Toxic Speech and Speech Emotions: Investigations of Audio-based Modeling and Intercorrelations

Wei-Cheng Lin*
Department of Electrical and Computer Engineering
The University of Texas at Dallas
Richardson, USA
wei-cheng.lin@utdallas.edu

Dimitra Emmanouilidou
Audio and Acoustics Group
Microsoft Research
Redmond, USA
dimitra.emmanouilidou@microsoft.com

***Abstract*—Content moderation (CM) systems have become essential following the monumental increase in multimodal and online social platforms; and while increasingly published work focuses on text-based solutions, there is still limited work on audio-based methods. In this study we aim to explore relationships between speech emotions and toxic speech, as part of a CM scenario. We first investigate an appropriate framework for combining speech emotion recognition (SER) and audio-based CM models. We then investigate which emotional aspects (i.e., attribute, sentiment, or attitude) could contribute the most in facilitating audio-based CM recognition platforms. Our experimental results indicate that conventional shared feature encoder approaches may fail to capture additional discriminative features for boosting audio-based CM tasks while utilizing SER learning. We further investigate performance trade-offs of late-fusion frameworks for combining SER and CM information. We argue that these observations could be attributed to an emotionally-biased distribution in the CM scenario, concluding that SER could in deed play a role in content moderation frameworks, given added application-specific emotional information.**

***Index Terms*—speech emotion recognition, audio-based content moderation, toxic language detection, sentiment detection**

I. Introduction

Increasing and widespread availability of social media platforms has enabled people to more freely interact with others online. Activities such as social networking or online gaming provide enriched and convenient interactions between people, but could also inevitably attract negative, aggressive and even toxic behaviors, leading to cyberbullying, harassment or hate speech phenomena [1], [2]. Traditional approaches adopted by social platforms for identifying these harmful behaviors rely heavily on manual examination by human moderators, which is costly, inefficient and non-scalable. Automated content moderation (CM) systems have recently welcomed machine learning techniques for resolving traditional limitations [3], thanks to advancements in machine learning and deep learning algorithms, and to increasingly available data resources.

Popular platforms such as Twitter and Facebook utilize conventionally text-based frameworks for CM, with the objective of identifying whether a post or comment contains toxic information based on some extracted textual features. For instance, Del *et al.* [4] utilized common morpho-syntactic and sentiment polarity features to build a long short-term memory (LSTM) hate speech recognition model for Facebook comments. Similarly, Nobata *et al.* [5] showed the combination of different standard natural language processing (NLP) features (e.g., N-gram, POS tags) and semantic embeddings (e.g., *word2vec*) could lead to outperforming performances over baselines that use single aspect feature on abusive language detection. More recently, many studies have exploited powerful pretrained language models (e.g., BERT) for extracting robust neural representation of words, and then perform additional finetuning [6] or transfer learning [7] stage for the downstream CM tasks, achieving state-of-the-art recognition performances.

However, when it comes to voice-based social platforms (e.g., multiplayer online gaming or voice/video chatting), purely text-based models may not be sufficient for comprehensively capturing a user's intent, especially for those cues that are embedded in acoustic-only features. A few recently published studies with a focus on acoustic features such as Mel-spectrogram [8] or Mel-frequency cepstrum coefficients (MFCCs) [9] have shown promising results for detecting toxic speech. Besides toxic-related acoustic patterns, people's emotional states have further been found to be closely related to their toxic behaviors [10]. Therefore, speech emotion recognition (SER) could be an important technique to facilitate the exploration in the audio-based CM research field.

In this study, we investigate whether speech emotions provide complementary information for identifying toxic speech in a content moderation scenario of online gaming. Our analysis is based on deep learning frameworks and our contributions are divided into three parts: 1) the modeling strategy for combining SER and audio-based CM; 2) the choice of the most relevant emotional view or aspect (attributes, sentiments or categories); 3) the implications of fusing SER information in an audio-based CM system. To the best of our knowledge, there is no prior work on a comprehensive automated CM framework, aimed at bringing insights into potential intercorrelations between speech emotions and toxic speech moderation.

II. Background

In NLP and text-based CM, employing additional information from emotional cues is not a new concept. Prior work has explored different approaches to exploit sentiment and emotion information; a straightforward method is to augment morpho-syntactic (e.g., part-of-speech, POS) and stylometric (e.g., function words, FW) patterns with emotion-based features, as

* This work was done as research intern at Microsoft Research, Redmond.

the input to train a CM recognition model. Markov *et al.* [11] adopted the NRC emotion lexicon to encode emotion-based features by emotional word counts and their associations of emotions (e.g., sadness). The lexicon helps extract emotion information contained in the text, and can also be directly applied to perform unsupervised sentiment analysis and clustering for identifying hate speech [12]. Besides handcrafted features, deep learning frameworks can directly extract the discriminative emotional features by treating the emotional information as a supervised recognition task [13]; or fine-tune pre-trained language models such as BERT using additional sentiment corpus to obtain emotion-relevant neural representations for enhancing the CM prediction accuracy [14].

However, there is still no literature that explores the integration of speech emotions in an audio-based CM task. Conventionally, SER is regarded as a sequence-to-one recognition task which aims to identify the emotional state of an input audio clip based on its acoustic patterns. Traditional handcrafted features often contain energy, spectral and frequency-based acoustic parameters [15] such as formants, MFCCs and pitch. The emotional labels are usually attributes (e.g., arousal, valence, dominance) [16] or categories (e.g., joy and angry) [17], annotated in sentence-level manner (i.e., one global label is assigned per spoken sentence). Recently, deep network architectures using raw waveform [18] or spectrogram [19] have achieved competitive performances compare to traditional handcrafted features. Below, we investigate the role of SER for identifying toxic speech in a content moderation scenario, and the potential of leveraging acoustic-based features of SER towards more effective CM techniques.

III. Resources and Investigation Settings

A. Datasets

- *MS-CM*: For the content moderation scenario, a private dataset of voice clips is used, obtained from an online gaming-related platform, and comprised of verbal communication that was self-reported for violating Microsoft's public user agreement policies as related to toxic, discriminatory or abusive behavior. The voice recordings had a maximum duration of 15 secs and were annotated by a policy-expert annotator into two classes namely toxic and non-toxic, according to company's moderation rules and guidelines. In total, there are 630K utterances with an imbalanced label distribution of 1:5 for the toxic and non-toxic classes. This corpus contains in-the-wild recordings with uncontrolled microphone and channel settings, and is riddled by contamination and artifacts such as background music playing and other noises [8].
- *1D-IEMOCAP*: For complimentary SER learning, we considered IEMOCAP [20], a popular public benchmark dataset that contains 12 hours of scripted or improvised dialogues from 10 actors. Both attribute and categorical labels will be used in this study, separately, from all utterance-split dialogues.
- *5D-Collection*: This augmented emotional corpus considers a collection of categorically-labeled speech segments from 5 public datasets, IEMOCAP, EmoDB, RAVDESS, VAM and eNTERFACE [20]–[24], in an effort to simulate a diversified dataset that comes closer to realistic needs of a SER system.
- *Atti-1D MS-Call*: Here we consider another view (attitude view) of the emotional information with speech segments coming from the purposefully-mismatched domain of support calls, yet close enough to the domain of emotions. The motivation is to incorporate sentiment information from different views or scenarios, and also to enlarge the available emotional data with more natural speech, since most public datasets are non-spontaneous (i.e., actors acting in controlled environments), and have a relatively small data size (i.e., less than 10K utterances per set). This corpus includes 110K utterances of private customer service support calls, labeled according to the attitude as negative, neutral or positive.

We use these four distinct sets of corpora to investigate which aspect of emotions could provide complementary information to our audio-based CM scenario.

TABLE I
The detailed model parameters.

Layer	Channels/Nodes	Kernel	Stride	Activation
Input	1	N/A	N/A	N/A
CNN-block	32	$(5, 5)$	1	ReLU
MaxPooling	N/A	$(4, 4)$	4	N/A
CNN-block	64	$(3, 3)$	2	ReLU
MaxPooling	N/A	$(2, 2)$	2	N/A
CNN-block	64	$(3, 3)$	2	ReLU
Flatten/Dropout	$p = 0.5$	N/A	N/A	N/A
FCN	256	N/A	N/A	ReLU
BLSTM	64	N/A	N/A	Tanh
Output Layers	depends	N/A	N/A	depends

B. Acoustic Features and Model Architecture

Data were downsampled to 16KHz, 16-bit mono, excluding segments beyond $[1, 15]$ secs duration. We extract logarithmic mel-spectrograms as input features for both the SER and audio-based CM models. The 512-dim magnitude spectrogram was computed over 32ms windows with 50% overlap, and then mapped onto the mel scale using 64 mel-frequency filters, followed by z-normalization.

We used the same core architecture and parameters throughout all experiments, consisting of the CNN-BLSTM model. We first adopt the dynamic segmentation formula proposed by [16] to split originally varied length (on the temporal dimension) input feature map (i.e., the log-mel-spectrogram) into a fixed number of equal-length data chunks. The CNN part of the model is responsible for encoding the coarse-level representations, while the concatenated bidirectional LSTM layer captures temporal information across different data chunks for summarizing the final sentence-level feature representation. The detailed model structure is shown in Table I; the CNN-block consists of a 2D-CNN with BatchNorm and ReLU activation function and the output layer consists of a fully connected layer (FCN) using ReLU activation and a top layer depending on the recognition tasks and their loss functions. More specifically, the audio-based CM model is a binary classifier using Sigmoid activation and binary cross entropy

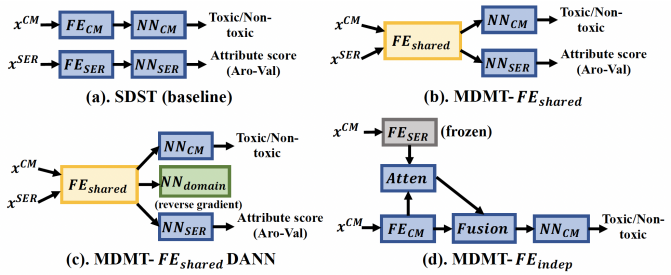

Fig. 1. Baseline models (a) and different modeling approaches for integrating SER in an audio-based CM model (b-d).

(BCE) loss, while the SER model can be either a multi-class classifier for emotional categories or an emotional regressor for attribute scores optimized with the CE or concordance correlation coefficient (CCC) loss function, respectively. We refer to the CNN-BLSTM layers as the feature extractor *FE* and the FCN output layers as *NN*, Figure 1(a).

We used the Adam optimizer at a fixed learning rate of 10^{-4} with 128 batch size, and 100 epochs maximum, sufficient for all models to converge. In addition, we used an uniform label sampler to obtain balanced label distribution for every batch in training. We applied early stopping criteria using a small development set. For corpora without predefined train, development and test sets, we performed 5-fold cross validation (CV) using speaker independent split sets and similar label distribution among sets. The average test results from all 5-CVs constitute the final model performance. The performance metrics used in this work are the weighted and unweighted F1-scores, and the precision and recall of the toxic class.

IV. Investigation Studies

A. Preliminary Study of Modeling Methodology

As the goal is to compare different modeling approaches for incorporating SER information into the audio-based CM, we first build two baseline models looking at each task as an independent recognition problem: one for CM using the MS-CM corpus; and one for SER using the IEMOCAP corpus for the arousal and valence regressors. We refer to the baseline problem formulation as single-(data) domain-single-task (**SDST**) in Figure 1(a), where the recognition model of each task is trained with its own task-specific data and labels. Table II contains the baseline or benchmark performance of the SER model (**SDST-SER**), demonstrating the effectiveness of the proposed model architecture; it achieves competitive performance with other state-of-the-art benchmarks on IEMOCAP. Since the focus of this work is mainly on how SER information could boost CM, we will move our focus from the SER performance to the CM performance while we incorporate different emotional information to the CM task, and we will compare the results to the CM baseline (**SDST-CM**).

We consider the combination of SER and audio-based CM as the multi-(data) domains-multi-tasks (**MDMT**) problem, involving two different recognition tasks from different datasets. One of the most common modeling methodology is to train

TABLE II
Performance comparison for baseline task SDST-SER.

Metric	*SDST-SER (IEMOCAP)*
CCC (ours)	arousal=**0.656**, valence=0.406
CCC [16]	arousal=0.629, valence=0.365
CCC [27]	arousal=0.590, valence=**0.421**

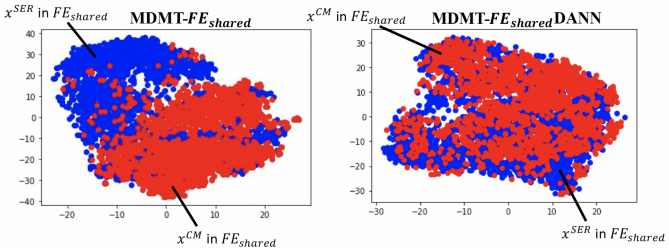

Fig. 2. tSNE plots of the shared *FE* models. SER datapoints appear in blue, and CM datapoints in red. Domain separation is mitigated with DANN (right).

a shared feature encoder (i.e., the FE_{shared} part of our model) for the two tasks, shown in Figure 1(b). Findings indicate that the trained model is trying to memorize domain-specific information of the two different tasks. This can be seen in the t-SNE [25] plot visualization in Figure 2 left panel, where we see a clear separation between the shared features *FE* extracted for the SER datapoints (blue color) and the CM datapoints (red color). To eliminate the task-specific separation, we extend the model with an additional domain-adversarial classifier, DANN [26] (**MDMT-FE**$_{shared}$ **DANN** in Figure 1(c)), using a reverse gradient operation to confuse the model. This indeed helps the model create shared features with no apparent separation for the two datasets as seen in Figure 2 right panel. However, moving from the **MDMT-FE**$_{shared}$ model to the **MDMT-FE**$_{shared}$ **DANN** model results in a small unexpected performance degradation, Table III: partial domain-specific information might be critical to capturing discriminating features, and that training shared features jointly may not always constitute a favorable option.

B. Different Emotional Aspects and their Impact

The shared *FE* model failed to illustrate benefits in combing SER and audio-based CM, Table III. In this section, we train the *FE* for each task independently and then perform late feature-fusion [28] to incorporate emotional information to CM, Figure 1(d), **MDMT-FE**$_{indep}$. We first pretrain on the SER task and freeze the trained *FE* model to extract emotional-relevant features on the *MS-CM* corpus; this will constitute the SER-related information channel. The other channel is a supervised audio-based CM model with weighted fusion for incorporating the SER-related channel. The weighted fusion is achieved by a standard channel-wise attention model with Softmax activation weights for explicitly capturing the contribution of the two information channels (i.e., weights are constraint to sum up to 1). Although there are various ways

TABLE III
COMPARISON OF RECOGNITION PERFORMANCES FOR DIFFERENT MODELING APPROACHES FOR INCORPORATING SER TO CM.

Metric	*SDST-CM*	FE_{shared}	FE_{shared} *DANN*
F1-weighted	**0.770**	0.765	0.756
F1-unweighted	**0.640**	0.611	0.607
Toxic-Precision	**0.385**	0.356	0.349
Toxic-Recall	**0.473**	0.376	0.396

TABLE IV
COMPARISON OF RECOGNITION PERFORMANCES FOR DIFFERENT EMOTIONAL ASPECTS CONTRIBUTING TO THE AUDIO-BASED CM MODEL.

Metric	*SDST-CM*	*Attr-1D*	*Senti-1D*	*Senti-5D*	*Atti-1D*
F1-weighted	0.770	0.771	**0.776**	**0.773**	0.769
F1-unweighted	0.640	0.637	0.640	**0.642**	0.640
Toxic-Precision	0.385	0.384	**0.393**	**0.391**	0.384
Toxic-Recall	**0.473**	0.457	0.453	0.467	0.474

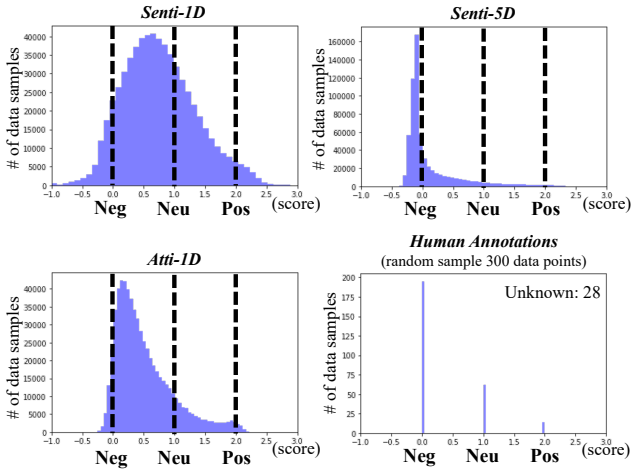

Fig. 3. Distributions of the predicted emotion for task CM using 3 different SER models, along with the human labeling for reference.

to fuse information (e.g., concatenation, mean fusion), we did not find significant performance difference among them. The major advantage of this modeling methodology is that one can flexibly replace the pretrained SER in Figure 1(d), and further facilitate the investigation of different SER emotional aspects contributing to the audio-based CM. Here, we divide the emotional aspects into three different types: *attribute* (Attr), *sentiment* (Senti) and *attitude* (Atti). These types will be defined in conjunction with the corpora of Section III-A.

- *Attr-1D*: This SER model is an emotion regressor trained on IEMOCAP arousal and valence attributes. The "1D" means only one dataset is used for training.
- *Senti-1D*: The SER model is a 3-class classifier. The categorical labels are mapped to three sentiment classes, positive, neutral and negative (e.g. fear, anger, sad and disgust are mapped to the negative sentiment class). The SER model is trained on IEMOCAP only.
- *Senti-5D*: Same 3-class classifier as *Senti-1D*, but here SER training includes the full *5D-Collection* corpora.
- *Atti-1D*: A 3-class classifier trained on negative, neutral and positive attitudes of *Atti-1D MS-Call* corpora.

The caveat is to verify that the pretrained SER model captures reliable emotion information for *MS-CM*, before using it as an emotion channel. As emotion labels are not available for the *MS-CM* corpus, we manually labeled a random sample of 300 audio clips using the three sentiment classes to form a human annotation reference of the emotion distribution. Hard-to-annotate clips (due to extreme artifacts or low SNR) were deemed *Unknown*. The last panel of Figure 3 portrays the reference distribution of emotion, with an obvious bias towards negative sentiment, which is not surprising considering the nature of the *MS-CM* corpus. Next, we turn into the model's sentiment predictions. We replace the objective function from CE to CCC loss (from classifier to regressor) to obtain continuous sentiment scores from emotional aspects. The regressor sets positives further apart from negatives, than neutral. The first 3 panels in Figure 3, illustrate a negative-biased distribution, in agreement with the manual annotations. This means that the pretrained SER models are effective for capturing the emotional trend of the *MS-CM* corpus.

Table IV shows performances on the four emotional aspects of the SER model, as compared to the CM baseline (**SDST-CM**). Main takeaways: (i) the *Senti* emotion generally obtains improved performances (except the Toxic-Recall) over the *Attr* and *Atti*. On average, *Senti-5D* achieves the best performance, arguably because it incorporates more diverse speech data, potentially closer resembling the realistic conditions of the MS-CM corpus. (ii) We see a trade-off trend of the improved precision but degraded recall performances for the toxic class under the *Senti* aspect of SER, compared to the baseline. A closer look helped us verify that this trade-off is a consistent trend even under different size of training data, Figure 4 left. In summary, the *Senti* emotional aspect is a more favorable integration to the audio-based CM model, and results in a trade-off trend of the toxic precision and recall performances. However, diversifying the SER training datasets (*Senti-5D* vs *Senti-1D*) can help remedy the recall degeneration.

Finally, we notice that the overall improved performance of the CM model is quite moderate after adding SER information. Taking a closer look at the attention weights during *Senti-5D* test phase, we can separate the SER contribution (blue color) and the audio-based CM (red color), Figure 4 right; high values indicate higher contribution. The distribution shows that SER contributes significantly less compared to the CM features w.r.t. the overall audio-based CM model (i.e., most attention weights for the SER channel fall below 0.2). The major contributor to this may be the negative sentiment-bias in the online CM scenario or the more controlled recording settings of the SER domain when compared to the CM corpus.

V. CONCLUSIONS

In this study, we investigated different automatic modeling methodologies for integrating speech emotion recognition

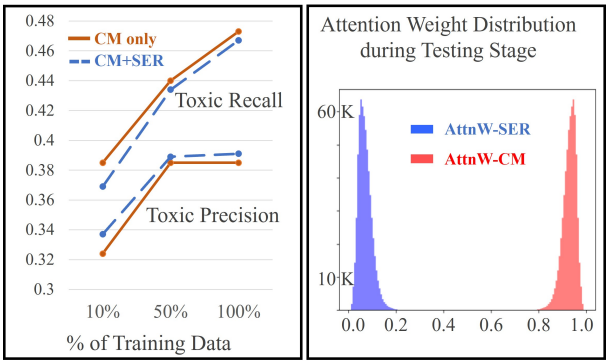

Fig. 4. Trade-off performance trend for different training data size (left); attention weights distribution during testing (right).

(SER) to an audio-based content moderation (CM) model. We found that a conventional shared feature encoder (FE_{shared}) framework is incapable of capturing additional or enhanced discriminative information for the CM domain. An independently trained *FE* with late fusion and attention leads to an improved precision with the trade-off of a degraded recall performance. We found that speech sentiment (negative, neutral, positive) is a more favorable emotion aspect for audio-based CM recognition compared to emotional attribute or attitude. Diversifying the speech emotion corpora in the SER module of the model leads to higher toxic recall performance for CM. Finally, we observed that an emotional-biased CM data distribution can limit the importance of SER contributions, deeming SER integration dependent on the CM scenario. Additional exploration is needed for understanding the true role of emotional features in automated content moderation frameworks, and a closer look to different content moderation scenarios is part of our future work. We hope that this study serves as reference for further exploration in this area.

REFERENCES

[1] J. A. Pater, M. K. Kim, E. D. Mynatt, and C. Fiesler, "Characterizations of online harassment: Comparing policies across social media platforms," in *Proceedings of the 19th international conference on supporting group work*, 2016, pp. 369–374.

[2] R. Slonje, P. K. Smith, and A. Frisén, "The nature of cyberbullying, and strategies for prevention," *Computers in human behavior*, vol. 29, no. 1, pp. 26–32, 2013.

[3] B. Gambäck and U. K. Sikdar, "Using convolutional neural networks to classify hate-speech," in *Proceedings of the first workshop on abusive language online*, 2017, pp. 85–90.

[4] F. Del Vigna12, A. Cimino23, F. Dell'Orletta, M. Petrocchi, and M. Tesconi, "Hate me, hate me not: Hate speech detection on facebook," in *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, 2017, pp. 86–95.

[5] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th international conference on world wide web*, 2016, pp. 145–153.

[6] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, "Hatebert: Retraining bert for abusive language detection in english," *arXiv preprint arXiv:2010.12472*, 2020.

[7] M. Mozafari, R. Farahbakhsh, and N. Crespi, "A bert-based transfer learning approach for hate speech detection in online social media," in *International Conference on Complex Networks and Their Applications*. Springer, 2019, pp. 928–940.

[8] M. Yousefi and D. Emmanouilidou, "Audio-based toxic language classification using self-attentive convolutional neural network," in *Proc. European Signal Processing Conf. (EUSIPCO)*. IEEE, 2021.

[9] A. Iskhakova, D. Wolf, and R. Meshcheryakov, "Automated destructive behavior state detection on the 1d cnn-based voice analysis," in *International Conference on Speech and Computer*. Springer, 2020, p. 184.

[10] C. Calvert, "Hate speech and its harms: A communication theory perspective," *Journal of Communication*, vol. 47, no. 1, pp. 4–19, 1997.

[11] I. Markov, N. Ljubešić, D. Fišer, and W. Daelemans, "Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection," in *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, 2021, pp. 149–159.

[12] A. Rodriguez, C. Argueta, and Y.-L. Chen, "Automatic detection of hate speech on facebook using sentiment and emotion analysis," in *2019 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*. IEEE, 2019, pp. 169–174.

[13] S. Rajamanickam, P. Mishra, H. Yannakoudakis, and E. Shutova, "Joint modelling of emotion and abusive language detection," *arXiv preprint arXiv:2005.14028*, 2020.

[14] A. Elmadany, C. Zhang, M. Abdul-Mageed, and A. Hashemi, "Leveraging affective bidirectional transformers for offensive language detection," *arXiv preprint arXiv:2006.01266*, 2020.

[15] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan *et al.*, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2015.

[16] W.-C. Lin and C. Busso, "Chunk-level speech emotion recognition: A general framework of sequence-to-one dynamic temporal modeling," *IEEE Transactions on Affective Computing*, 2021.

[17] S. Mirsamadi, E. Barsoum, and C. Zhang, "Automatic speech emotion recognition using recurrent neural networks with local attention," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2017, pp. 2227–2231.

[18] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *Proc. IEEE Intl. Conf. on Acoustics, Speech and Signal Process. (ICASSP)*. IEEE, 2016, pp. 5200–5204.

[19] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in *2017 international conference on platform technology and service (PlatCon)*. IEEE, 2017, pp. 1–5.

[20] C. Busso, M. Bulut, C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. Chang, S. Lee, and S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Journal of Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, December 2008.

[21] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *9th European Conference on Speech Communication and Technology (Interspeech'2005 - Eurospeech)*, Lisbon, Portugal, September 2005, pp. 1517–1520.

[22] M. Grimm, K. Kroschel, and S. Narayanan, "The Vera am Mittag German audio-visual emotional speech database," in *IEEE International Conference on Multimedia and Expo (ICME 2008)*, Hannover, Germany, June 2008, pp. 865–868.

[23] S. Livingstone and F. Russo, "The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," *PLOS ONE*, vol. 13, no. 5, pp. 1–35, May 2018.

[24] O. Martin, I. Kotsia, B. Macq, and I. Pitas, "The enterface'05 audio-visual emotion database," in *22nd International Conference on Data Engineering Workshops (ICDEW'06)*. IEEE, 2006, pp. 8–8.

[25] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[26] Y. Ganin, E. Ustinova, H. Ajakan, P. Germain, H. Larochelle, F. Laviolette, M. Marchand, and V. Lempitsky, "Domain-adversarial training of neural networks," *The journal of machine learning research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[27] B. T. Atmaja and M. Akagi, "Improving valence prediction in dimensional speech emotion recognition using linguistic information," in *2020 23rd Conference of the Oriental COCOSDA International Committee for the Co-ordination and Standardisation of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2020, pp. 166–171.

[28] C. G. Snoek, M. Worring, and A. W. Smeulders, "Early versus late fusion in semantic video analysis," in *Proceedings of the 13th annual ACM international conference on Multimedia*, 2005, pp. 399–402.