Text vs. Speech Analysis – Detecting Sentiment of Customer Calls

Manuel Weber
WHU - Otto Beisheim School of Management
Christian Schlereth
WHU - Otto Beisheim School of Management

Cite as:
Weber Manuel, Schlereth Christian (2023), Text vs. Speech Analysis – Detecting Sentiment of Customer Calls. *Proceedings of the European Marketing Academy*, 52nd, (114205)

Paper from the 52nd Annual EMAC Conference, Odense/Denmark, May 23-26, 2023

Text vs. Speech Analysis – Detecting Sentiment of Customer Calls

Abstract:

In recent years, sentiment analysis has been adopted in customer service to better address customer needs during call center calls. We aim to learn about the value of text-based vs. speech-based sentiment analysis to understand customer satisfaction with a call center call. We apply a pre-trained transformer model to classify customer sentiment of mock call transcripts. Then, we train a convolutional neural network on a public speech dataset and use it to classify the sentiment of call recordings. We find that the "simple" text-based model is more accurate in predicting customer sentiment than the speech-based model (~84% vs. ~53%). While the further predicts sentiment best using the entire call transcripts, the latter predicts sentiment best using just the beginning of call recordings. For some calls, the speech-based model detects sentiment more accurately, which indicates that both approaches could complement each other.

Keywords: Sentiment Analysis, Machine Learning, Natural Language Processing

Track: Methods, Modelling & Marketing Analytics

1. Introduction

In the past, companies heavily relied on surveys to collect customer feedback and identify customer satisfaction (CSAT). Nowadays, the usage of CSAT surveys is still the standard for most companies, even though they are costly, and the results may be biased toward dissatisfied respondents participating (Qualtrics, 2022). Surveys are also used in customer service. After an interaction between a customer and a customer service agent, hereafter referred to as the agent, has taken place, sending out a survey to the customer asking for feedback on the interaction is common practice. However, most customers do not respond to these surveys (Chung, 2022), which makes it impossible to measure CSAT. Besides, some customers might not even indicate their true satisfaction, and it also takes time until customers complete such surveys. Hence, there is a need for a modern CSAT measurement approach.

At the same time, natural language processing has been increasingly adopted in customer service, e.g., chatbots. Multiple text and speech analysis approaches have been developed, which can be used to measure customer sentiment (CSENT) and emotions during a customer-agent interaction, e.g., a customer call. These approaches, which are mainly based on machine learning (ML) and require customer call recordings or transcripts of these recordings as input data, solve most of the problems outlined above: 1.) CSENT can be measured for any customer, 2.) its computations are objective, and 3.) it can be computed on a real-time basis, which could even signal agents to change their behavior during a call. However, CSENT conceptually differs from CSAT and would only serve as one component in automated CSAT detection. Other components would include historical customer data. To detect CSENT in a customer call, approaches based on speech are far more complex than those based on text. Thus, we want to understand if it is worth applying more complex, speech-based approaches or if text-based approaches are more suitable for detecting the CSENT of customer calls.

For real-time CSENT detection during a call, companies would need to build a prediction model on a shorter slice of a call. Call center operators may also wonder if they need to analyze the entire call or portions, which might lead to a loss in prediction quality. Therefore, we apply the thin-slice methodology (see Ambady et al., 2006) to analyze if the prediction accuracy suffers from the corresponding information loss. It might be the case that the ending of a call is more relevant for CSENT prediction, as it includes the outcomes of a call. In contrast, the beginning of a call might be more insightful as the customer could already enter the call with a negative or positive attitude toward the company. For instance, Hall et al. (2014) find that CSAT predictions based on transcripts of the first 120 seconds of a call

outperform predictions based on the entire transcripts. However, the authors used human raters instead of ML to classify CSAT. Consequently, we raise two essential research questions about the design of CSENT detection mechanisms. First, we aim to identify which ML approach performs better in detecting CSENT of customer calls, one based on text or speech analysis. Second, we slice each call to understand if accuracy improves or suffers as compared to if the entire call is analyzed.

Our research contributes to the existing literature. While previous literature applied either text analysis or speech analysis for CSENT detection, we apply both approaches and compare model performances. We also contribute to the literature stream of multimodal sentiment analysis by applying it not only to text but also to speech data, although the latter has been mainly used for emotion detection (e.g., Abbaschian et al., 2021; Poria et al., 2015; Shaw et al., 2016). Moreover, no other studies applied the thin-slice methodology to investigate if CSENT can be detected by ML applications (cf., Hall et al., 2014).

2. Literature Review

Previous research primarily applied text-based approaches to measure the sentiment of customer interactions. Balducci and Marinova (2018) and Berger et al. (2020) provide overviews of different text analysis applications in marketing. Text-based sentiment analysis is done using ML approaches or dictionary-based approaches (see Hartmann et al., 2019; Kübler et al., 2020). For ML approaches, the raw text data, e.g., call transcripts, needs to be tokenized, i.e., transformed into numerical vectors that can be interpreted by a computer. Dictionary-based approaches can be directly applied to the raw text data. For instance, the negativity of a transcript could be determined by counting the number of negative words that can be found in a dictionary of negative words divided by the total number of words in the transcript. In contrast, ML approaches are not trained on word semantics but on "how words and word combinations … are tied together" (Kübler et al., 2020, p. 139). Most text-based ML sentiment detection studies have used naïve Bayes, support vector machines, or neural networks as classification algorithms (Kübler et al., 2020). However, also pre-trained transformer models exist, which can be used in the case of small sample sizes (e.g., BERT by Devlin et al., 2018; RoBERTa by Liu et al., 2019).

Speech analysis is still a developing research area in computer science and has not been applied in the marketing literature to measure specifically the CSENT of customer calls. While most speech analysis studies focused on speech emotion detection (see Abbaschian et al., 2021), some studies have applied speech-based models and multimodal models to detect

sentiment in YouTube (YT) videos (e.g., Pereira et al., 2016; Perez Rosas et al., 2013; Poria et al., 2016). We are unaware of any pre-trained speech-based sentiment detection models. In speech analysis, abstract speech features, such as loudness, speed, or voice quality, must be extracted from the raw audio files and converted into numerical values. Transforming the audio files into interpretable features for CSENT detection is far more complex than transforming transcripts into tokenized vectors. As the semantics of the spoken language get lost in this process, one can only use ML-based approaches to analyze audio features. Hence, we only apply ML-based models in this research to enable a direct comparison between speech and text-based approaches. Most speech-based sentiment or emotion recognition studies have used neural networks as classification algorithms (see Abbaschian et al., 2021).

We expect that both approaches, text and speech analysis, can be used to predict the CSENT of customer calls and that sometimes text-based models and sometimes speech-based models perform better, as speech provides non-verbal cues that go over and beyond the pure word semantics. For instance, it also includes intonation or loudness, enabling inferences about a person's mood, social background, or personality. In contrast, text-based models are simpler. Additionally, we expect that the beginning of a call might be most insightful for CSENT detection (see Hall et al., 2014).

3. Research Design

3.1 Data

We collected publicly available mock customer call data to answer the research questions. Using the package *youtube_dl*, we scraped a list of 38 calls from five different YT channels of customer service coaches as audio files. Using the package *spleeter*, we separated speech from background noise and removed the latter. We aligned the sampling rate, sampling width, and the number of channels for each call and cut off parts from each audio that did not belong to the actual call (e.g., introduction). We identified speaker changes per call using the package *pyannote_audio* (Bredin et al., 2019), which includes a cutting-edge algorithm for speaker diarization. Based on the identified timestamps of speaker changes, we created a set of audio clips per call containing only the utterances[1] which belong to the customer. Speech analyses are typically done using data at the utterance level, as they perform better when trained on short audio files. To use the speech data in ML applications, we

[1] "An utterance is the vocalization (speaking) of a word or words that represent a single meaning to the computer. Utterances can be a single word, a few words, a sentence, or even multiple sentences" (Indian Institute of Technology Kanpur , 2022, para. 3).

extracted audio features from the customer utterances using the audio processing package *librosa*. Following Shaw et al. (2016), we normalized the volume range across all utterances and extracted and averaged 20 components of Mel-frequency cepstral coefficients (MFCCs) for each frame of an utterance. We had to reduce the length of each utterance to three seconds to get a consistent dimension for each audio vector. MFCCs are low-level, spectral audio features to detect emotions in speech (see Abbaschian et al., 2021; Poria et al., 2015).

Additionally, we scraped the automatically generated transcripts for each video from YT using the package *youtube_transcript_api* and removed irrelevant parts. Based on the identified timestamps of speaker changes, we created a set of transcripts per call containing only the utterances which belong to the customer and merged them into a single transcript containing all customer utterances per call. We applied further text cleansing steps, such as removing stop words, special characters, and punctuation, to the transcripts.

3.2 Methodology

We followed three analysis steps to predict CSENT in the set of mock calls: First, we labeled CSENT in each call to get the ground truth for our prediction algorithms. For that purpose, three researchers labeled CSENT in each call as "negative" or "positive," independent of each other. The three resulting labels per call were aggregated into a single label, either "negative" or "positive," if at least two judges picked the respective label. Thereby, 17 calls were labeled as "negative," and 21 calls as "positive."

Second, we created a text-based model to predict CSENT using the customer transcripts and a speech-based model to predict CSENT using the audio features of the customer utterances. Our dataset was too small to train an ML application, so we used a pre-trained text-based model. For that purpose, we relied on SiEBERT (Hartmann et al., 2022), which is fine-tuned for sentiment analysis but based on the general-purpose RoBERTa model. Note that RoBERTa has been trained on "five English-language corpora of varying sizes and domains, totaling over 160GB of uncompressed text" (Liu et al., 2019, p. 3), and SiEBERT has been fine-tuned on 15 additional, sentiment-labeled datasets with different types of texts (e.g., movie reviews, tweets, etc.). SiEBERT has been applied in several replication studies, resulting in a validation accuracy of over 93%, which outperforms most competitor models. The model automatically tokenizes input text.

For the speech analysis, we followed the procedure outlined in Puthran (2021) and trained a convolutional neural network (CNN) on the publicly available RAVDESS dataset (Livingstone & Russo, 2018), which contains ~1,500 audio recordings of 24 actors reading

short sentences in different emotions. We had to reduce the dataset to ~500 recordings containing only “angry” and “happy” utterances. We re-labeled them as “negative” and “positive” because sentiment was solely unambiguous for these two emotions. As the sentiment labels correspond with emotion labels, it makes sense to use MFCCs as input features to the speech-based model. The CNN consists of an input layer, four one-dimensional, convolutional layers, a one-dimensional pooling layer, four activation layers with *ReLU* as the activation function, and an output layer with *Softmax* as the loss function. The model was optimized using the *RMSprop* algorithm and trained for 700 epochs, using a batch size of 128. To train the CNN, we executed an 80-20 train-test split. Figure 1 illustrates the development of the model accuracy on the train and test datasets by epochs. After 700 epochs, the accuracy on the train dataset is ~87%, and the validation accuracy on the test dataset is ~76%, thereby performing similarly to other models which were trained on the RAVDESS dataset (cf., Abbaschian et al., 2021, pp. 20-22).

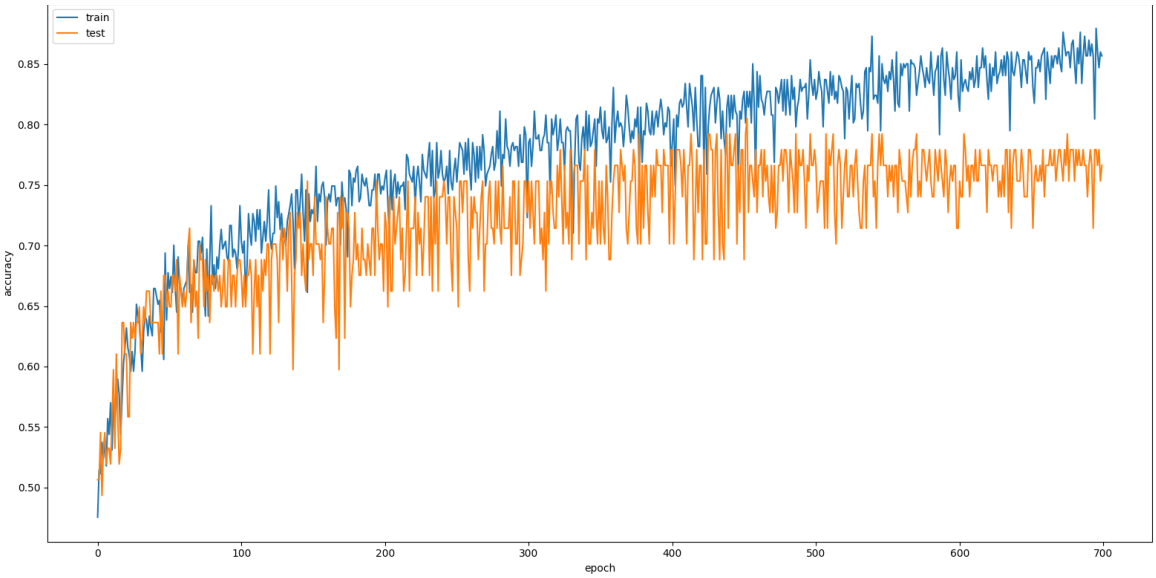

Figure 1. Accuracy of the Speech-Based CNN on the RAVDESS Train and Test Datasets

Third, we applied the text-based SiEBERT model and the speech-based CNN on the customer call dataset. While SiEBERT classified CSENT in each entire customer transcript, the CNN classified CSENT in each customer audio utterance. Hence, we had to aggregate the CSENT predictions of the individual utterances by call. If “negative” (“positive”) CSENT occurred most often in the utterance-level predictions, the entire call was classified as “negative” (“positive”). To answer the second research question, we split each call into thirds and only used the respective slices as input data for the models. We evaluated the performance of the models using a set of metrics that are based on the popular confusion matrix, revealing if the models are better at predicting “negative” or “positive” CSENT.

4. Results

4.1 Model performances

To evaluate which model performs best, we report the predictions of the classification models and aggregate these using confusion matrices and related metrics, such as accuracy, precision, and recall. We present the confusion matrices for the two models in Figure 2. While the SiEBERT model classifies CSENT in 32 of the 38 calls correctly (accuracy = 84.21%), the CNN only classifies 20 of the 38 calls correctly (accuracy = 52.63%), revealing a better performance of the text-based model in terms of accuracy. The confusion matrices also show that both models are remarkably accurate in correctly predicting calls with negative CSENT, with true negative rates of 88.24% and 82.35%, respectively. Moreover, the CNN overpredicts negative CSENT (29 predicted values out of 17 actual values), which results in a small recall of 28.57%. This might indicate that the audio features, which we extracted from the RAVDESS train dataset, were more distinctive toward predicting negative sentiment.

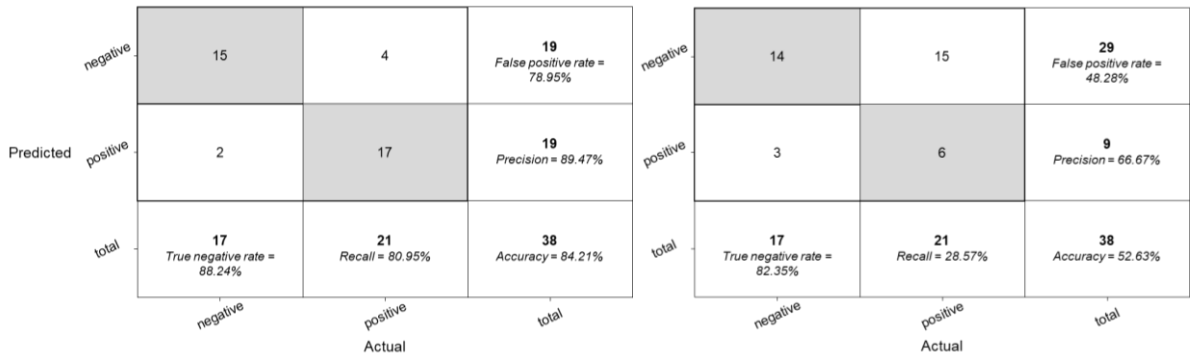

Figure 2. Confusion Matrices for the Results of the Text (left) and Speech Analysis (right)

Table 1 summarizes the prediction results compared to actual CSENT grouped by classification pattern. Except for one call (*Call ID:* 0), CSENT is correctly classified in all calls by at least one of the two models. While Table 1 again shows that SiEBERT performs better than the CNN, it also reveals that the CNN classifies some calls correctly, which are incorrectly classified by SiEBERT (*Call IDs*: 13, 21, 26, 27, and 37). Although for most calls, CSENT is more accurately classified by the text-based model, for some calls, the speech-based model prevails. Hence, the models complement each other.

Aggregated Call IDs	# of Hits (# of Misses)	Actual CSENT	Predicted CSENT	
			SiEBERT (Text Analysis)	*CNN (Speech Analysis)*
1, 6, 12, 14, 15, 20, 22, 23, 25, 28, 29, 31, 32, 34	14	**positive**	**positive**	negative
2, 5, 7, 8, 11, 16, 18, 19, 24, 30, 33, 35	12	**negative**	**negative**	**negative**
3, 4, 9	3	**negative**	**negative**	positive
10, 17, 36	3	**positive**	**positive**	**positive**
13, 21, 37	3	**positive**	negative	**positive**
26, 27	2	**negative**	positive	**negative**
0	0 (1)	positive	negative	negative
-	0	negative	positive	positive

Notes. We highlight all hits for which the CSENT predictions equal actual CSENT in bold with a grey background. We aggregate the calls by hit pattern (positive, positive, negative; negative, negative, negative; etc.) and count the number of correctly classified calls by at least one model, reported as the number of hits.

Table 1. CSENT Predictions of the Classification Models Compared to Actual CSENT

Finally, we provide an overview of the most important performance metrics of the models in Table 2. To answer the second research question, we compare the models' performances to those that only received slices as inputs to predict CSENT in the entire call. For the text analysis, the model which uses the entire transcript as input outperforms the models which use slices of the calls in almost all metrics (except recall). However, for the speech analysis, the model based on the first third of a call slightly outperforms all other speech-based models in all metrics. Although the performance improvement is not large, the results could indicate that, in speech analysis, it might already be sufficient to analyze the beginning of a call to predict CSENT in the entire call.

Model Type	*Accuracy*	*Precision*	*FPR*	*Recall*	*TNR*
SiEBERT (Text Analysis)					
Entire transcript (n=38)	**84.21%**	**89.47%**	**78.95%**	80.95%	**88.24%**
1st third of transcript (n=38)	63.16%	70.59%	57.14%	57.14%	70.59%
2nd third of transcript (n=35)	68.57%	78.57%	61.90%	57.89%	81.25%
3rd third of transcript (n=34)	67.65%	65.38%	75.00%	**89.47%**	40.00%
CNN (Speech Analysis)					
Entire recording (n=38)	52.63%	66.67%	48.28%	28.57%	**82.35%**
1st third of recording (n=38)	**55.26%**	**70.00%**	**50.00%**	**33.33%**	**82.35%**
2nd third of recording (n=34)	50.00%	55.56%	48.00%	27.78%	75.00%
3rd third of recording (n=38)	44.44%	46.15%	43.48%	31.58%	58.82%

Notes. FPR = False Positive Rate, *TNR* = True Negative Rate; The best-performing models per performance metric are highlighted in bold. We excluded calls that did not have any customer utterances in the 2nd or 3rd thirds of the calls from the respective analyses, creating a reduced number of observations.

Table 2. Performance Metrics of the Models

4.2 Discussion

In the following, we present the main implications of the model results and discuss them accordingly. The performance comparison of the models reveals that SiEBERT achieves high accuracy levels for CSENT detection, whereas the CNN only achieves mediocre accuracy levels. This observation might imply that more established, text-based algorithms better predict CSENT of customer calls. However, we should avoid deducing generalizations from the finding that, in our case, the text-based model outperforms the speech-based model, as

both models are pre-trained on different kinds of datasets. To improve the direct comparison of the two approaches and to check the robustness of our results, we would need to replicate the results of our study using other publicly available datasets, preferably conversational datasets. Furthermore, we would need to extract additional audio features as model inputs and test performance levels in detecting sentiment in conversations.

Our results indicate that speech-based models sometimes outperform text-based models in detecting CSENT. We speculate that this finding is based on the additional information conveyed by speech compared to text, and we highly encourage further research in this direction. It might be beneficial to develop models that use common text and speech analysis elements to detect sentiment ("multimodal fusion;" e.g., Pereira et al., 2016; Perez Rosas et al., 2013; Poria et al., 2016; Wöllmer et al., 2013). Nevertheless, the high complexity of speech-based models may pose a barrier to their application in practice. In contrast, text-based models can be easily applied using available transcription services to process the data and pre-trained sentiment analysis tools to analyze it.

Finally, the result that the CNN performs best using only data from the beginning of a call indicates that speech-based models require less information to classify CSENT correctly than text-based models. A conversation's tone is often already set at the beginning of a call, which could explain why the CNN based on the first third of a call performs best among the speech-based models. Hence, it might be sufficient for companies to use data from the beginning of a call in detecting the CSENT of the entire call, which would be a desirable research outcome for real-time sentiment detection.

5. Conclusion

Our research reveals that ML-based sentiment analysis can be effectively used to predict the CSENT of customer calls. While text-based models outperform speech-based models, we conclude that multimodal models could achieve even higher accuracy levels. Therefore, our subsequent research step will be to combine both approaches. Moreover, we are currently setting up a cooperation with a call agency to apply our proposed models to real-life customer call data. In addition, we are preparing a survey-based CSAT study to compare traditional CSAT measurement to ML-based measurement. Thereby, we aim to quantify the bias of a CSAT survey and evaluate if ML-based sentiment analysis is suitable not only for CSENT detection but also for CSAT detection.

6. References

Abbaschian, B. J., Sierra-Sosa, D., & Elmaghraby, A. (2021). Deep Learning Techniques for Speech Emotion Recognition, from Databases to Models. *Sensors (Basel, Switzerland)*, *21*(4).

Ambady, N., Krabbenhoft, M. A., & Hogan, D. (2006). The 30-Sec Sale: Using Thin-Slice Judgments to Evaluate Sales Effectiveness. *Journal of Consumer Psychology*, *16*(1), 4–13.

Balducci, B., & Marinova, D. (2018). Unstructured data in marketing. *Journal of the Academy of Marketing Science*, *46*(4), 557–590.

Berger, J., Humphreys, A., Ludwig, S., Moe, W. W., Netzer, O., & Schweidel, D. A. (2020). Uniting the Tribes: Using Text for Marketing Insight. *Journal of Marketing*, *84*(1), 1–25.

Bredin, H., Yin, R., Coria, J. M., Gelly, G., Korshunov, P., Lavechin, M., Fustes, D., Titeux, H., Bouaziz, W., & Gill, M.-P. (2019, November 4). *pyannote.audio: neural building blocks for speaker diarization*. Retrieved from http://arxiv.org/pdf/1911.01255v1. (Last accessed: November 2, 2022).

Chung, L. (2022). *What is a good survey response rate for online customer surveys?* Delighted. Retrieved from https://delighted.com/blog/average-survey-response-rate. (Last accessed: November 3, 2022).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018, October 11). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Retrieved from https://arxiv.org/pdf/1810.04805. (Last accessed: November 2, 2022).

Hall, J. A., Verghis, P., Stockton, W., & Goh, J. X. (2014). It Takes Just 120 Seconds: Predicting Satisfaction in Technical Support Calls. *Psychology & Marketing*, *31*(7), 500–508.

Hartmann, J., Heitmann, M., Siebert, C., & Schamp, C. (2022). More than a Feeling: Accuracy and Application of Sentiment Analysis. *International Journal of Research in Marketing.* Advance online publication.

Hartmann, J., Huppertz, J., Schamp, C., & Heitmann, M. (2019). Comparing automated text classification methods. *International Journal of Research in Marketing*, *36*(1), 20–38.

Indian Institute of Technology Kanpur. (2022, December 3). *Speech Recognition HOWTO*. Retrieved from https://www.iitk.ac.in/LDP/HOWTO/Speech-Recognition-HOWTO/introduction.html. (Last accessed: November 3, 2022).

Kübler, R. V., Colicev, A., & Pauwels, K. H. (2020). Social Media's Impact on the Consumer Mindset: When to Use Which Sentiment Extraction Tool? *Journal of Interactive Marketing, 50*(1), 136–155.

Liu, Y., Ott, M., Goyal, N., Du Jingfei, Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 26). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. Retrieved from https://arxiv.org/pdf/1907.11692. (Last accessed: November 3, 2022).

Livingstone, S. R., & Russo, F. A. (2018). 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*, *13*(5), e0196391.

Pereira, M. H. R., Pádua, F. L. C., Pereira, A. C. M., Benevenuto, F., & Dalip, D. H. (2016, April 10). *Fusing Audio, Textual and Visual Features for Sentiment Analysis of News Videos*. Retrieved from http://arxiv.org/pdf/1604.02612v1. (Last accessed: November 3, 2022).

Perez Rosas, V., Mihalcea, R., & Morency, L.-P. (2013). Multimodal Sentiment Analysis of Spanish Online Videos. *IEEE Intelligent Systems*, *28*(3), 38–45.

Poria, S., Cambria, E., Howard, N., Huang, G.-B., & Hussain, A. (2016). Fusing audio, visual and textual clues for sentiment analysis from multimodal content. *Neurocomputing*, *174*, 50–59.

Poria, S., Cambria, E., Hussain, A., & Huang, G.-B. (2015). Towards an intelligent framework for multimodal affective data analysis. *Neural Networks: The Official Journal of the International Neural Network Society*, *63*, 104–116.

Puthran, M. (2021). *Speech Emotion Analyzer.* GitHub. Retrieved from https://github.com/MiteshPuthran/Speech-Emotion-Analyzer. (Last accessed: November 3, 2022).

Qualtrics. (2022, May 9). *Survey Bias: Common Types of Bias and How to Avoid Them*. Retrieved from https://www.qualtrics.com/uk/experience-management/research/survey-bias. (Last accessed: November 2, 2022).

Shaw, A., Kumar, R., & Saxena, S. (2016). Emotion Recognition and Classification in Speech using Artificial Neural Networks. *International Journal of Computer Applications*, *145*(8), 5–9.

Wöllmer, M., Weninger, F., Knaup, T., Schuller, B., Sun, C., Sagae, K., & Morency, L.-P. (2013). YouTube Movie Reviews: Sentiment Analysis in an Audio-Visual Context. *IEEE Intelligent Systems*, *28*(3), 46–53.