Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 6549–6555
Marseille, 11–16 May 2020

A Large Scale Speech Sentiment Corpus

Eric Y. Chen[1], Zhiyun Lu[2]*, Hao Xu[1], Liangliang Cao[1], Yu Zhang[1], James Fan[1]
[1]Google, Inc.,
New York, NY
{erchen, hxx, llcao, ngyuzh, jjfan}@google.com

[2]University of Southern California
Los Angeles, CA
zhiyunlu@usc.edu

Abstract

We present a multimodal corpus for sentiment analysis based on the existing Switchboard-1 Telephone Speech Corpus released by the Linguistic Data Consortium. This corpus extends the Switchboard-1 Telephone Speech Corpus by adding sentiment labels from 3 different human annotators for every transcript segment. Each sentiment label can be one of three options: positive, negative, and neutral. Annotators are recruited using Google Cloud's data labeling service and the labeling task was conducted over the internet. The corpus contains a total of 49500 labeled utterances covering 140 hours of audio. To the best of our knowledge, this is the largest multimodal Corpus for sentiment analysis that includes both speech and text features.

Keywords: sentiment, switchboard, multimodal, speech

1. Introduction

Sentiment analysis is the task of recognizing the sentiment of a given input. While most of the previous works done in sentiment analysis are using text based inputs, voice inputs are becoming more important as the adoption of voice based user agents, such as smart assistants and mobile voice control becomes more prevalent. In this paper we present the first large scale publicly available speech sentiment corpus, based on the Linguistic Data Consortium (LDC) Switchboard-1 Corpus. For the remainder of this paper, we shall refer to the original Switchboard corpus as *Switchboard* and our extension as *Switchboard Sentiment*.

Each audio file in Switchboard is first split into segments based on the start and end time of the officially released transcript turns. These segments are henceforth referred to as utterances. We extend Switchboard by annotating each utterance with at least three human workers. Each worker may assign one of three sentiment labels to each utterance: *positive*, *neutral*, or *negative*. Workers are recruited over the internet using Google Cloud's data labeling service (Google, 2019). The data labeling task involves playing the audio corresponding to every utterance and ask workers to independently assign the most likely sentiment label. Every worker's annotation is recorded and released with this corpus.

Conflicting sentiment labels are a natural occurrence. We propose using a simple majority voting scheme to select the most probably sentiment label as the ground-truth. Based on this approach, the corpus has 30.4% positive utterances, 17% negative utterances, and 52.6% neutral utterances. Using the highest voted sentiment label as ground-truth, we measure the average accuracy human workers at assigning the correct (highest voted) sentiment label as 85%. This number is lower than many of the other speech sentiment corpora. We suspect the reason is because most speech sentiment corpora either force participants to act out emotions following certain scripts (Busso et al., 2008), or create scenarios to incite strong emotional reaction from participants (Ringeval et al., 2013). Switchboard on the other hand, was not designed to elicit intense emotional response.

The rest of this paper is outlined as follows. Section 2. summarizes existing work and related corpora. Section 3. describes the Switchboard Sentiment corpus in detail, as well as its annotation guidelines and format. Section 4. analyses the corpus in detail, including sentiment label distribution and worker accuracy. Section 5. evaluates several state of the art sentiment prediction models using the Switchboard Sentiment corpus. Finally, Section 6. concludes.

2. Related work

Speech sentiment analysis is a well researched problem that involves assigning a discrete sentiment label (representing a human emotion) to a segment of speech. There have been two schools of work in the field of speech sentiment analysis. Single modality models make sentiment predictions are made using single modal features such as acoustic features (Li et al., 2019; Li et al., 2018; Wu et al., 2019; Xie et al., 2019), raw waveforms (Tzirakis et al., 2018), or transcript texts (Lakomkin et al., 2019). Multimodality models combines both acoustic and text features to make predictions to maximize the mutual information (Kim and Shin, 2019; Cho et al., 2018; Gu et al., 2018; Eskimez et al., 2018; Zhang et al., 2019). The Switchboard Sentiment corpus described in this paper is suitable for both single modal and multimodal research because it contains both acoustic and text features.

Text based sentiment corpora – Several text-only sentiment corpora are available; such as movie review data (Pang and Lee, 2004; Pang and Lee, 2005; Maas et al., 2011), product review data (Dredze and Blitzer, 2009), Tweets (Go et al., 2009), and paper review data (Keith et al., 2017). One major difference between the text portion

*The author performed the work while interned at Google.

of the Switchboard Sentiment corpus (if we only consider the transcripts) and these aforementioned text corpora is that Switchboard transcripts are conversational speech as opposed to written reviews.

Multimodal sentiment corpora – SEMAINE (McKeown et al., 2012) is a audiovisual database that contains conversations between human participants and artificial agents. IEMOCAP (Busso et al., 2008) is a similar audiovisual database but with actors performing scenarios selected to elicit emotional expressions. RECOLA (Ringeval et al., 2013) is a corpus of recordings where participants were asked to reach consensus on how to survive in a disaster scenario, and ask to self report their emotion using the positive and negative effective schedule. MOSEI (Zadeh and Pu, 2018) contains sentiment annotations of Youtube videos from more than 1000 speakers. ICT-MMMO (Zadeh et al., 2018) is a video review corpus annotated at video level for sentiment.

Aforementioned corpora suffer from several drawbacks. They are either scripted (Busso et al., 2008), small in volume (Ringeval et al., 2013; McKeown et al., 2012), single speaker monologues (Zadeh and Pu, 2018; Zadeh et al., 2018), or a combination of each.

The Switchboard Sentiment dataset contains free-form conversations that bare closer resemblance to natural dialogue. It is also the largest speech sentiment database to date. Table 1 provides a comparison between aforementioned speech sentiment corpora and Switchboard Sentiment.

3. The Corpus

The Switchboard corpus (Godfrey et al., 1992) is a well studied speech corpus composed of 2400 two-sided telephony conversations from 543 speakers around the US. Topics of conversation are selected by a computer-driven operator among 70 options. Both the topic and the two participants of a conversation are selected such that no two participants could converse more than once, and no participant can discuss the same topic more than once.

Each conversation in the Switchboard corpus can last up to 5 minutes; and every recording has 2 channels, one for each speaker. Our goal is to extend the Switchboard Corpus with sentiment labels. To make annotations easy and precise, each conversation is split into small segments. The granularity of segments is defined by the Switchboard's officially released transcripts. That is, we consider each transcript turn as a unit of speech that can be assigned a sentiment label. Sometimes, a speech segment could contain a short exchange between both speakers. We intentionally keep audio from both channels when playing it to annotators so the conversation's context is captured. However, one side effect of this is that there are cases where the two speakers have conflicting sentiments, an example of this is illustrated in Table 3b.

This section describes the process of assigning sentiment labels and the storage format of sentiment labels.

3.1. Annotation Guideline

Annotators are recruited using Google Cloud's data labeling service (Google, 2019). This section describes the guideline provided to the annotators.

3.1.1. Output Annotation

Annotators are asked to classify each conversation segment into one of the following categories:

1. **Positive** – One or both participants shows signs of happiness and positive attitude, such as laughing or smiling or using positive words. e.g., "this is great.", "Thank you (laughter)", "I'm pretty happy about it." Positive sentiments can be expressed in the form of:
   - encouragement,
   - joy,
   - or lists of positive traits/features.
2. **Negative** – One of both participants shows signs of negative emotions such as raising voice in anger, being dismissive or using negative words. e.g., "I hate it.", "(sigh) I don't know what to do about it." Negative sentiment is not necessarily directed at the other speaker but could be at the topic be discussed. Negative sentiment can be expressed in the form of:
   - disparagement,
   - doubts/seriously questioning/suspicious,
   - sarcasm,
   - anger,
   - or lists of negative traits/features.

   Note that disagreement itself can viewed as neutral as long as the speaker is calm and objective.
3. **Neutral** – No emotional or lexical cues to indicate the speakers sentiment one way or another. e.g., "Turn right after the first traffic light."

Table 2 includes some real examples of each sentiment label.

3.1.2. Annotation Task

The annotation task is two-fold. First, annotators are asked to produce a sentiment classification (from Section 3.1.1.. Second, if the classification is non-neutral (i.e., either positive or negative), the annotator is also asked to provide a justification which can be based on the flowchart steps below.

3.1.3. Annotation Flowchart

Here is the flowchart that annotators are asked to follow:

1. Does the segment contain clear emotional markings to indicate sentiment (e.g. laughter for positive, yelling for negative)?
   - **Yes** – Use the marking to annotate accordingly. Note that:
     (a) Sometimes people laugh to reduce the awkwardness of saying something negative, in such cases, mark it as neutral.
     (b) Sometimes people sneer (smile or laugh in a mocking tone), mark the sentiment as negative in that case.

Dataset	# of hours	# of classes	Type of speech
SEMAINE	6.5	27	conversations with non-human agents
RECOLA	3.8	5	natural human conversations
MOSEI	65	5	video monologues
ICT-MMMO	14	5	video monologues
IEMOCAP	11.5	4	scripted human conversations
SWBD Sentiment	140	3	natural human conversations

Table 1: A comparison of common speech sentiment corpora

Reasons	Transcript
Agreement	That's exactly right.
Positive belief	The metric system is kind of easy to me.
Happiness	That's great (laughter).

(a) Positive

Reasons	Transcript
Frustration	I don't know how do we end this thing.
Strong disagreement	I have strong objections to that.
Negative fact	Obviously it didn't work in California.

(b) Negative

Reasons	Transcript
Fact	Drive cars with catalytic converters and all that.
Neutral Disagreement	Not a large group of them but just a few.
Question	Doesn't Wisconsin have a state medicare program?

(c) Neutral

Table 2: Examples of sentiment labels

- **No** – See Step 2.

2. Is the segment an objective description of facts?
    - **Yes** – If it lists a number of positive facts/features (e.g. ample parking spots, more durable than industry standard), then it's positive. If it lists a number of negative facts/features then it's negative. Otherwise, it's neutral
    - **No** – See Step 3.
3. Does the segment express a preference?
    - **Yes** – If the subjective opinion/preference expresses a like or dislike or positive (e.g. it's great that ...), or negative, then annotate it accordingly.
    - **No** – It's neutral.
4. If the utterance is too short to determine the sentiment, mark it as neutral.

3.2. Corpus Format

The corpus is stored in a single tab-delimited CSV file. Each row is arranged in a fixed 4 columns format. The contents of these columns are described below.

1. An audio marker that maps directly to a Switchboard audio file name.
2. Start time in seconds of the audio recording that was played to annotators.
3. End time in seconds of the audio recording that was played to annotators.
4. Sentiment annotations from 3 annotators.

The fourth column contains a serialized string that represents the sentiment annotation. This string has the following grammar.

$$\begin{array}{rcl} \langle\text{label}\rangle & \models & \texttt{"Positive"} \mid \texttt{"Neutral"} \mid \\ & & \texttt{"Negative"} \\ \langle\text{reason}\rangle & \models & \texttt{[Any string]} \\ \langle\text{annotation}\rangle & \models & \langle\text{label}\rangle\ \texttt{"-\{"}\ \langle\text{reason}\rangle\ \texttt{"\}"} \\ \langle\text{annotation}\rangle & \models & \langle\text{annotation}\rangle\ \texttt{"\#"}\ \langle\text{annotation}\rangle \end{array}$$

4. Analysis

Sentiment is often subjected to the listener's own interpretation. Sometimes there is no clear cut sentiment label for a given utterance. To reduce bias from individual annotators, we choose to get 3 annotators to annotate each utterance. The set of annotators for different utterances may be different. When disagreement occurs, it's up to the user of this corpus to decide on how to reconcile disagreements. Here is a simple annotation reconciliation strategy based on majority voting:

- **3 way agreement** – For the case that all annotators agree on a single sentiment label, nothing needs to be done.
- **2 way agreement** – When the majority of annotators agree on a sentiment label, use it.
- **3 way disagreement** – When every annotator disagrees on the sentiment label, ignore this segment.

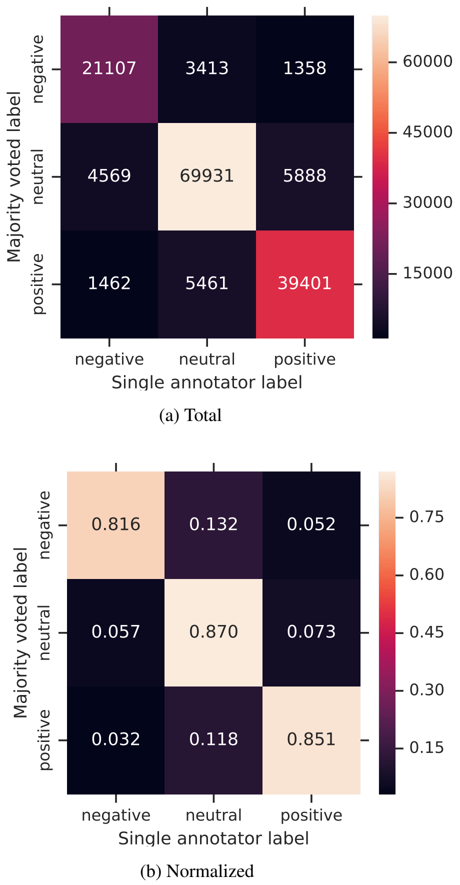
Table 3 illustrates some examples of inter-annotator disagreements. To better present emotional cues embedded in the audio component of these samples, we augment each transcript with our own interpretation of disagreements.

Using the above majority voting strategy, the distribution of sentiment labels is as follows – 30.4% of the speech segments are labelled as positive, 17% of the segments are labelled as negative, and 52.6% of the segments are labelled as neutral. For comparison, the IEMOCAP (Busso et al., 2008) corpus has 30.9% positive (happy), 49.5% negative (angry or sad), and 19.6% neutral. The intuition behind the larger representation of neutral sentiment labels in Switchboard Sentiment is as follows: because Switchboard participants are paid to converse on a given topic **not** of their choosing, it is less likely for the participants to have strong emotional attachments to their topic, and participants are less inclined to have extreme emotional responses such as heated arguments and intense joy. This lack of extreme emotional variance makes labels in the Switchboard Sentiment corpus harder to compare to other corpora. Section 5. includes some evaluations of existing sentiment analysis models using the Switchboard Sentiment corpus.

Figure 1 illustrates the distribution of inter-annotator disagreements. We only consider samples with a clear majority label (i.e., samples with 3-way disagreement are discarded). The vertical dimension represents sentiment of the most voted label, which can be interpreted as the ground truth. The horizontal dimension represents sentiment labels from individual annotators. The top-left to bottom-right diagonal of Figure 1b can be interpreted as the accuracy of human annotators. That is, the likelihood that any single human annotator can produce the same sentiment label voted for by the majority of annotators. The average human accuracy of the Switchboard Sentiment corpus is around 85%. For contrast, the IEMOCAP (Busso et al., 2008) corpus has a 91% human accuracy. This suggests that Switchboard Sentiment is a harder sentiment prediction dataset even for humans.

5. Evaluation

The Switchboard Sentiment corpus presented in this paper is already used to evaluate state of the art speech sentiment models (Lu et al., 2019). In this section, we refer closely to the evaluation results in (Lu et al., 2019), which is directly referenced in Table 4. This evaluation compares the Switchboard Sentiment corpus against the IEMOCAP (Busso et al., 2008) corpus. It measures the sentiment prediction accuracy of models trained using 3 types of features – acoustic features (mel-scaled spectrograms), acoustic and language features (mel-scaled spectrograms and transcripts), and embedded multimodal features taken from an intermediary layer of a pre-trained end-to-end ASR model (Graves, 2012; Rao et al., 2017).



(a) Total

(b) Normalized

Figure 1: Confusion matrix for inter-annotator agreement.

Due to the uneven distribution of sentiment labels, 2 accuracy metrics were computed – **Weighted Accuracy (WA)** represents the conventional classification accuracy (i.e., true positives over total), and **Unweighted Accuracy (UA)** represents the average accuracy of each sentiment class. Aside from Switchboard Sentiment being overall harder to predict, a notable observation from Table 4 is that speech signals in Switchboard Sentiment do not carry as much weight as they do in other corpora for speech sentiment prediction.

To further dissect model performance on Switchboard Sentiment, we present prediction results of three baseline sentiment analysis models as confusion matrices shown in Figure 2. Each model used in this evaluation consumes a different feature type:

1. **Using only acoustic features** – CNN with 3 layers of convolution filters with max pooling.
2. **Using only transcripts** – RNN with word embedding layer followed by 2 LSTM layers and 2 fully connected layers.
3. **Using multimodal features (acoustic features and transcripts)** – 1 LSTM layer followed by a multi-head self-attention layer (Vaswani et al., 2017).

Several observations can be made. First, the model trained using only acoustic features failed to learn from negative

Context	Transcript
The sentiment of “old” here (referring to a song) is open for interpretation.	It was really old.
Switching sentiment	I do think the jury system works, but I also feel . . .
Confusing tone	I should say on the west side, I mean everything is on the west side . . .

(a) 3-way disagreement

Context	Transcript
Slight doubt in response to a positive statement	Speaker 1: I think they will need me more when they are older (laughter). Speaker 2: Well (questioning tone).
Influenced by the religious orientation of annotators	Videos, like music videos that go along with songs about churches and Jesus . . .

(b) 2-way agreed positive

Context	Transcript
Switching sentiment	I really don’t like this stuff, but my husband does, he loves to cook it . . .
Ambiguous tone	Gee (grumbling tone), we have so much going on here . . .

(c) 2-way agreed negative

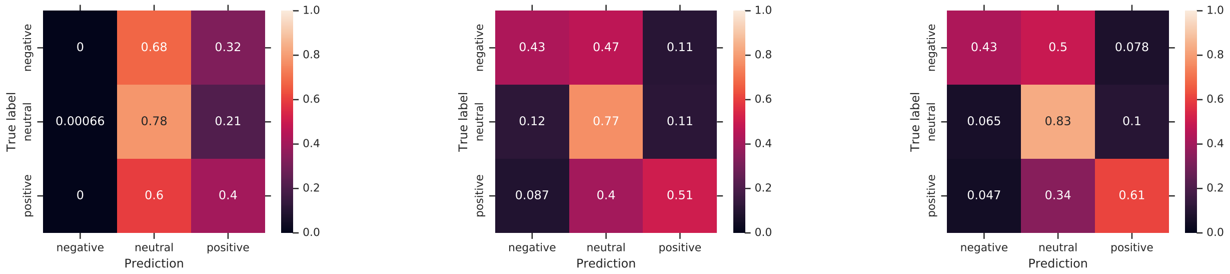
Context	Transcript
“Advantage” is a positive word.	He would have family close by, there are advantages . . .
Slightly positive tone	I imagine where you live, you wear warm clothing quite a bit of the year.

(d) 2-way agreed neutral

Table 3: Examples of disagreements between annotators

	IEMOCAP dataset			SWBD-senti dataset		
Input features	Architecture	WA (%)	UA (%)	Architecture	WA (%)	UA (%)
acoustic	DRN + Transformer (Li et al., 2019)	-	67.4	CNN	54.23	39.63
acoustic + text	DNN (Kim and Shin, 2019)	66.6	68.7	CNN and LSTM	65.65	54.59
e2e ASR	RNN w/ attention	71.7	72.6	RNN w/ attention	70.10	62.39
-	human	91.0	91.2	human	85.76	84.61

Table 4: Evaluation result directly taken from (Lu et al., 2019).



(a) Acoustic features

(b) Transcript features

(c) Multimodal features

Figure 2: Confusion matrix for evaluation using different input feature types.

examples. This seems to suggest that our baseline model could not effectively utilize acoustic signals to predict negative sentiment. Furthermore, comparing Figure 2b with Figure 2c, we can see that adding additional information on top of transcript features does not improve the accuracy of negative sentiment prediction (top left cell). Human annotators seem to have an easier time making such prediction (81.6% in Figure 1b comparing to 43% in Figure 2b and Figure 2c). One possible explanation behind this is that switchboard participants are paid to conduct free-form conversations and have less incentives to get into heated arguments or display strong negative emotions; and humans excel at detecting subtle hints of negativity. Effectively detecting negative sentiment in Switchboard Sentiment is an interesting problem that warrants future research.

6. Conclusion

We present Switchboard Sentiment, a large scale, multimodal speech sentiment corpus leveraging the existing Switchboard-1 Telephone Speech Corpus. Switchboard Sentiment is the largest multi-speaker conversational speech sentiment corpus to date with 49500 labels and 140 total hours. Unlike most existing speech sentiment corpora, Switchboard Sentiment participants were not explicitly asked to elicit strong emotional behavior, but rather encouraged to have natural conversations. This makes predictions on Switchboard Sentiment much harder compare to similar corpora. We believe Switchboard Sentiment can be used as a new benchmark for multimodal speech sentiment analysis.

7. Bibliographical References

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. (2008). Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335.

Cho, J., Pappagari, R., Kulkarni, P., Villalba, J., Carmiel, Y., and Dehak, N. (2018). Deep neural networks for emotion recognition combining audio and transcripts. In *Interspeech*, pages 247–251.

Eskimez, S. E., Duan, Z., and Heinzelman, W. (2018). Unsupervised learning approach to feature analysis for automatic speech emotion recognition. In *ICASSP*, pages 5099–5103. IEEE.

Google, (2019). *AI Platform Data Labeling Service*. Google Inc. https://cloud.google.com/data-labeling/docs/.

Graves, A. (2012). Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*.

Gu, Y., Yang, K., Fu, S., Chen, S., Li, X., and Marsic, I. (2018). Multimodal affective analysis using hierarchical attention strategy with word-level alignment. In *Proc. ACL (Volume 1: Long Papers)*, pages 2225–2235.

Kim, E. and Shin, J. W. (2019). Dnn-based emotion recognition based on bottleneck acoustic features and lexical features. In *ICASSP*, pages 6720–6724. IEEE.

Lakomkin, E., Zamani, M. A., Weber, C., Magg, S., and Wermter, S. (2019). Incorporating end-to-end speech recognition models for sentiment analysis. *arXiv preprint arXiv:1902.11245*.

Li, P., Song, Y., McLoughlin, I. V., Guo, W., and Dai, L. (2018). An attention pooling based representation learning method for speech emotion recognition. In *Interspeech*, pages 3087–3091.

Li, R., Wu, Z., Jia, J., Zhao, S., and Meng, H. (2019). Dilated residual network with multi-head self-attention for speech emotion recognition. In *ICASSP*, pages 6675–6679. IEEE.

Lu, Z., Cao, L., Zhang, Y., Chiu, C.-C., and Fan, J. (2019). Speech sentiment analysis via pre-trained features from end-to-end asr models. *arXiv preprint arXiv:1911.09762*.

Rao, K., Sak, H., and Prabhavalkar, R. (2017). Exploring architectures, data and units for streaming end-to-end speech recognition with rnn-transducer. In *Proc. of ASRU*, page 193–199.

Ringeval, F., Sonderegger, A., Sauer, J. S., and Lalanne, D. (2013). Introducing the recola multimodal corpus of remote collaborative and affective interactions. pages 1–8. IEEE Computer Society.

Tzirakis, P., Zhang, J., and Schuller, B. W. (2018). End-to-end speech emotion recognition using deep neural networks. In *ICASSP*, pages 5089–5093. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wu, X., Liu, S., Cao, Y., Li, X., Yu, J., Dai, D., Ma, X., Hu, S., Wu, Z., Liu, X., et al. (2019). Speech emotion recognition using capsule networks. In *ICASSP*, pages 6695–6699. IEEE.

Xie, Y., Liang, R., Liang, Z., Huang, C., Zou, C., and Schuller, B. (2019). Speech emotion classification using attention-based lstm. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(11):1675–1685.

Zhang, Z., Wu, B., and Schuller, B. (2019). Attention-augmented end-to-end multi-task learning for emotion prediction from speech. In *ICASSP*, pages 6705–6709. IEEE.

8. Language Resource References

Busso, Carlos and Bulut, Murtaza and Lee, Chi-Chun and Kazemzadeh, Abe and Mower, Emily and Kim, Samuel and Chang, Jeannette N and Lee, Sungbok and Narayanan, Shrikanth S. (2008). *IEMOCAP: Interactive emotional dyadic motion capture database*. Springer.

Mark Dredze and John Blitzer. (2009). *Multi-Domain Sentiment Dataset*.

Go, Alec and Bhayani, Richa and Huang, Lei. (2009). *Twitter sentiment classification using distant supervision*.

Godfrey, John J. and Holliman, Edward C. and McDaniel, Jane. (1992). *SWITCHBOARD: Telephone Speech Corpus for Research and Development*. IEEE Computer Society, ICASSP'92.

Keith, Brian and Fuentes, Exequiel and Meneses, Claudio. (2017). *A Hybrid Approach for Sentiment Analysis Applied to Paper Reviews*.

Maas, Andrew L. and Daly, Raymond E. and Pham, Peter T. and Huang, Dan and Ng, Andrew Y. and Potts, Christopher. (2011). *Learning Word Vectors for Sentiment Analysis*. Association for Computational Linguistics.

McKeown, Gary and Valstar, Michel and Cowie, Roddy and Pantic, Maja and Schroder, Marc. (2012). *The SEMAINE Database: Annotated Multimodal Records of Emotionally Colored Conversations Between a Person and a Limited Agent*. IEEE Computer Society Press.

Bo Pang and Lillian Lee. (2004). *A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts*.

Bo Pang and Lillian Lee. (2005). *Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales*.

Ringeval, Fabien and Sonderegger, Andreas and Sauer, Jürgen S. and Lalanne, Denis. (2013). *Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions.* IEEE Computer Society.

Zadeh, Amir and Pu, Paul. (2018). *Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph*.

Zadeh, A and Liang, PP and Poria, S and Vij, P and Cambria, E and Morency, LP. (2018). *Multi-attention Recurrent Network for Human Communication Comprehension*.