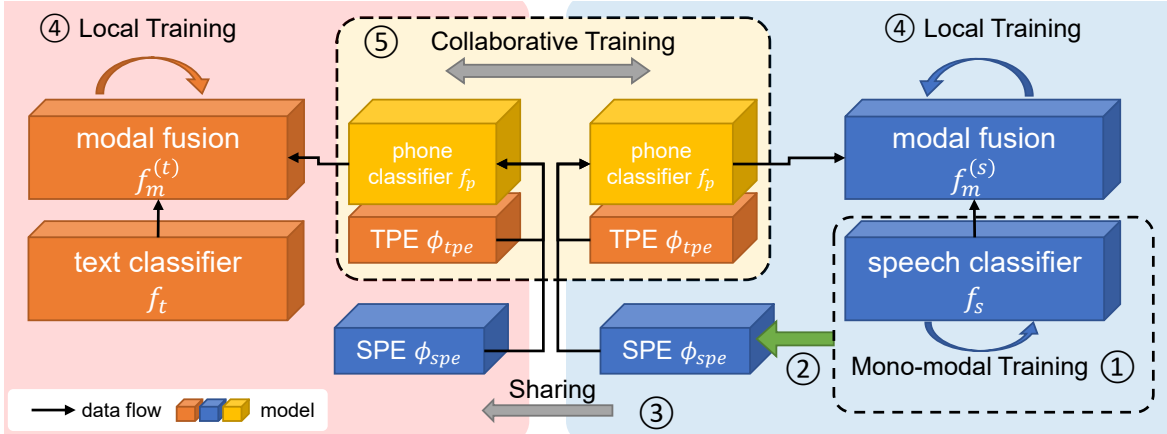


Figure 1: The Joint-view of Cross-modal Transfer Learning Framework of Text and Speech module

2 Related Work

2.1 Multi-modal Sentiment Analysis

Multi-modal sentiment and emotion analysis often involve multiple modalities, such as text, speech, and vision [3, 4]. Different approaches have been proposed to leverage the information from multiple modalities more effectively [5, 6, 10], showing that well-designed fusion strategies significantly affect the model performance. Other research works have further explored two-modal sentiment analysis, such as visual-audio fusion [11, 3] and textual-audio fusion [1, 7]. These works typically focus on model architecture design or representation extraction. In contrast, our proposed method enhances mono-modal sentiment classification by a novel federated cross-modal transfer framework, where each party in the federation owns data in only one modality.

2.2 Cross-modal Transfer

Cross-modal transfer has been used in image classification since 2013 [12]. It allows an image classifier to classify a given image class, even if it has not been seen in the training data. Cross-modal transfer leverages the textual description of the object, forces the textual modal and visual modal embedding to map into the same space, and hence allows zero-shot inference with only a textual description of the unseen class. Existing cross-modal transfer methods for sentiment classification often focus on fusing audio and visual modalities based on the correlation between speech and facial expression [13, 14]. Our approach differs in that it focuses on the semantic complementarity between speech and spoken words. We use phonetic features as an intermediary to connect audio and textual information. For example, when in intense emotion, the pronunciation of words may have distortion, which cannot be observed in the text generated by automatic speech recognition (ASR) but can be captured in the phone sequence.

2.3 Federated Learning

Federated learning was first proposed by Google researchers in 2016 [15] as a way to learn a global model without explicitly gathering data from different clients. Federated learning has since been extended to various architectures [9, 16], such as horizontal federated learning, vertical federated learning, and federated transfer learning, with different privacy protection approaches, such as differential privacy [17], homomorphic encryption [18], and multiparty security computation [19]. Among these privacy and security protection techniques, differential privacy has emerged as a widely adapted approach for deep learning because it is practical in terms of computational efficiency. Differential privacy for deep learning is a mechanism that injects deliberately generated noise into the model during the training phase, which offers strong and robust guarantees to bound the probability of revealing raw data under a reasonable threshold. However, there is a tradeoff between high utility and high privacy protection, as injecting more noise into the model leads to less utility. To achieve higher privacy protection, we apply differential privacy to our proposed method, and to maintain high utility, the noises are only injected into a small portion of the model parameters that are shared across parties. Our proposed framework is well-suited for federated learning as it allows two parties to collaborate and share knowledge while keeping sensitive data locally, which effectively relieves the challenges of data scarcity and data privacy in sentiment classification.

3 Federated Cross-modal Transfer

3.1 Settings and Notations

For simplification, we describe our setting as consisting of two parties, and we can easily extend it to the multi-party scenario. We summarize our cross-modal transfer framework in the form of the flow chart in Fig. 1,

where the red block stands for text modality and the blue one for speech modality. The circled numbers state the general steps of cross-modal joint learning.

Denote datasets for text-side and speech-side parties as $\mathcal{D}_t = \{\mathbf{w}_i, y_i\}_{i=1}^{N_t}$ and $\mathcal{D}_s = \{\mathbf{s}_i, y_i\}_{i=1}^{N_s}$, respectively, consisting of pairs of the text \mathbf{w} or speech \mathbf{s} and the label y , where N_t and N_s are the numbers of training instances of the text-side party and the speech-side party. We use subscripts $(\cdot)_s$, $(\cdot)_t$ to distinguish where a variable belongs to. As a decentralized federated learning setting, this framework does not involve a central orchestrator, two parties directly set up connections and exchange parameters with each other.

3.2 Phone as Intermedium

We use phone-level information as an intermedium between text and speech modalities in this work. Phones are the units of speech that constitutes the pronunciation of words. The phone sequence captures the speech pronunciation of a sentence or a word, including the tones and pronunciation variation, such as stress and rhythm, which implies a speaker’s sentiment. Phonetic cues have been used to enhance multi-lingual text classification [20] and text representation learning [21], as phones entail more semantic and sentiment cues than text. On the other hand, the phone sequence is obtainable from both speech and text sides, making it a good bridge to connect the two modalities. For example, a classic DNN-HMM hybrid ASR system can conduct the `speech-to-phone` transformation to produce the phone sequence of a given speech as the byproduct. The text can also be transformed into a phone sequence with the lexicon representing the “standard” pronunciation of the text, denoting this process as `text-to-phone`.

3.3 Federated Cross-modal Transfer Framework

As shown in Fig. 1, our federated cross-modal transfer framework involves two parties and can be summarized in 5 steps. For clarity, we denote the speech-side (blue block) as *client-s*, and the text-side (red block) as *client-t*. *Client-s* and *client-t* are two independent institutes that collaboratively enhance their mono-modal sentiment classifier with federated cross-modal transfer with each following the Alg. 2 and 3 respectively.

Algorithm 1 Federated cross-modal transfer

Require: two parties *client-s* and *client-t* with speech sentiment corpus and text sentiment corpus respectively;

- 1: *client-s* conducts lines 1-5 in Alg. 2 for initialization;
 - 2: *client-t* conducts lines 1-2 in Alg. 3 for initialization;
 - 3: **while not** reach maximum rounds **do**
 - 4: *client-s* conducts lines 7-12 in Alg. 2 and *client-t* conducts lines 4-9 in Alg. 3 **parallelly**;
 - 5: *client-s* sends $\phi_{tpe}^{(s)}, f_p^{(s)}, f_m^{(s)}$ to *client-t*;
 - 6: *client-t* sends $\phi_{tpe}^{(t)}, f_p^{(t)}, f_m^{(t)}$ to *client-s*;
 - 7: *client-s* conducts lines 13 to update $\phi_{tpe}^{(s)}, f_p^{(s)}, f_m^{(s)}$;
 - 8: *client-t* conducts lines 10 to update $\phi_{tpe}^{(t)}, f_p^{(t)}, f_m^{(t)}$;
 - 9: **end while**
-

Step 1. The workflow of the proposed framework starts with *client-s* training a mono-modal speech sentiment classifier. Then *client-s* build a Speech-Phone-Extractor (SPE) that embeds the speech signal-level information to phonemes. The speech classifier (f_s) and text classifier (f_t) are implemented with mono-modal models mentioned in Section 4.4.

Step 2. To avoid missing the focus on our framework, we propose a vanilla SPE practice, which represents a phone with its responding f_s averaged by utterances where this appears.

Step 3. Sharing SPE helps client- t build a homogeneous phone classifier as client- s , which allows two clients to start federated learning, even though their input modalities are different.

Step 4-5. Client- s and client- t have different input modalities and mono-modal models while sharing the same phone classifier and the modal-fusion weights. The input of client- s is a piece of speech features \mathbf{x} , noted as $\mathbf{s} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_T]$, extracted from the raw audio wave, where T is the number of frames. The input of client- t is a sentence, i.e., a word sequence $\mathbf{w} = [w_1, w_2, \dots, w_L]$, where L is the length of the sentence. We hence obtains phone sequences from \mathbf{s} and \mathbf{w} with

$$\begin{aligned}\mathbf{p}_s &= \text{speech-to-phone}(\mathbf{s}), \\ \mathbf{p}_t &= \text{text-to-phone}(\mathbf{w}).\end{aligned}\tag{1}$$

Thereafter, each client has two input sequences: speech/text sequence and phone sequence. The speech/text sequence is fed into the mono-modal model, i.e.,

$$\mathbf{o}_s = f_s(\mathbf{s}), \quad \mathbf{o}_t = f_t(\mathbf{w}).\tag{2}$$

Processing phone sequence is identical for the two clients, so we ignore the superscript of the models $\phi_{tpe}^{(\cdot)}, f_p^{(\cdot)}, f_m^{(\cdot)}$. We extract the distributed speech-phone representation \mathbf{r}_{sp} and text-phone representation \mathbf{r}_{tp} of the given phone sequence \mathbf{p} ,

$$\mathbf{r}_{sp} = \phi_{spe}(\mathbf{p}), \quad \mathbf{r}_{tp} = \phi_{tpe}(\mathbf{p}).\tag{3}$$

The phone representations from two modalities are concatenated and fed into the phone classifier, i.e.,

$$\mathbf{r}_p = [\mathbf{r}_{sp}; \mathbf{r}_{tp}], \quad \mathbf{o}_p = h_p(\mathbf{r}_p)\tag{4}$$

We apply the late fusion (a.k.a., decision-level fusion) strategy [5] to combine information from different modalities, i.e.,

$$y = f_m(\mathbf{o}_s, \mathbf{o}_p), \quad y = f_m(\mathbf{o}_t, \mathbf{o}_p),\tag{5}$$

We compute the cross-entropy loss with the output y and label l as well as the gradients of parameters through Eq. (1)-(5). The training data of text and speech are store at each client isolatedly, and the two clients only share $\phi_{tpe}^{(\cdot)}, f_p^{(\cdot)}$, and $f_m^{(\cdot)}$ to each other. $\phi_{tpe}^{(\cdot)}$ represents $\phi_{tpe}^{(s)}$ and $\phi_{tpe}^{(t)}$, for simplicity, as well as $f_p^{(\cdot)}$, and $f_m^{(\cdot)}$. All models are updated with the Stochastic Gradient Descent optimization algorithm (SGD). To achieve differentially private models, the shared models are updated with the differentially private version of SGD that provides privacy protection of the released model [22], noted as DP-SGD. In more detail, we inject Gaussian noise [23] to the parameters during optimization. The scale of noise is related to the privacy budget (ϵ, δ) indicating the probabilities of leaking privacy. A larger privacy budget allows less model perturbation.

Each client conducts local training using their local data, and conducts weighted average over parameters of $\phi_{tpe}^{(\cdot)}, f_p^{(\cdot)}$, and $f_m^{(\cdot)}$. As a decentralized federated learning scheme, two parties directly exchange differentially private parameters of shared models with each other without the need for a central server [16]. Within each party, the weighted averaging follows Eq. (6), i.e.,

$$\begin{aligned}\phi_{tpe}^{(s)} &= \phi_{tpe}^{(t)} = \frac{(|\mathcal{D}_s| \cdot \phi_{tpe}^{(s)} + |\mathcal{D}_t| \cdot \phi_{tpe}^{(t)})}{(|\mathcal{D}_s| + |\mathcal{D}_t|)}, \\ f_p^{(s)} &= f_p^{(t)} = \frac{(|\mathcal{D}_s| \cdot f_p^{(s)} + |\mathcal{D}_t| \cdot f_p^{(t)})}{(|\mathcal{D}_s| + |\mathcal{D}_t|)}.\end{aligned}\tag{6}$$

Algorithm 2 DP cross-modal transfer (**client-s**)

Require: local training data \mathcal{D}_s ;

- 1: initialize the local speech classifier (f_s);
 - 2: train f_s with local training data;
 - 3: build SPE $\phi_{spe}^{(s)}$ with f_s according to Alg. 4;
 - 4: share $\phi_{spe}^{(s)}$ to client- t ;
 - 5: initialize TPE $\phi_{tpe}^{(s)}$, phone-classifier $f_p^{(s)}$, and model-fusion function $f_m^{(s)}$;
 - 6: **while not** reach maximum rounds **do**
 - 7: **for** local training iteration i **do**
 - 8: sample (s, l) from local training data;
 - 9: compute gradients of f_s , $\phi_{tpe}^{(s)}$, $f_p^{(s)}$, and $f_m^{(s)}$ with (s, l) according to Eq. (1)-(5);
 - 10: update parameters of f_s with SGD;
 - 11: update parameters of $\phi_{tpe}^{(s)}$, $f_p^{(s)}$, and $f_m^{(s)}$ with DP-SGD;
 - 12: **end for**
 - 13: send differentially private parameters of $\phi_{tpe}^{(s)}$ and $f_p^{(s)}$ to client- t ;
 - 14: receive differentially private parameters of $\phi_{tpe}^{(t)}$ and $f_p^{(t)}$ from client- t ;
 - 15: update $\phi_{tpe}^{(s)}$ and $f_p^{(s)}$ according to Eq. (6);
 - 16: **end while**
 - 17: **return** $\phi_{spe}^{(s)}$, $\phi_{tpe}^{(s)}$, $f_p^{(s)}$, and $f_m^{(s)}$;
-

Algorithm 3 DP cross-modal transfer (**client-t**)

Require: local training data \mathcal{D}_t ;

- 1: receive $\phi_{spe}^{(s)}$ from client- t ;
 - 2: initialize $\phi_{tpe}^{(t)}$, $f_p^{(t)}$, $f_m^{(t)}$, and text classifier (f_t);
 - 3: **while not** reach maximum rounds **do**
 - 4: **for** local training iteration i **do**
 - 5: sample (w, l) from local training data ;
 - 6: compute gradients of f_t , $\phi_{tpe}^{(t)}$, $f_p^{(t)}$, and $f_m^{(t)}$ with (w, l) according to Eq. (1)-(5);
 - 7: update parameters of f_t with SGD;
 - 8: update parameters of $\phi_{tpe}^{(t)}$, $f_p^{(t)}$, and $f_m^{(t)}$ with DP-SGD;
 - 9: **end for**
 - 10: send differentially private parameters of $\phi_{tpe}^{(t)}$ and $f_p^{(t)}$ to client- s ;
 - 11: receive differentially private parameters of $\phi_{tpe}^{(s)}$ and $f_p^{(s)}$ from client- s ;
 - 12: update $\phi_{tpe}^{(t)}$ and $f_p^{(t)}$ according to Eq. (6);
 - 13: **end while**
 - 14: **return** $\phi_{tpe}^{(t)}$, $f_p^{(t)}$, and $f_m^{(t)}$;
-

4 Experiments

4.1 Experimental Setup

In this study, we assess the performance of our proposed framework on two public multi-modal datasets, namely MOSI [24] and MOSEI [25], which are available in the CMU Multimodal Data SDK ¹.

4.2 Dataset Description

Table 1 presents a summary of the datasets, including their size and label distributions. For both datasets, we leverage the training set to train the model and the validation set to fine-tune the hyperparameters. The evaluation of the models is performed on the test set using the hyperparameters chosen through the validation set.

4.3 Baseline Models and Evaluation Metrics

To demonstrate the efficacy of our framework, we compare it with two classes of representative mono-modal sentiment classification methods, i.e., Textual Modal Model and Audio Modal Model, referred to as **baseline** models. We also perform experiments in the centralized **oracle** settings. To control for variables, the oracle settings employ the neural network model proposed in this study, but its inputs (text and speech) are aligned per utterance. The oracle settings are ideal but impractical, and they indicate the model’s full potential. To further validate the compatibility of our framework with state-of-the-art models, such as BERT [26] and Transformer, we substitute the mono-modal text model with pre-trained BERT_{Large} and the speech model with a standard Transformer [27].

All models are tuned with the validation set, and the evaluation metrics are **F1-scores** reported on the testing sets, which includes four testing sets obtained from two modalities and two datasets.

Table 1: The statistics of the reference label distribution

Dataset	MOSI	MOSEI
Train	1283 605:678 [‡]	16331 8279:8052
Valid	299 105:124	1871 939:932
Test	686 409:277	5057 2375:2287

[‡] positive-negative count of reference labels

4.4 Implementation

We provide details of the implementation of our proposed framework, specifically focusing on the phone feature extractor and the acoustic phone feature mapper.

The textual phone feature extractor (ϕ_{tp}) leverages the Forward-Maximum-Matching (FMM) algorithm to translate an utterance \mathbf{w} to the corresponding phone sequence \mathbf{p}_t , with the help of a lexicon.

To recover the acoustic feature from the phone sequence, we propose a concise acoustic phone feature mapper (ϕ_{sp}). We build a mapper from each context-free phone to an acoustic feature, either from MFCCs or openSMILE [28], following Algorithm 4. For a given phone sequence \mathbf{p}_s , we generate a sequence of acoustic feature vectors \mathbf{R}_s . We obtain an utterance vector by taking the mean over \mathbf{R}_s .

As we focus on the sentiment classification task, the sentiment labels in the datasets are normalized to positive (intensity > 0) and negative (intensity ≤ 0). For the acoustic model, we use a handcrafted feature set

¹<https://github.com/A2Zadeh/CMU-MultimodalDataSDK>

extracted by openSMILE² [28]. This produces a set of features that indicate intensity, loudness, Mel-frequency cepstral coefficients (MFCCs), and pitch. openSMILE extracts 384-dimensional acoustic features for every 100ms-frame. The maximum sequence lengths for words, phones, and acoustic features are all limited to 800.

Algorithm 4 Build speech-phone-extractor

Require: an utterance-level acoustic feature extractor M_s for speech classification, and a set of speech \mathbf{X}_s , a vocabulary of phone V

- 1: $\mathbf{P}_s = \text{speech-to-phone}(\mathbf{X}_s)$;
- 2: **for** each phone p in V **do**
- 3: $\mathbf{P}_{sp} = \{\mathbf{p}_s \mid p \in \mathbf{p}_s, \forall \mathbf{p}_s \in \mathbf{P}_s\}$;
- 4: $\mathbf{R}_{sp} = \{M_s(\mathbf{p}_s) \mid \mathbf{p}_s \in \mathbf{P}_{sp}\}$;
- 5: $\mathbf{r}_{sp} = \frac{\sum_{\mathbf{r} \in \mathbf{R}_{sp}}(\mathbf{r})}{|\mathbf{R}_{sp}|}$;
- 6: $\text{map}[p] = \mathbf{r}_{sp}$;
- 7: **end for**
- 8: $\phi_{spe}(\mathbf{p}) := (\text{map}[p_j] \mid p = [p_1, \dots, p_j, \dots, p_N])$;
- 9: **return** ϕ_{spe}

We have implemented the Textual Modal Model using a multi-layer bi-directional GRU model [29], referred to as **BiGRU**. For the Audio Modal Model, we have used a Convolutional Neural Network [30] (**CNN**). Furthermore, to ensure a fair comparison between different settings, we have limited the training iteration at each epoch to 1000.

The proposed framework and baselines have been implemented using PyTorch [31], and the differential private algorithm has been applied using PyVacy [32]. We have used the differentially private optimizer [22] to train the shared phone classifier. In the differential privacy settings, we have set all δ to 10^{-5} , and we have varied ϵ to evaluate both protection and performance. All experiments have been conducted on a machine equipped with an Intel(R) Xeon(R) CPU E5-2630, 128 GB RAM, and 4 NVIDIA GeForce GTX TiTian XP GPUs.

4.5 Experimental Results

Table 2: The F1-score of cross-modal transfer under different privacy budgets

Dataset	Baseline (Mono-modal)	Oracle (Centralized)	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 15$	$\epsilon = \infty$	Avg. #turn
MOSI-MOSI	55.61	59.23	57.59	57.56	57.53	57.50	56.77	3
MOSEI-MOSEI	62.02	66.91	63.54	65.84	65.82	65.71	65.73	4.8
MOSI-MOSEI	55.61	N/A	57.65	57.53	58.04	58.01	57.88	7.2
MOSEI-MOSI	62.02	N/A	65.86	65.56	65.88	65.86	65.58	5.8
MOSI-MOSI	57.42	65.23	63.32	61.59	61.33	60.81	61.97	4.2
MOSEI-MOSEI	68.53	70.91	70.30	70.54	70.51	69.98	69.73	5
MOSI-MOSEI	68.53	N/A	69.74	70.04	69.95	69.73	69.76	6.8
MOSEI-MOSI	57.42	N/A	62.01	62.86	62.31	60.16	60.38	6.8

Table 2 presents the performance of four data settings with varying privacy budgets. The first column provides the details of the four settings in the form of *speech-side* and *text-side* data. In addition to transferring

²We used the IS09 configuration from <https://github.com/naxingyu/opensmile/blob/master/config/>.

between identical data distributions (MOSI-MOSI and MOSEI-MOSEI), we also validate our framework’s effectiveness for non-IID and imbalanced distributions (MOSEI-MOSI and MOSI-MOSEI), which are more realistic in real-world scenarios ³. The columns from $\epsilon = 1$ to $\epsilon = \infty$ display the results under different parameter settings. The larger the value of ϵ , the weaker the privacy protection, and the less noise injected into the framework. According to [33], ϵ between 6 and 14 is practical in real-world applications. When $\epsilon = \infty$, the model is trained without any differential privacy protection (e.g., FedAvg).

The results in Table 2 show that our federated cross-modal transfer framework enhances performance in all settings. Moreover, we observe that our framework effectively improves the performance of mono-modal inputs to reach the ideal level of aligned multimodal inputs, as compared to the oracle settings. For instance, under-setting $\epsilon = 5$, our framework achieves absolute accuracy improvements of 3.82% and 2.01% over the baselines in MOSEI and MOSE datasets, respectively. The results also suggest that the improvement is more significant for the side with less training data (as shown in the MOSI-MOSI and MOSEI-MOSI results from Table 2). The numbers in the last column demonstrate that our framework converges within a few rounds, indicating low network overload. Furthermore, we find that stricter privacy protection does not necessarily lead to performance decreases. In fact, smaller privacy budgets may provide excellent performance, where the injected noise acts as regularization for the model, avoiding overfitting. Importantly, our proposed framework is fully compatible with secure-enhanced schemes such as homomorphic encryption [34] and secure multi-party computing [35].

Table 3: F1-score of large-scale models under different privacy budgets

Dataset	Baseline (Mono-modal)	Oracle (Centralized)	$\epsilon = 1$	$\epsilon = 5$	$\epsilon = 10$	$\epsilon = 15$	$\epsilon = \infty$	Avg. #turn
MOSI-MOSI	56.98	66.13	60.77	60.96	60.32	59.53	57.17	2
MOSEI-MOSEI	62.01	70.22	66.14	66.81	66.22	66.44	66.73	4.4
MOSI-MOSEI	56.98	N/A	60.12	60.33	61.11	60.01	59.88	6
MOSEI-MOSI	62.01	N/A	67.16	67.86	66.28	65.98	65.38	4.6
MOSI-MOSI	72.26	75.32	73.12	72.99	72.53	71.81	70.97	4.2
MOSEI-MOSEI	74.20	78.38	76.60	77.45	77.23	76.89	76.73	8
MOSI-MOSEI	74.20	N/A	76.47	77.23	76.95	76.83	76.76	10.8
MOSEI-MOSI	72.26	N/A	73.20	74.16	73.93	73.71	73.88	6.8

In order to evaluate the potential of our proposed framework on large-scale models, we replaced the simple CNN speech model and BiGRU text model with a standard Transformer and a pre-trained BERTLarge model, respectively. The results are presented in Table 3, which shows four settings with two testing sets under different privacy budgets, following the same format as the baseline models. Our findings are interesting. Firstly, we observed that the performance on the text-side testing set is significantly improved by using a pre-trained model. Nevertheless, our federated cross-modal transfer still helps improve the mono-modal performance, approaching the oracle performance. On the other hand, upgrading the speech-side model to a Transformer does not yield a remarkable improvement, and the performance boost provided by our framework remains consistent with the baseline models. Furthermore, we noted that when the models are larger, weaker DP protections are more likely to result in overfitting (as seen in the last few columns of Table 3).

³Note that the training data of speech-side and text-side are neither parallel nor aligned

5 Conclusions

In this paper, we have proposed a framework for privacy-preserving cross-modal sentiment classification, which leverages the power of multi-modal input features and models trained on different modalities to enhance performance. Our experimental results, both on classic and large-scale models, demonstrate the efficacy of our framework in achieving higher accuracy in sentiment classification, alleviating data scarcity issues, and preserving data privacy for all parties involved. Our work represents a promising initial exploration into knowledge transfer among private modality data, which may hold great potential for improving mono-modal performance via federated cross-modal transfer. However, we acknowledge that this approach also presents challenges, such as privacy protection and communication efficiency, which require further research. We hope that this work will inspire future investigations in this area.

References

- [1] B. Li, D. Dimitriadis, and A. Stolcke, “Acoustic and lexical sentiment analysis for customer service calls,” in *ICASSP*. IEEE, 2019, pp. 5876–5880.
- [2] A. Korhonen, D. R. Traum, and L. Màrquez, Eds., *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*. Association for Computational Linguistics, 2019. [Online]. Available: <https://www.aclweb.org/anthology/P19-1000/>
- [3] M. Wöllmer, F. Weninger, T. Knaup, B. Schuller, C. Sun, K. Sagae, and L.-P. Morency, “Youtube movie reviews: Sentiment analysis in an audio-visual context,” *IEEE Intelligent Systems*, vol. 28, no. 3, pp. 46–53, 2013.
- [4] V. Rozgić, S. Ananthakrishnan, S. Saleem, R. Kumar, and R. Prasad, “Ensemble of svm trees for multi-modal emotion recognition,” in *Proceedings of The 2012 Asia Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2012, pp. 1–4.
- [5] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, “Tensor fusion network for multimodal sentiment analysis,” *arXiv preprint arXiv:1707.07250*, 2017.
- [6] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, “Multimodal sentiment analysis using hierarchical fusion with context modeling,” *Knowledge-Based Systems*, vol. 161, pp. 124–133, 2018.
- [7] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, “Context-dependent sentiment analysis in user-generated videos,” in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 873–883.
- [8] P. Voigt and A. Von dem Bussche, “The eu general data protection regulation (gdpr),” *A Practical Guide, 1st Ed.*, Cham: Springer International Publishing, 2017.
- [9] Q. Yang, Y. Liu, T. Chen, and Y. Tong, “Federated machine learning: Concept and applications,” *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 10, no. 2, pp. 1–19, 2019.
- [10] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional mkl based multimodal emotion recognition and sentiment analysis,” in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

- [11] A. Metallinou, S. Lee, and S. Narayanan, “Audio-visual emotion recognition using gaussian mixture models for face and voice,” in *2008 Tenth IEEE International Symposium on Multimedia*. IEEE, 2008, pp. 250–257.
- [12] R. Socher, M. Ganjoo, C. D. Manning, and A. Ng, “Zero-shot learning through cross-modal transfer,” in *Advances in neural information processing systems*, 2013, pp. 935–943.
- [13] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, “Emotion recognition in speech using cross-modal transfer in the wild,” in *2018 ACM Multimedia Conference on Multimedia Conference, MM 2018, Seoul, Republic of Korea, October 22-26, 2018*, 2018, pp. 292–301. [Online]. Available: <https://doi.org/10.1145/3240508.3240578>
- [14] S. H. Dumpala, I. Sheikh, R. Chakraborty, and S. K. Kopparapu, “Audio-visual fusion for sentiment classification using cross-modal autoencoder,” *NIPS2018 ViGIL Workshop*, 2018.
- [15] H. B. McMahan, E. Moore, D. Ramage, S. Hampson *et al.*, “Communication-efficient learning of deep networks from decentralized data,” *arXiv preprint arXiv:1602.05629*, 2016.
- [16] Q. Yang, Y. Liu, Y. Cheng, Y. Kang, T. Chen, and H. Yu, *Federated Learning*. Morgan & Claypool Publishers, 2020.
- [17] C. Dwork, F. McSherry, K. Nissim, and A. Smith, “Calibrating noise to sensitivity in private data analysis,” in *Theory of cryptography conference*. Springer, 2006, pp. 265–284.
- [18] R. L. Rivest, L. Adleman, M. L. Dertouzos *et al.*, “On data banks and privacy homomorphisms,” *Foundations of secure computation*, vol. 4, no. 11, pp. 169–180, 1978.
- [19] A. C.-C. Yao, “Protocols for secure computations,” in *FOCS*, vol. 82, 1982, pp. 160–164.
- [20] S. K. Singh and M. K. Sachan, “Classification of code-mixed bilingual phonetic text using sentiment analysis,” *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol. 17, no. 2, pp. 59–78, 2021.
- [21] H. Peng, Y. Ma, S. Poria, Y. Li, and E. Cambria, “Phonetic-enriched text representation for chinese sentiment analysis with reinforcement learning,” *Information Fusion*, vol. 70, pp. 88–99, 2021.
- [22] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep learning with differential privacy,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*. ACM, 2016, pp. 308–318.
- [23] C. Dwork, A. Roth *et al.*, “The algorithmic foundations of differential privacy,” *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.
- [24] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, “Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos,” *arXiv preprint arXiv:1606.06259*, 2016.
- [25] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, “Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph,” in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2018, pp. 2236–2246.
- [26] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

- [27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” in *Advances in neural information processing systems*, 2017, pp. 5998–6008.
- [28] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: the munich versatile and fast open-source audio feature extractor,” in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 1459–1462.
- [29] J. Chung, C. Gulcehre, K. Cho, and Y. Bengio, “Empirical evaluation of gated recurrent neural networks on sequence modeling,” *arXiv preprint arXiv:1412.3555*, 2014.
- [30] Y. Kim, “Convolutional neural networks for sentence classification,” *arXiv preprint arXiv:1408.5882*, 2014.
- [31] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in *NIPS Autodiff Workshop*, 2017.
- [32] C. Waites, “Pyvacy: Towards practical differential privacy for deep learning,” *Article*, 2019.
- [33] A. Orr, “Google’s differential privacy may be better than apple’s,” 2017. [Online]. Available: <https://www.macobserver.com/analysis/google-apple-differential-privacy/>
- [34] Y. Liu, Y. Kang, C. Xing, T. Chen, and Q. Yang, “A secure federated transfer learning framework,” *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 70–82, 2020.
- [35] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, “Practical secure aggregation for privacy-preserving machine learning,” in *proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*, 2017, pp. 1175–1191.