Project Vāc: Can a Text-to-Speech Engine Generate Human Sentiments?

Shivam Kulkarni*, Luis Barbado*, Jordan Hosier†‡, Yu Zhou†, Siddharth Rajagopalan§, and Vijay K. Gurbani†*
*Illinois Institute of Technology. {skulkarni17,lbarbado}@hawk.iit.edu / vgurbani@iit.edu
†Vail Systems, Inc. {jhosier,yzhou,vgurbani}@vailsys.com
‡Northwestern University. jordanhosier2020@u.northwestern.edu
§siddharth.rajagopalan@gmail.com

***Abstract*—Sentiment analysis is an important area of natural language processing (NLP) research, and is increasingly being performed by machine learning models. Much of the work in this area is concentrated on extracting sentiment from textual data sources. Clearly however, a textual source does not convey the pitch, prosody, or power of the spoken sentiment, making it attractive to extract sentiments from an audio stream. A fundamental prerequisite for sentiment analysis on audio streams is the availability of reliable acoustic representation of sentiment, appropriately labeled. The lack of an existing, large-scale dataset in this form forces researchers to curate audio datasets from a variety of sources, often by manually labeling the audio corpus. However, this approach is inherently subjective. What appears "positive" to one human listener may appear "neutral" to another. Such challenges yield sub-optimal datasets that are often class imbalanced, and the inevitable biases present in the labeling process can permeate these models in problematic ways. To mitigate these disadvantages, we propose the use of a text-to-speech (TTS) engine to generate *labeled* synthetic voice samples rendered in one of three sentiments: positive, negative, or neutral. The advantage of using a TTS engine is that it can be abstracted as a function that generates an infinite set of labeled samples, on which a sentiment detection model can be trained. We investigate, in particular, the extent to which such training exhibits acceptable accuracy when the induced model is tested on a separate, independent and identically distributed speech source (i.e., the test dataset is not drawn from the same distribution as the training dataset). Our results indicate that this approach shows promise and the induced model does not suffer from underspecification.**

***Index Terms*—Sentiment, Sentiment Detection, Text-to-Speech (TTS), Regression Models.**

I. Introduction and Problem Statement

Sentiment analysis is an active research area in the paralinguistic processing community. Traditionally, this line of research is rooted in the investigation of techniques for computationally detecting the polarity (e.g. positive, negative, neutral) towards products, services, organizations, individuals, and events by mining textual sources like blogs, Twitter feed, and other social networking platforms [1]. Sentiment analysis is distinct from *emotion* analysis as noted by Munezero et al. [2], and best illustrated by an example:[1] the sentences "this product wasn't what I expected", and "I hate this product with the white hot fury of a thousand suns" are both negative sentiments. Yet, clearly, the second sentence packs more emotion than the first one. In a sentiment analysis model, it suffices to detect the polarity of an utterance without attributing an emotive state to it.

A representative example of using a sentiment detection model is customer service in call centers. Automatically detecting sentiment contained in human interactions can aid in proactively sensing potential problems in real-time as well as informing post-hoc analyses of customer-agent phone interactions. Typically, sentiment analysis models developed for call center applications are driven by post-hoc textual analysis of the resulting transcript [3], which today is often generated by an automatic speech recognition (ASR) engine. However, in a telephony setting, directly analyzing raw audio can prove advantageous and less costly than performing speech-to-text for subsequent lexical analyses. While traditional lexical approaches analyze the words spoken, acoustic models of sentiment analysis rely on acoustic feature extraction to determine the manner in which a message is conveyed (i.e. using valence, and other prosodic cues), moving away from the actual words used in the conversation. Results in recent literature [4], [5] show that relying on the joint use of linguistic and acoustic modalities in the prediction of sentiment outperforms traditional, purely text based methods. As an added benefit, this approach preserves privacy of the conversation as the audio stream is not subject to transcription and subsqeuent evaluation.

Despite such benefits, sentiment extraction from natural audio streams is challenging. The two primary challenges are first, there does not exist any large-scale, labeled, audio-only dataset of curated human sentiments that can be used for training in the linguistic community. The publicly available datasets traditionally used for audio sentiment detection — CMU-MOSI [6], RAVDESS [7], and IEMOCAP [8] — are multi-modal (audio, video, images) datasets more suited to emotion detection than for sentiment detection. Second, the

[1]Example courtesy Shahbaz Anwar, "Sentiment Analysis vs. Emotional Analysis: Same or Different," Sept. 2016, LinkedIn Blog, visited Jun 15, 2021. https://www.linkedin.com/pulse/sentiment-analysis-versus-emotional-same-different-shahbaz-anwar/

978-1-6654-2786-9/21/$31.00 ©2021 IEEE

labeling process itself is subjective as the same utterance that appears "positive" to one listener may be classified as "neutral" by another. Motivated by these challenges, we outline our work[2] that has the following contributions:

- We propose the use of a text-to-speech (TTS) engine to curate a *labeled* dataset of utterances. Each utterance is generated and labeled with one of the three sentiments: *negative, neutral,* or *positive*. Practically, such a TTS engine acts as a source to generate unlimited, labeled utterances in male or female voices (Section III).
- We examine the accuracy of a machine learning model trained on the corpus generated by a TTS engine and tested on a samples from a separate distribution that generated the TTS corpus (Section IV). Our approach controls for underspecification [9], an increasingly common problem with machine learning models.

II. Related Work

Compared to text-based sentiment detection, automatic sentiment detection from an audio stream remains a relatively underexplored area of research [10]. There is active work in the area of multi-modal sentiment analysis using text, audio, and video, see Soleymani et al. [11] for a representative survey, and recent works that uses recurrent neural networks [12] and attention models [13], [14] to mine sentiment from multi-modal data. However, requiring multi-modal data for sentiment analysis is not appropriate for many domains of interest. In call centers, for example, it is advantageous to perform sentiment analysis on the audio stream only, without even incurring the latency associated with transcribing the conversation for textual sentiment detection. In this domain, Li et al. [5] mine sentiments using both audio streams and the transcribed text corresponding to the audio stream. Our work, by contrast, is focused on extracting sentiment from only an audio stream without using any adjutant modes like text, video, or images.

Multi-modal sentiment analysis uses datasets curated primarily from YouTube videos or movies. For example, Agarwal et al. [12] perform multi-modal sentiment analysis on a multi-modal dataset called CMU-MOSI [6]. This dataset contains 2,199 utterances culled from 93 videos obtained from YouTube. Bhuiyan et al. [15] curate their data from YouTube videos; Alhujaili et al. [16] provide a comprehensive catalog of papers that uses YouTube to curate datasets.

The process of curating multi-modal datasets from YouTube is taxing as it involves viewing many videos to make a decision to include a subset of them for training and testing, writing crawlers to download the videos, and complex post-processing to annotate and label the multi-modal content appropriately. Consequently, the training set is usually small as curating the datasets manually does not scale easily and is a time consuming endeavor. Our approach on dataset curation, discussed in Section III, relies only on a single modality and is far more automated. Nonetheless, it results in the generation of an unlimited number of labeled samples stratified across male and female voices. Furthermore, our approach avoids copyright and privacy issues associated with downloading content from platforms such as YouTube. Even though the content on YouTube may be public, we remain conservative and assume the purpose limitation principle, i.e., the content owner may not consider it acceptable to use their likeness or sound for anything other than viewing the content on the platform on which it (the content) was uploaded.

Finally, we note that our result on using a simpler model (logistic regression) trained on the a dataset generated from a TTS engine is comparable in terms of accuracy to the literature we reviewed [5], [12]–[15], which contains more complex models (recurrent neural networks and attention-based models) operating across a multi-modal dataset to detect sentiment.

III. TTS Output As Labeled Sentiment Data

A. Training Data Generation

The data that will serve as training samples for sentiment determination is based on samples produced programmatically by a TTS engine using a RESTful service [17]. This configuration allows users to generate infinitely many training examples rich in variable vocabulary and sentence structure and more importantly, objectively labeled. In this study, the samples are generated using Voicery, a TTS provider[3]. Voicery provides an API that can be used to generate the audio samples in .wav format from any given text input. The platform provides options for both male and female voices spoken with eight different accents. Voicery also provides twelve emotions in which the samples can be rendered. We use only three sentiments for this study: Positive, Neutral, and Negative. Each were rendered in both male and female voices in an American accent. In total, we generated 4,860 .wav files, 50% of which are male and 50% of which are female, with an even distribution between Positive, Negative, and Neutral. Programmatic access through well-defined application programming interfaces allows the generation of unlimited training data appropriately labeled in one of the three sentiments. This approach lowers the barrier tremendously to obtaining quality labeled data for training sentiment analysis systems.

B. Test Data

To test the performance of a model induced on the training data as described above, we created two datasets. The first test dataset was drawn from the same Voicery distribution as described in Section III-A. The second dataset is an independent, out-of-sample human speech dataset consisting of acted speech, called The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [7]. The second test dataset is an important check on the performance of the model as it controls for underspecification [9]. Briefly, standard machine learning pipelines are characterized by a model induced from a training dataset, and an independent and identically distributed (iid) evaluation dataset that is drawn from the same

[2]Vāc is the Sanskrit word for speech.

[3]https://www.voicery.com/. Voicery ceased operations in October 2020; however, links to datasets to replicate this work are provided in Section III-C.

distribution as the training dataset. At times, such a pipeline leads to the model learning spurious correlations in the data that do not show up in the test data prevalent in the deployment domain; this is due to differences in causal structure between training and deployment domains. Underspecification, thus, impairs the generalization ability of the induced model. To ensure against underspecification, we use a separate, iid test dataset from the RAVDESS repository.

The RAVDESS repository contains 24 professional actors (12 male, 12 female) vocalizing two lexically matched statements in a North American Accent with 7 emotions, each with two intensity levels and two repetitions. For this study, we retain only the strong angry, neutral and strong happy emotions to represent negative, neutral and positive sentiments, respectively. This results in a test set consisting of 96 samples for each sentiment class, 288 samples in total. The distribution of sentiments across both datasets is summarized in Table I.

Audio Source	Negative	Neutral	Positive	Total Count
Voicery TTS	1620	1620	1620	4860
RAVDESS	96	96	96	288

TABLE I
DISTRIBUTION OF POSITIVE, NEGATIVE, AND NEUTRAL SAMPLES ACROSS BOTH DATASETS.

C. Feature Extraction

Modeling sentiment directly from an audio stream requires first extracting speech features relevant to the detection of sentiment. While humans are quite adept at recognizing sentiment based on acoustic properties of the speech signal, modeling sentiment directly from the acoustic features produced is nontrivial. In recent years, largely due to the drawbacks of sentiment analysis drawn solely from text, there has been a push to perform such analyses directly on the signal itself [18]. This process, while complex, can reduce latency and exploit the rich, prosodic information carried in the signal itself.

To perform the feature extraction, we used openSMILE [19], a modular and flexible audio analysis toolkit used frequently for feature extraction geared at sentiment analysis. We used one of its many available configurations, the IS13_comPaRE.conf, to extract features from the audio files generated by Voicery and RAVDESS. This configuration provides a well-developed set for automatic recognition of paralinguistic phenomena and is the official set of the 2013 INTERSPEECH Computational Paralinguistics Challenge (ComParE) [20]. This feature set promotes reproducibility and is a recognized standard in speech emotion recognition. Our datasets consisted of 7,573 dimensions for each observation (audio file). The OpenSMILE configuration consisted of 77 low-level audio descriptors and their first-order derivatives; several statistical parameters such as quartiles, range, mean value of the peaks, etc. In all, the base features and the derived features added to 7,573 dimensions.

The Voicery training dataset and the RAVDESS test dataset are available at the following URLs:

1) Voicery training dataset is available at http://www.cs.iit.edu/~vgurbani/datasets/sped/Voicery_4860.csv.gz.
   SHA-1: 94c01a7af05b647512f53121770194d5295719c7
2) RAVDESS testing dataset is available at http://www.cs.iit.edu/~vgurbani/datasets/sped/RAVDESS_288.csv.gz.
   SHA-1: b9122cd4964a570702a649ff15120c4a61e05b6f

IV. SENTIMENT ANALYSIS EXPERIMENTS

A. Model

Multinomial Logistic Regression (MLR) model is used in this study for sentiment classification, producing an output of positive, neutral or negative sentiment for an input vector of audio features. We used other more complex learning algorithms, including neural networks, but they produced identical results on the dataset we used. Following the principle of Occam's Razor, we preferred the simpler model.

With 7,573 features extracted from each audio sample, Principal Component Analysis (PCA) is performed first for dimensionality reduction that lead to top 1,174 principal component loadings explaining 95% of the total variance. Using this PCA variance ratio, the modeling result is very similar to that without PCA, with less than 0.5% difference in model prediction metrics including precision, recall and balanced accuracy. All experimental results presented in subsequent sections are obtained using the 95% PCA variance ratio.

B. Model Trained with TTS Samples

In this experiment, the TTS dataset of 4,680 samples generated from Voicery is used for training and testing an MLR model. Each sample consists of 7,573 features extracted from an audio clip for a specified sentiment as descried in Section III. The dataset is split into 80% training and 20% testing sets. After performing PCA, the top 1,174 principal components of each sample are retained for model training and testing. Resulting confusion matrix and model performance metrics for the test set are shown in Table II.

		Predicted		
		Negative	Neutral	Positive
True	Negative	319	5	6
	Neutral	10	301	0
	Positive	3	0	328

Sentiment	Negative	Neutral	Positive
Precision	0.961	0.984	0.982
Recall	0.967	0.968	0.991
Balanced Accuracy	0.973	0.980	0.991
Overall Accuracy	0.975		

TABLE II
TTS TEST SET CONFUSION MATRIX AND PERFORMANCE METRICS

The high accuracy in the test result indicates that the features extracted from TTS-generated audio clips contain sufficient information for sentiment analysis, and the MLR model is able to classify the sentiment using these features as input. The model is further evaluated using stratified k-fold cross validation, with $k = 10$. The mean and standard deviation of performance metrics are listed in Table III, which

shows the consistency and low variance in model prediction results when then MLR model is applied to the TTS dataset.

Sentiment		Negative	Neutral	Positive
Precision	Mean	0.971	0.983	0.985
	Std.Dev	0.010	0.010	0.007
Recall	Mean	0.975	0.981	0.982
	Std.Dev	0.010	0.009	0.004
Balanced Accuracy	Mean	0.980	0.987	0.987
	Std.Dev	0.006	0.005	0.003

TABLE III
K-FOLD CROSS VALIDATION OF MLR MODEL ON TTS DATASET

It should be noted the accuracies shown in Table II are significantly higher than those of other sentiment analysis models with acoustic feature inputs [21], [22]. This is not surprising because compared to real-world audio recordings, a sample generated by a TTS engine contains minimal noise and carries a reliable ground truth label. There is also the possibility that some sentiment-specific artifacts are introduced by the TTS engine in its audio output and subsequently captured in the extracted features. If a model has been trained with a dataset containing these artifacts, then testing the model on in-domain samples, i.e. audio clips generated by the same TTS engine, will yield an artificially high accuracy. To guard against such underspecification in model training, we conduct further experiments with the induced model on a separate iid test dataset as discussed next.

C. Model Evaluation on Separate IID Samples: Multi-class Prediction Model

This experiment evaluates the sentiment classification model trained with TTS dataset on real-world human voices using a set of test samples selected from RAVDESS. As described in Section III, we use a test set consisting of 96 samples from each of positive, neutral and negative sentiment classes, 288 samples in total. An MLR model is trained using all 4,680 samples in TTS dataset, then tested on the RAVDESS set. Table IV lists the test result.

		Predicted		
		Negative	Neutral	Positive
True	Negative	76	4	16
	Neutral	0	23	73
	Positive	10	6	80

Sentiment	Negative	Neutral	Positive
Precision	0.884	0.697	0.473
Recall	0.792	0.240	0.833
Balanced Accuracy	0.870	0.594	0.685
Overall Accuracy	0.622		

TABLE IV
RAVDESS TEST SET CONFUSION MATRIX AND PERFORMANCE METRICS

The balanced accuracies of 87.0% and 68.5% for negative and positive sentiments respectively are particularly interesting and demonstrate a reasonable efficacy in sentiment classification. A particular domain where such a model can be used is customer service in call-center operations where customers can converse with conversational artificial intelligent agents. In such an operation, it is imperative to automatically sense the frustration of a customer so he or she can be transitioned to talk to a live human agent. For this purpose, detecting the negative sentiment is more important than detecting the positive one. The results of Table IV demonstrate that the model induced from TTS dataset is able to capture the negative sentiment with increased accuracy compared to the positive sentiment. Despite a suppressed overall accuracy due to mis-classifying a number of neutral samples as positive, this result supports the approach of using TTS-generated synthetic audio samples to train a sentiment analysis model and applying it to real world audio recordings. (For completeness, in call-center operations a neutral sentiment serves to keep the customer engaged with the conversational artificial intelligent agent.)

D. Model Evaluation on Separate IID Samples: Binary Prediction Model

Due to the importance of detecting the negative sentiment in call-center operations, this experiment uses the same training and testing datasets as Section IV-C but induces a binary model, i.e., classify each test observation as negative or non-negative (the non-negative class consists of observations classified as positive or neutral). Result of the binary sentiment model is shown in Table V.

		Predicted	
		Negative	Non-Negative
True	Negative	72	24
	Non-Negative	8	184

Sentiment	Negative	Non-Negative
Precision	0.900	0.885
Recall	0.750	0.958
Balanced Accuracy	0.854	0.854
Overall Accuracy	0.889	

TABLE V
RAVDESS TEST SET BINARY CLASSIFICATION

The overall accuracy and precision of the binary sentiment model (88.9% and 90.0%, respectively) is higher than their counterparts in the multi-class prediction model (62.2% overall accuracy and 88.4% precision for the negative class, cf. Table IV). This result resembles those from similar studies using more complex multi-nodal sentiment analysis methods [21], and further validates the potential of our hypothesis on using synthetic, labeled samples generated from a TTS engine to fit sentiment generated by humans.

V. DISCUSSION

Comparing Tables II and IV, there is a noticeable drop in prediction accuracies when applying a sentiment analysis model trained on TTS dataset to a separate iid test dataset with human voice samples. There are several possible contributing factors:

1) *Data Quality*
   The dataset curated from a TTS engine has an advantage in data quality, including negligible noise in the audio and consistency in ground truth sentiment labels.

Even though the RAVDESS dataset is produced in a studio, the noise level of that controlled environment is still no match to the synthesized audio generated by a TTS engine. Moreover, consistent sentiment labeling is notoriously difficult due to its subjective nature. For example, different RAVDESS performers may interpret and vocalize an emotion differently, and this difference may cause the model to missclassify a sentiment. Nonetheless, as Table V demonstrates, there appears to be enough of a signal in the RAVDESS test observations that is appropriately picked up by a model induced on a dataset generated by a TTS engine.

2) *Limitation of Synthetic Speech*
Even though a state of the art TTS engine can synthesize speech that seems indistinguishable from a human voice recording, the feature-rich human voice is still not easily reproducible; certain prosodic cues associated with human speech may be absent from the synthetic TTS dataset. As a result a sentiment analysis model trained on TTS samples is limited to abstracting information that can be produced by the TTS engine only. This limitation can be mitigated in two ways: one, by restricting the response class depending on the specific domain the model is being used in, as we demonstrate in Section IV-D; and two, by ensuring that the induced model is generalizable and does not suffer from underspecification by testing it on a separate iid test dataset (as we did with the human-generated RAVDESS dataset).
3) *Potential TTS-induced Artifacts*
There is a chance that the Voicery TTS engine may latch on to sentiment-specific artificial signal in the generated audio files; such artifacts may subsequently propagate through extracted audio features and influence model training. To guard against such an eventuality, we demonstrate that the TTS induced model is robust and generalizes beyond test samples drawn from the same distribution as the one that contributed to the training samples. To do so, we used a separate iid human-generated RAVDESS test dataset. The generalization capability of the TTS induced model is evident in Tables IV and V.

Finally, for the sake of completion we note that RAVDESS dataset is produced by performers in controlled studio environment. For audio recordings from real world applications such as call center conversations, where the conversation may sometimes occur on noisy or lossy channels, the effects of above factors may be more pronounced. Therefore some difference in model prediction accuracy is expected if the separate iid test data to fit the model is obtained from such channels.

VI. Conclusion

High-quality datasets required to train models for automatic sentiment analysis can prove both costly and unreliable. These limitations can restrict a model's ability to generalize. In this paper, we contribute to the growing body of literature aimed at sentiment prediction drawn from acoustic data. We propose a method that allows for the generation of infinitely many objectively labeled samples, through the use of TTS renderings – in our case, provided by Voicery. We perform feature extraction on the salient acoustic features for the purpose of sentiment classification using a Multinomial Logistic Regression Model after performing Principal Component Analysis for dimensionality reduction.

The experiments presented demonstrate that our method of curating unlimited labeled data through a TTS engine shows great promise. Our model's high performance when tested on TTS samples generated by the same distribution as the one that generated training samples is not surprising. The more interesting result lies in a balanced accuracy of 87.0% and 68.2% for negative and positive sentiments respectively, when tested on sentiment data produced by humans using RAVDESS, a separate iid test dataset. Further, in a multi-label classification, the precision of 88.4% for the negative class is particularly noteworthy – as callers displaying intensely negative sentiment may require special action. We can further improve on this precision using a binary classification scheme, which demonstrates increased precision of 90.0%. Our result using a simple MLR trained on TTS generated data is comparable to accuracy in the literature reviewed in Section II, which contain more complex models as well as a more complex training data collection process.

Finally, we note that while there exist synthetic speech systems rated as having near human naturalness, the feature-rich human voice is not easily synthesized and there are some features of human speech not easily represented by a TTS engine. As a result, training on a TTS generated dataset absent of such features leads to some limitations we discussed in Section V. Another important caveat is that the RAVDESS samples, while produced by humans, is not entirely representative of naturally produced, spontaneous speech, as would be the case in a telephony setting. Acted speech, much like synthesized speech, contains many residual artifacts not present in naturally produced speech. We recognize that employing a model trained and tested on synthetic and acted speech can introduce some noise not present in other, naturally produced human speech datasets. Nonetheless, this work demonstrates a promising avenue of research in curating labeled training data from a TTS engine across multiple sentiments, and using these to induce a sentiment model with good generalization capabilities as demonstrated by testing the model on a separate iid test dataset.

VII. Future Work

We note that a logistic regression model trained and tested on data generated by Voicery achieves excellent performance; this however is expected as the training and testing datasets are fairly homogeneous because they are drawn from the same distribution. When the model trained on synthetic TTS engine is fitted to a separate iid test dataset, we observe a small drop in performance, which is both expected and tolerable as the accuracy and precision of the class of interest (the negative

class) is high, 88.9% and 90.9% respectively (cf. Table V). We plan to further investigate the drop to determine if it can be attributed artifacts of the TTS voices that do not exist in human speech, or some other factors. Without explicitly identifying and removing the artifacts, it is still possible to reduce their impact by adding some non-TTS samples to the training set, so that the training loss incurs a penalty when the model relies on the artifact signal for sentiment classification. While some exploratory work appears promising, the RAVDESS dataset used in this study is too small for a comprehensive experiment to validate it.

As we collated the results from the set of experiments described in this paper, it appeared that certain features show more promise in predicting a particular sentiment than others. We plan to follow up on this by discriminating the selection of features that yield an increased prediction accuracy for a particular class of interest. Finally, we expect that ongoing research on TTS engines trained using neural methods will likely yield better prosodies; we plan to investigate such TTS engines to determine their viability as generators of unlimited, labeled voice samples.

REFERENCES

[1] B. Liu, *Sentiment analysis: mining opinions, sentiments, and emotions*. The Cambridge University Press, 2015.

[2] M. Munezero *et al.*, "Are they different? affect, feeling, emotion, sentiment, and opinion detection in text," *IEEE Transactions on Affective Computing*, vol. 5, no. 2, pp. 101–111, 2014.

[3] S. Ezzat *et al.*, "Sentiment analysis of call centre audio conversations using text classification," *Intl. Journal of Computer Information Systems and Industrial Management Applications*, vol. 4, pp. 619–627, Dec 2012.

[4] Z. Luo, H. Xu, and F. Chen, "Audio sentiment analysis by heterogeneous signal features learned from utterance-based parallel neural network," in *AffCon@ AAAI*, 2019.

[5] B. Li *et al.*, "Acoustic and lexical sentiment analysis for customer service calls," in *ICASSP 2019-2019 IEEE Intl. Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5876–5880.

[6] A. Zadeh *et al.*, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[7] S. Livingston 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," in *PLoS One*, vol. 13, no. 5, 2018.

[8] C. Busso *et al.*, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[9] A. D'Amour *et al.*, "Underspecification presents challenges for credibility in modern machine learning," *CoRR*, vol. abs/2011.03395, 2020. [Online]. Available: https://arxiv.org/abs/2011.03395

[10] L. Kaushik *et al.*, "Automatic sentiment detection in naturalistic audio," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 8, pp. 1668–1679, 2017.

[11] M. Soleymani *et al.*, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing.

[12] A. Agarwal *et al.*, "Multimodal sentiment analysis via RNN variants," in *2019 IEEE Intl. Conference on Big Data, Cloud Computing, Data Science Engineering (BCD)*, 2019, pp. 19–23.

[13] T. Kim and B. Lee, "Multi-attention multimodal sentiment analysis," in *Proceedings of the 2020 Intl. Conference on Multimedia Retrieval*, 2020, pp. 436–441.

[14] Q.-T. Truong and H. W. Lauw, "Vistanet: Visual aspect attention network for multimodal sentiment analysis," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, 2019, pp. 305–312.

[15] H. Bhuiyan *et al.*, "Retrieving youtube video by sentiment analysis on user comment," in *2017 IEEE Intl. Conference on Signal and Image Processing Applications (ICSIPA)*. IEEE, 2017, pp. 474–478.

[16] R. Alhujaili *et al.*, "Sentiment analysis for youtube videos with user comments: Review," in *2021 Intl. Conference on Artificial Intelligence and Smart Systems (ICAIS)*, 2021, pp. 814–820.

[17] L. Richardson and S. Ruby, *RESTful web services*. " O'Reilly Media, Inc.", 2008.

[18] M. G. Huddar, S. S. Sannakki, and V. S. Rajpurohit, "A survey of computational approaches and challenges in multimodal sentiment analysis," *Int J Comput Sci Eng*, vol. 7, no. 1, pp. 876–883, 2019.

[19] 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*, 2010, pp. 1459–1462.

[20] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi *et al.*, "The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism," in *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*, 2013.

[21] Y. Jia, "A deep learning system for sentiment analysis of service calls," in *Proceedings of The 3rd Workshop on e-Commerce and NLP*. Seattle, WA, USA: Association for Computational Linguistics, Jul. 2020, pp. 24–34. [Online]. Available: https://www.aclweb.org/anthology/2020.ecnlp-1.4

[22] N. Sato and Y. Obuchi, "Emotion recognition using mel-frequency cepstral coefficients," *Information and Media Technologies*, vol. 2, no. 3, pp. 835–848, 2007.