Me, myself, and ire: Effects of automatic transcription quality on emotion, sarcasm, and personality detection

John Culnan, Seongjin Park, Meghavarshini Krishnaswamy, Rebecca Sharp
University of Arizona, Tucson, Arizona, USA
{jmculnan, seongjinpark, mkrishnaswamy, bsharp}@email.arizona.edu

Abstract

In deployment, systems that use speech as input must make use of automated transcriptions. Yet, typically when these systems are evaluated, gold transcriptions are assumed. We explicitly examine the impact of transcription errors on the downstream performance of a multi-modal system on three related tasks from three datasets: emotion, sarcasm, and personality detection. We include three separate transcription tools and show that while all automated transcriptions propagate errors that substantially impact downstream performance, the open-source tools fair worse than the paid tool, though not always straightforwardly, and word error rates do not correlate well with downstream performance. We further find that the inclusion of audio features partially mitigates transcription errors, but that a naive usage of a multi-task setup does not. We make available all code and data splits needed to reproduce all of our experiments.[1]

1 Introduction

With the large amount of available speech data, multimodal approaches to classic natural language processing tasks are becoming increasingly prevalent. Many of these proposed systems, however, demonstrate their gains under the assumption of gold transcriptions (e.g., Poria et al., 2019; Ghosal et al., 2020; Castro et al., 2019), a condition which is highly unrealistic in real-world scenarios. In practice, deployed systems will need to utilize an automatic speech recognition (ASR) tool to obtain transcriptions. However, when selecting the best tool given the constraints of the use case, the results of an *intrinsic evaluation* such as the word error rate (WER) are not necessarily correlated with *extrinsic performance* on the downstream task of interest (Faruqui et al., 2016). This issue is exacerbated when considering that ASR tools may perform quite differently across different domains (Georgila et al., 2020).

In this initial work, we explore how much transcription errors affect performance of a multimodal system on several related but distinct downstream tasks and domains. We compare two open-source and one paid transcription tools and evaluate on three multimodal English-language tasks: emotion, sarcasm, and personality detection. We explicitly compare intrinsic and extrinsic evaluations, and discuss the utility of WER as an indicator of task-performance.

Our specific contributions are:

1. An exploration of the relationship between transcription WER and performance on downstream multimodal tasks. We show that overall, the transcriptions from the paid tool may be more useful than those of the open-source tools, but that they perform worse than gold. Further, we show that differences in WER do not describe well the differences in the downstream tasks.

2. We explore using multitask (MT) training for mitigating issues with transcription quality. We find that, in this setting, MT does not help model performance, suggesting that problems with transcription quality need a more thoughtful approach to overcome. On the other hand, we show that inclusion of the audio modality does improve performance for all datasets when using automatic transcriptions, indicating that additional modalities can be helpful to mitigate transcription errors.

2 Related Work

Multimodal language work frequently makes use of video, audio, and text modalities to examine emo-

[1] https://github.com/clulab/tomcat-speech

Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 250–256
April 19, 2021. ©2021 Association for Computational Linguistics

tion (Zadeh et al., 2018), sentiment (Soleymani et al., 2017), and personality (Rissola et al., 2019). Previous work in emotion recognition has used speech features alone (Latif et al., 2020), speech and text (Atmaja and Akagi, 2020), and speech, text, and video (Tsai et al., 2018) to make predictions. Ghosal et al. (2020) use transformer-based models to establish a new state of the art on multiple multimodal datasets. Recent work in personality detection has often examined the OCEAN traits (openness, conscientiousness, extraversion, agreeableness, and neuroticism; Ponce-López et al., 2016), using either apparent (as perceived by others) (Yan et al., 2020) or self-reported traits (Celli et al., 2014).

ASR systems have also been emphasized in recent decades. Some can be custom trained or used with pretrained models (Povey et al., 2011; Lamere et al., 2003), while others are extensively trained but limited in their customization (Bano et al., 2020). Georgila et al. (2020) examine the performance of different ASR transcription tools on multiple ASR datasets, providing a strong reference for WERs; however, to the best of our knowledge, we are the first to compare this form of intrinsic performance to performance on the downstream tasks of emotion, personality, and sarcasm detection.

3 Approach

When multimodal data including speech is used in a deployed system, the related text features generally come from automatic transcriptions of the speech itself. Here, we compare how errors in these transcriptions affect the performance of a neural network model on three distinct tasks. Further, we compare results from both a single-task and a multitask (MT) network composed of data from two datasets, a common strategy for mitigating issues with limited or flawed data (Schulz et al., 2018).

As we are not trying to define a new state of the art, we use a simple model for our experiments. This model takes as input text and audio features from MELD and FirstImpr (Section 4), and feeds them through a late-fusion network (i.e., one which processes the modalities separately, then concatenates them before predicting). In the MT setting, the final layers are task specific, and the base layers share parameters.

We generate transcriptions for each dataset using three separate ASR systems (Section 4.3) and find the intrinsic performance (WER[2]; Section 6.1) as well as the extrinsic performance for each (i.e., performance on the downstream tasks of interest; Section 6).

4 Data

For our experiments we use three datasets covering distinct tasks. For each data point, we extract acoustic features and obtain three different transcriptions, in addition to the dataset-provided gold transcriptions, which form the basis for our comparative study. Dataset sizes are shown in Table 1.[3]

Dataset	Train	Dev	Test
MELD	8878	2220	2610
MUStARD	414	138	138
FirstImpr	6000	2000	2000

Table 1: Number of utterances in each data partition.

4.1 Datasets

Our selected datasets represent distinct tasks that may have different levels of reliance upon each modality for successful prediction. For example, while emotions and personality may be expressed through word choice as much as pronunciation, sarcasm detection should rely much more heavily upon acoustics.

Multimodal Emotion Lines Dataset (MELD; Zahiri and Choi, 2017; Poria et al., 2019): MELD provides 13708 annotated utterances (<5 words, with an average length of 3.59s) from 1,433 dialogues from the TV series *Friends*. Utterances are annotated for emotion (anger, disgust, sadness, joy, neutral, surprise and fear) and sentiment.

Multimodal Sarcasm Detection (MUStARD; Castro et al., 2019): MUStARD is a collection of 690 utterances (average of 14 tokens and 5.22s) from *Friends, The Golden Girls, The Big Bang Theory*, and *Sarcasmaholics Anonymous*. Each is gold-annotated as sarcastic or non-sarcastic.

First Impressions V2 dataset (FirstImpr; Ponce-López et al., 2016): FirstImpr contains 10,000 English utterances (average length 15s) taken from 3,000 YouTube video blogs and annotated for the OCEAN personality traits.

[2] We use the JiWER Python library (https://pypi.org/project/jiwer/) and the dataset-provided transcriptions to calculate WER.

[3] Note that for MELD we redistributed the train and dev partitions to make dev closer in size to test (we did not modify test in any way).

4.2 Acoustic features

We extract acoustic features with the Open-source Speech and Music Interpretation by Large-space Extraction toolkit v2.3.0 (OpenSMILE; Eyben et al., 2010). We use the INTERSPEECH 2010 (IS10; Schuller et al., 2010) or 2013 Paralinguistics Challenges (IS13; Schuller et al., 2013) features, using the set that performed the best on a task's development partition. These sets contain low-level descriptors (such as MFCCs and fundamental frequency) and the associated functionals, extracted at 10ms intervals (total 76 for IS10, 141 for IS13). We capture features only from the middle 50 percent of each audio file, and calculate mean feature values per utterance for both feature sets, with minimum, maximum, and mean plus/minus standard deviation concatenated to this for IS10.

4.3 Text features

Our text features consist of tokens extracted directly from the transcripts using the basic english tokenizer in `torchtext`.[4] We compare the transcriptions from the following:

CMU Sphinx Open Source Toolkit (Sphinx; Lamere et al., 2003): We utilize the open-source, lightweight PocketSphinx[5] version of Sphinx for transcription, using the pretrained acoustic and language models provided by CMU.

Google Cloud Speech-to-Text[6] (Google): Google Cloud Speech-to-Text is a commercial tool trained on data collected by Google and provided by users who have used Speech-to-Text and agreed to share their data. We use synchronous speech recognition.

Kaldi Speech Recognition Toolkit (Kaldi; Povey et al., 2011): Kaldi is an open-source tool for speech recognition. We use the Librispeech ASR model,[7] which is trained on Librispeech (Panayotov et al., 2015), and the online-decoding function.

Gold: Each dataset also provides a gold transcription. For MELD this is extracted from subtitles, MUStARD's comes both from subtitles and manual transcription, and FirstImpr's comes from a professional transcription service. With these we calculate the WERs (Section 6.1) and a ceiling performance in our extrinsic evaluations (Section 6).

5 Models

To evaluate the impact of transcription on performance, we create (a) baseline models that use only text or only audio features, plus (b) a multimodal model that uses audio and text. Note that as our goal here is not to achieve a new state of the art, but rather to explore the impact of real-world options for transcriptions, we use straightforward models that are not particularly architecturally tuned.

Text-only baseline: The text baseline consists of a two-layer LSTM (Hochreiter and Schmidhuber, 1997). 300d GloVe embeddings trained on 42B words (Pennington et al., 2014) are concatenated with 30d trainable text embeddings and fed through the network. The 100d output vectors representing each utterance are then fed through a prediction layer with a cross-entropy loss function[8].

Audio-only baseline The audio baseline is a simple feedforward neural network, where extracted audio features are averaged over the course of the utterance,[9] and used as input into two fully connected layers. The first layer decreases the audio to a 50-dimensional vector, while the second increases it back to its original size. The vector is finally fed through a prediction layer. We use IS10 features for MUStARD and IS13 for MELD and FirstImpr, chosen based on dev performance.

Multimodal model (MM) For our multimodal model, we concatenate the 100d output from the text component to the 141d (IS13) or 380d (IS10) output of the acoustic layers. These vectors are then fed through two fully connected layers and a prediction layer. This model works in both single task and multitask settings. As a multitask network (MM-MT), the two final layers are distinct for each dataset.

6 Experiments and Results

We evaluate how the automated transcriptions from different ASR tools affect performance on three distinct downstream tasks. We show the intrinsic

[4] https://pytorch.org/text/stable/index.html

[5] https://github.com/cmusphinx/pocketsphinx

[6] https://cloud.google.com/speech-to-text

[7] https://kaldi-asr.org/models/m13

[8] used for this and all other models

[9] While we also experimented with a version that used an RNN over the acoustic features, we found that it did not affect performance and it was far slower to train.

performance of the tools (in terms of WERs) and compare it to the extrinsic performance.

Dataset	Sphinx	Kaldi	Google
MELD	114.9	104.4	82.6
MUStARD	96.8	73.0	54.3
FirstImpr	97.4	83.5	63.8

Table 2: WER of MELD, MUStARD, and FirstImpr; results show the difficulty of ASR tools with these datasets' speech.

6.1 Intrinsic Transcription Evaluation

We use the provided gold transcriptions to calculate WERs, shown in Table 2. We see that Google consistently has lower WERs than either of the open-source tools, and that Kaldi consistently outperforms Sphinx, though the difference is smaller in MELD than the other datasets.

6.2 Extrinsic Evaluation

The results of our models trained on each transcription type for each task is given in Table 3. We report all results as weighted average F1 scores over all classes and we evaluate statistical significance by calculating p-values with bootstrap resampling over 10,000 iterations on model predictions with Bonferroni correction applied.

We include the comparable[10] state of the art (SotA) performance for reference for MELD and MUStARD. As FirstImpr was previously evaluated as a regression task, and here we perform maximum-class prediction for consistency, there is no relevant SotA to include. For MELD, Poria et al. (2019), use a CNN over GloVe embeddings and an LSTM over audio features from OpenSMILE.[11] For MUStARD, Castro et al. (2019) employ SVMs trained over the BERT (Devlin et al., 2018) encoding of the utterance, combined with extracted audio features, averaged over the utterance. We include this result only to give context to what we report; recall that here we evaluate on a test split, whereas they used 5-fold cross validation.

For text-only results, ASR transcriptions show significantly lower performance than the gold transcriptions in MELD ($p<0.001$) and, to a smaller degree, MUStARD, although the Kaldi transcriptions are better in FirstImpr ($p<0.001$). With audio included, i.e., the multimodal systems, we again see that gold transcription models achieve higher performance than those with ASR transcriptions for MELD ($p<0.001$) and MUStARD. For FirstImpr, the Google transcriptions yield the best performance, although this difference is not significant.

	MELD (7cls)	MUStARD (2cls)	FirstImpr (5cls)
Poria et al. (2019)	55.51	–	—
Castro et al. (2019)	–	71.80	—
Aud	37.38	**75.42**	32.99
Txt (Kaldi)	33.94	52.61	32.35
Txt (Sphinx)	32.68	57.12	30.11
Txt (Google)	37.60	58.11	31.66
Txt (Gold)	**57.32**	59.48	30.20
MM (Kaldi)	39.47	69.51	34.22
MM (Sphinx)	37.90	70.89	32.71
MM (Google)	40.94	70.12	**35.98**
MM (Gold)	56.28	73.98	35.05
MM-MT (Kaldi)	39.36	—	32.99
MM-MT (Sphinx)	39.30	—	32.81
MM-MT (Google)	40.17	—	34.75
MM-MT (Gold)	56.57	—	34.21

Table 3: Extrinsic task performance of our single-task system using acoustic features (Aud), text features (Txt), or multimodal (MM). We also show performance of the multimodal multitask system (MM-MT) that utilizes MELD and FirstImpr. All results presented are weighted average F1.

	MELD	MUStARD	FirstImpr
Gold	139580	10991	523387
	(95.7%)	(98.7%)	(98.4%)
Google	88913	7315	436272
	(99.8%)	(99.8%)	(100%)
Kaldi	89082	7039	369201
	(100%)	(99.9%)	(100%)
Sphinx	103659	6891	419214
	(100%)	(100%)	(100%)

Table 4: GloVe coverage (in tokens) for each transcription type with each dataset. Percentage of all corpus tokens appearing in GloVe are shown in parentheses.

However, when only considering the ASR models (i.e., in the deployment scenario), the inclusion of audio features improves all models, indicating that these features *are* able to mitigate some of the noise arising from imperfect transcriptions. That said, it is worth noting that for MUStARD, the best performance is achieved by using audio features only, though again this difference is not significant.

Among the ASR systems, we see that despite the substantial differences in WERs (Table 2), on the extrinsic evaluation, the story is more nuanced. Google transcription models do the best on MELD and FirstImpr, but on MUStARD, Sphinx transcriptions have the best performance. Further, none of

[10] Since we do not use the surrounding context or external resources in this initial work, we provide the results of the previous best performing system that did likewise.

[11] We are reading between the lines on this, as the paper does not provide clear explanation of this model, which they call *cMKL*.

	Anger	Disgust	Fear	Joy	Sadness	Surprise	All Emotions
Gold	934 / 207	747 / 159	1398 / 239	2306 / 198	1182 / 220	1080 / 138	4428 / 656
Google	805 / 208	612 / 158	1056 / 245	1919 / 198	978 / 202	863 / 135	3666 / 644
Kaldi	664 / 216	521 / 157	1282 / 270	2008 / 241	1054 / 260	978 / 143	3861 / 741
Sphinx	973 / 239	665 / 168	1126 / 279	1806 / 211	1070 / 249	1063 / 145	3704 / 712

Table 5: Number of tokens (left) and types (right) of words from the NRC Word-Emotion Association Lexicon (Mohammad and Turney, 2010, 2013) for each emotion type appearing in the MELD transcriptions. Each word may be associated with more than one emotion, so the overall count is lower than the sum of the individual emotions.

these differences are statistically significant. Thus, depending on the task of interest, there may be no large advantage to more expensive tools. We also experiment with using a multi-task (MT) setup to determine whether using MT can mitigate transcription errors. We use only MELD and FirstImpr, as they both did best with the IS13 features. Further, the majority of utterances in MUStARD appear in MELD and cannot be restricted to the training partition, so it cannot be used fairly in a multitask system with MELD. For all transcription types, the MT setup performs similarly to the corresponding multimodal single-task models, suggesting that a more thoughtful approach may be needed to leverage external data to mitigate transcription noise, which is beyond the scope of the current work.

6.3 Error analysis

While transcription WERs are not intrinsically linked to model performance on a global level, it is possible that additional factors are at play. As such, we analyze GloVe's coverage for each transcription of each dataset and examine emotion words present in each transcription for MELD.

Results of Table 4 reveal that gold transcriptions contain both a larger number of overall tokens and a smaller percentage of coverage by GloVe than all other datasets. This larger number of tokens may allow a system to make fine-grained distinction, indicating potential need for caution with automatic transcription selection depending upon the dataset of interest. Google has the second highest number of tokens covered in GloVe for FirstImpr and MUStARD, while Sphinx has the second most for MELD. As models using sphinx transcriptions perform numerically worst with MELD, this indicates that GloVe coverage alone does not always correspond to downstream task performance.

For more fine-grained detail, we examine transcription success in one particular domain: identification of emotion words. To do this, we determine the number of emotion words identified for each transcription using the NRC Word-Emotion Association Lexicon (EmoLex; Mohammad and Turney, 2010, 2013).

As shown in Table 5, between the gold and automatic transcriptions, we see different patterns between the *type* and *token* frequencies of emotion words. Overall, the token frequency of emotion words is lower in the automatic transcripts than in the gold transcriptions, largely due to the increase in tokens of words expressing joy and fear, but the *type* frequency is higher, particularly with Kaldi and Sphinx. That is, there are fewer distinct emotion words in the gold transcriptions but they have more mentions. These frequency patterns may be partially responsible for the worse performance of models using Kaldi and Sphinx transcriptions, as some data points contain spurious false positive mentions of emotion words, and other data points are missing mentions of emotion words, due to transcription error.

7 Conclusions

We demonstrated that selecting the appropriate transcription service for a task is dependent upon the task and models to be used. Paid transcriptions may result in better model performance, but differences in task performance may be much smaller than suggested by only considering WER. Further, depending on the task, the noise introduced from automatic transcriptions may be mitigated by including additional input modalities, such as audio.

Acknowledgments

Research was sponsored by the Army Research Office, Grant Number W911NF2010002. The views and conclusions are those of the authors and should not be interpreted as representing official policies, expressed or implied, of the Army Research Office or U.S. Government. Rebecca Sharp declares a financial interest in lum.ai. This interest has been properly disclosed to the University of Arizona Institutional Review Committee and is managed in accordance with its conflict of interest policies.

References

Bagus Tris Atmaja and Masato Akagi. 2020. Dimensional speech emotion recognition from speech features and word embeddings by using multitask learning. *APSIPA Transactions on Signal and Information Processing*, 9.

Shahana Bano, Pavuluri Jithendra, Gorsa Lakshmi Niharika, and Yalavarthi Sikhi. 2020. Speech to text translation enabling multilingualism. In *2020 IEEE International Conference for Innovation in Technology (INOCON)*, pages 1–4. IEEE.

Santiago Castro, Devamanyu Hazarika, Verónica Pérez-Rosas, Roger Zimmermann, Rada Mihalcea, and Soujanya Poria. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Fabio Celli, Elia Bruni, and Bruno Lepri. 2014. Automatic personality and interaction style recognition from facebook profile pictures. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1101–1104.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference ofthe North American Chapter of the Association for Computational Linguistics: Human LanguageTechnologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the Munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM international conference on Multimedia*, pages 1459–1462.

Manaal Faruqui, Yulia Tsvetkov, Pushpendre Rastogi, and Chris Dyer. 2016. Problems with evaluation of word embeddings using word similarity tasks. *arXiv preprint arXiv:1605.02276*.

Kallirroi Georgila, Anton Leuski, Volodymyr Yanov, and David Traum. 2020. Evaluation of off-the-shelf speech recognizers across diverse dialogue domains. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 6469–6476.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Paul Lamere, Philip Kwok, Evandro Gouvea, Bhiksha Raj, Rita Singh, William Walker, Manfred Warmuth, and Peter Wolf. 2003. The cmu sphinx-4 speech recognition system.

Siddique Latif, Rajib Rana, Sara Khalifa, Raja Jurdak, Julien Epps, and Bjórn Wolfgang Schuller. 2020. Multi-task semi-supervised adversarial autoencoding for speech emotion recognition. *IEEE Transactions on Affective Computing*.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Saif M Mohammad and Peter D Turney. 2013. Crowdsourcing a word–emotion association lexicon. *Computational intelligence*, 29(3):436–465.

V. Panayotov, G. Chen, D. Povey, and S. Khudanpur. 2015. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Víctor Ponce-López, Baiyu Chen, Marc Oliu, Ciprian Corneanu, Albert Clapés, Isabelle Guyon, Xavier Baró, Hugo Jair Escalante, and Sergio Escalera. 2016. Chalearn lap 2016: First round challenge on first impressions-dataset and results. In *European conference on computer vision*, pages 400–418. Springer.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*. IEEE Signal Processing Society. IEEE Catalog No.: CFP11SRW-USB.

Esteban Andres Rissola, Seyed Ali Bahrainian, and Fabio Crestani. 2019. Personality recognition in conversations using capsule neural networks. In *IEEE/WIC/ACM International Conference on Web Intelligence*, pages 180–187.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Eleventh Annual Conference of the International Speech Communication Association*.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. 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*.

Claudia Schulz, Steffen Eger, Johannes Daxenberger, Tobias Kahse, and Iryna Gurevych. 2018. Multi-task learning for argumentation mining in low-resource settings. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. 2017. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14.

Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2018. Learning factorized multimodal representations. *ICLR*.

Shen Yan, Di Huang, and Mohammad Soleymani. 2020. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 361–369.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1).

Sayyed M Zahiri and Jinho D Choi. 2017. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *Proceedings ofthe AAAI Workshop on Affective Content Analysis*, pages 44–51, New Orleans, LA.