INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Detecting Domain-specific Credibility and Expertise in Text and Speech

Shengli Hu

Dataminr. Inc.
sh2264@cornell.edu

Abstract

We investigate and explore the interplay of credibility and expertise level in text and speech. We collect a unique domain-specific multimodal dataset and analyze a set of acoustic-prosodic and linguistic features in both credible and less credible speech by professionals of varying expertise levels. Our analyses shed light on potential indicators of domain-specific perceived credibility and expertise, as well as the interplay in-between. Moreover, we build multimodal and multi-task deep learning models that outperform human performance by 6.2% in credibility and 3.8% in expertise level, building upon state-of-the-art self-supervised pre-trained language models. To our knowledge, this is the first multimodal multi-task study that analyzes and predicts domain-specific credibility and expertise level at the same time.[1]

Index Terms: multimodal learning, multi-task learning, computational social science, computational paralinguistics

1. Introduction

We are faced with such dilemmas in social situations at all times: assessing the credentials of an acquaintance who claims to be a research scientist specializing in speech recognition, determining if a sommelier who recommends a particular wine is trustworthy and knowledgeable, choosing which lawyer to handle your case based on consultation and related documents, and deciding whether to trust your mechanic to replace a part of your car that you know nothing about. What all of these scenarios share in common is the fact that without enough domain knowledge, it would be challenging to make informed judgments or decisions, because of the difficulty of accurately gauging the level of expertise and credibility of the supposed expert. Despite the practical relevance and importance of detecting credibility and levels of expertise in text and speech, research on this front has been scarce. It is partly because large-scale datasets with groundtruth labels of credibility and expertise are difficult to come by. It is only in rare cases can we clearly define the level of expertise and disambiguate credibility in the wild. In the present study, we present a unique domain-specific corpus of text and speech, collected from field experiments with groundtruth labels, which allows us to investigate automatic detection of credibility and expertise. More specifically, we address the following questions:

1. What are the respective indicators of credibility and expertise, and how do they differ or overlap?
2. Are there any significant individual differences in such indicators?
3. Can we improve on human performance with a multimodal and multi-task classification framework?

[1] Work done at Cornell University.

2. Related Work

In terms of detecting expertise, the most relevant research stream is perhaps meeting analysis, where research activity has focused on speech recognition in meetings [1, 2], topic detection [3], role and expertise detection [4, 5], among others. The current study deviates in that it studies a unique but more general setting where a professional is pitching to a target client a product or service with a question and answering session that follows. What separates the current study apart is the unique insights into the interaction of expertise and credibility from a multimodal perspective, as well as the performance gain over human non-experts with a multi-task framework.

Credibility detection has been vetted in other contexts: [6] predicted credibility in community question answering forums using text, whereas the current study focuses on both acoustic-prosodic and linguistic features. Meanwhile, discreditable behavior such as deception and hiding information has been studied too. In this sense, this paper relates to detecting concealed information [7] and deception detection in both text [8, 9, 10] and speech [11, 12, 13], including perceived deception detection [14]. As opposed to deception or information concealment, credibility is inherently subjective and therefore more susceptible to idiosyncrasies, making it arguably less predictable. However, it is likely that the same rationale of changes in cognitive load applies in the current setting.

Automatic fact-checking attracted much interests [15, 16] over time, especially NLP in the contexts of journalism and politics. So is the research stream on factual accuracy most evident in evaluating NLP tasks such as summarization and/or translation [17, 18]. While we defined and automated the measurement of domain-specific *factuality* detailed in Section 4.2, our contributions lie in proposing and demonstrating *factuality* to be a statistically significant indicator and useful feature of automatic detection of credibility and expertise level, how it affects credibility and expertise predictions differently, and how to leverage it to better detect credibility and expertise levels.

Methodologically, our multimodal deep learning model extends the memory-efficient BERT [19] — ALBERT [20] — with a speech segment encoder. Perhaps closest in spirit to ours is SpeechBERT [21], which is an end-to-end cross-modal transformer-based pre-trained language model for spoken question answering. While applicable, our problem and setting are different in that audio signals provide additional useful information for downstream tasks, instead of as inputs to ASR for training end-to-end with downstream NLP tasks. Moreover, we couple it with a single multi-task model with auxiliary tasks and a dynamic training schedule to avoid catastrophic forgetting.

3. Data Collection and Processing

3.1. Sales Pitch Dataset

We collected a unique multimodal dataset of wine professionals at different expertise levels and perceived credibility pitching

http://dx.doi.org/10.21437/Interspeech.2020-1518

wines to and fielding questions by potential customers. Specifically, each session of sales pitch starts with a wine professional describing a wine of interest by introducing various aspects of it including its flavor profile, the grape varietal, the region, the vintage, and the producer, among other relevant information. It is followed by additional questions from an informed customer and the wine professional's responses.

We recorded 81 sessions with a total of 43 professionals and collected factual information sheets of wines featured in each session via authoritative channels such as producers' or importers' websites complemented by [22, 23]. We also collected information about the speakers' native language (77% American English, the rest French, Italian, and German), gender (44% female), professional credentials, years of industry experience, and any specific information of the wine such as vintage, country, region, grape growing or winemaking techniques, topography of vineyard site, soil types, climate, producer philosophy, and history, etc. Additional datasets that were collected for multi-task learning will be detailed in Section 5.3.

3.2. Annotation and Preprocessing

Groundtruth labels of expertise levels were obtained by calculating a weighted average of professional credentials and years of experience, and converting to a binary variable of an even split in consultation with a certified sommelier by CMS and WSET diploma holder, who also helped to annotate the dataset by assigning every session a binary credibility score to the wine professional in question. Two illustrative examples of transcriptions and assigned labels are shown in Table 1.

Table 1: *Examples of samples and corresponding annotations of credibility (Cr) and expertise level (Ex).*

Transcription	Cr	Ex
"... They also make incredible Syrahs, which I believe were some of their first grapes planted, and also Pinot Noirs. They just make beautiful wines, they make Riesling, gorgeous, but I do love their Chardonnays, a little bit richer, a little bit fuller than Chablis...."	0	0
"... The other reason, more scientifically is, because Syrah has so much color pigmentation, there's not enough juice in the grape itself to extract all the color out, so you need more juice, so when you crush Syrah... As an ode to the northern Rhone tradition...."	1	1

We aligned the audio samples with speaker id and product id using Praat. We discarded sections unrelated to the sales pitch, such as small talk, segments of others talking. The remaining audio samples were transcribed with Google Speech API for automatic speech recognition, and hand-corrected afterward. We segmented each session into turn units, where a turn is defined as a maximal sequence of inter-pausal units (pause-free segments separated by a minimum pause length of 50 ms) from a single speaker without any interlocutor speech that is not a back channel. Labels of speaker id (and speaker metadata), wine identity (and product metadata) were assigned to each turn accordingly. We define single turn segments as individual turns of a speaker in any session separately and aggregate them by speaker and wine as multiple turn segments.

Our classification is performed on both segmentation results of the data, whereas statistical analyses are on multiple turn segments. The resulting corpus totaled 132 hours, and 2534 multiple turn and 7934 single turn segments. 41% of turn segments were labeled as 0 (lower credibility) and the rest 1 of credibility, and the percentages associated with expertise levels are 52% for 0 (lower expertise), and 48% for 1 (higher expertise), respectively. We randomly split our entire set into training, development, and testing sets at the ratio of 70:10:20 separately for single and multiple turn. Evaluation results were based on 5-fold cross-validation.

4. Feature Extraction

4.1. Acoustic-prosodic Features and Indicators

We extract 8 low-level acoustic features: intensity mean and max, pitch mean and max, voice quality features (shimmer, jitter, noise-to-harmonics ratio), and speaking quality, along with 13 Mel-Frequency Cepstral Coefficients (MFCCs) per window of 256 frames and stride of 100 frames with softwares Praat and Parselmouth. Following previous studies, we use OpenSMILE to extract the 2013 Computational Paralinguistics Challenge baseline feature set [24], and the 2009 Emotion challenge baseline feature set [25]. The two feature sets were used in our machine learning classification tasks. All the audio features are z-score normalized by speaker.

Table 2 shows the statistically significant low-level acoustic features for both the credible and the less credible based on paired t-tests between the features of the two groups, corrected for family-wise Type I error by controlling the false discovery rate (FDR) at $\alpha = 0.05\%$.

We observe an increase in speaking duration, and a decrease in maximum intensity and speaking rate, across all speakers (the last column), suggesting that when speakers are perceived more credible, they on average tend to speak with lower maximum intensity, rate, and longer duration.

To understand the individual differences in speech of perceived credibility, we report the same test statistics for speakers grouped by gender, native language, and expertise level. We find that maximum pitch is lower for male speakers perceived to be more credible, as well as those of higher expertise levels, but not for female speakers or individuals of lower expertise levels. In contrast, for female speakers, perceived credibility is associated with speaking at a lower rate, for longer, and interestingly with better voice quality. Individuals with lower expertise appear to be more credible when they speak with a lower rate. English native speakers are perceived as more credible when they speaker with a lower maximum intensity.

Some of these results appear consistent with previously reported acoustic-prosodic indicators in deception [12] or information concealment [7]. For instance, changes in maximum pitch, maximum intensity, and speaking rate have been found associated with untruthful speech, and gender differences have been highlighted in various studies [12, 26, 7]. With the caveat that perceived credibility is not necessarily equivalent to credibility, however, with a politically neutral domain expert of ethnic minority providing the labels, we believe the distribution to be as close to true credibility as it could be.

We conduct the same series of statistical tests for expertise levels, the results of which are shown in the same table 2 indicated with E, (E), and (-E). We did not identify statistically significant acoustic indicators of expertise levels with corrected p-values. Without correction, however, a higher speaking rate and shorter duration appear to be associated with higher expertise level especially for male English native speakers, and shorter duration

Table 2: *Low-level Acoustic Indicators of Credibility and Expertise. C indicates Credibility and E Expertise. (C) and (E) indicates significant uncorrected p-values. - indicates negative correlation.*

Feature	Male	Female	Low Exp	High Exp	English	French	Italy	All
Pitchmax	(-C)			(-C)				
Pitchmean								
Intensity					-C		(E)	-C
Intensity								
Rate		(-C) E	(-C)					(-C) (E)
Duration		C (-E)			(-E)	(-E)	(-E)	C (-E)
VQ		(C)						

appears to be indicative of higher expertise except in female speakers, all of which appear intuitively plausible.

4.2. Linguistic Features and Indicators

LIWC: Following relevant literature, we extract 93 semantic classes using LIWC 2015 [27, 28]. They include standard linguistic dimensions, grammar, psychological processes, time orientation, relativity, and formality.

Linguistic: We extract 11 linguistic features based on results from previous literature. Included are binary and numeric features capturing **hedging** [29], linguistic and syntactical **distinctiveness**, **subjectivity**, **sentiment** (valence, intensity), **contraction**, **specificity** [30], **breadth** of knowledge, **depth** of knowledge, and **factuality** [17, 18].

We measure linguistic and syntactical distinctiveness in the same way as in [7] using a large-scale review corpus in [31]. Subjectivity and sentiment measures were extracted with TextBlob software.

We measure the depth of knowledge by building domain-specific rule-based algorithms following hierarchical study guides and practice problem sets used by wine professionals. Our method counts both the number and percentage of named entities belonging to different tiers of study materials , weighted by pre-specified weights calculated based on a weighting scheme identical to term frequency inverse document frequency (tf-idf), to assign each named entity a depth score. We apply max pooling by speaker and session, normalized to be within 0 and 1.

We measure the breadth of knowledge by topic diversity using *topic entropy* [32] where topic distributions are derived from a *seeded LDA* [33], seeded with keywords for six categories: terrior (climate, geography, etc.), grape growing, winemaking, maturation, wine law, and wine business.

We measure speech factuality (the extent to which the speech is true to facts) by calculating the *factual consistency* of transcripts against information sheets we collect for each session, using an automatic generative method. We first train an LSTM entity tagger trained on a large wine corpus where a vocabulary from the index pages of [23] was used to create a set of groundtruth labels of entities of location, person, organization, product, and else. Based on the domain-specific entity tagger, we follow [34] to automatically generate questions from the information sheets, based on which we apply question answering [35] to the speech transcription, and the information sheets respectively. We measure *factuality* as the distance between the generated answers from corresponding information sheets and speech transcriptions by F1 score.

Length and **Ngrams**: We include the average total number of words in total and per sentence, the average length of words in total, per sentence, and per word. Unigrams, bigrams, and trigrams are extracted.

Table 3 shows (1) the top n-gram features for perceived credible and high expertise classes from a logistic regression classifier, which yields an F1-score of 64.01% for expertise level, and 58.72% for credibility; (2) the statistically positively (negatively) significant LIWC, linguistic, and other features for both perceived credibility and expertise based on the same statistical tests as detailed in Section 4.1.

Table 3: *Linguistic Indicators of Credibility and Expertise. Those in parentheses are of negative statistical significance, as opposed to positive statistical significance without parentheses.*

Feature	Credibility	Expertise
N-grams	elevation, perhaps, nose, natural, intense, right there (value, crisp, not unlike)	plow, savory, lay down, natural, percent, tension
LIWC	feel, discrep, verb, affect, focuspast, tentat, ingest, compare, cause; (posemo, cogproc, funct)	compare, cause, certain; (leisure)
Linguistic	intensity, valence; (hedging, subjectivity)	ling_distinct, specificity
Other	factuality, specificity	depth, #word factuality

Consistent with results on perceived deception [14], we find LIWC dimensions such as *clout* or *certain* not significant indicators of perceived credibility, however, *certain* was indicative of higher expertise level. Likewise, more *hedging* is found to be associated with less perceived credibility, echoing findings that more filled pauses and hedge words are believed to be more likely deceptive [14]. Contrary to previous literature [14, 7] where *specificity* is shown associated with deceptive or information concealing behaviors, however, we find greater *specificity* to be a significant indicator of perceived credibility and higher expertise level in our domain-specific context, an example of and potential explanation for which is shown in Table 1. Another linguistic feature found statistically significant for both is *factuality*, as opposed to *subjectivity*, which we find to be an indicator of less credibility. Such could be explained by the fact that experts are more likely to cite relevant facts in their speech and thus appear less subjective. Interestingly, among statistically significant n-grams for either credibility or expertise, *natural* shows up as an indicator for both. We hypothesize that it could be due to an inherent bias that most participants, as well as expert annotators, are proponents of the natural wine movement in the domain. *Linguistic distinctiveness* and *depth* of knowledge are also found to be statistically significant for expertise prediction, so is the number of words in total and per sentence.

5. Classification Experiments

We balance our dataset by random upsampling for credibility classification, since the number of positive labels is almost twice of negative labels.

5.1. Baseline: BiLSTM + MLP

To establish multimodal baselines , we train Bidirectional Long Short-Term models (BiLSTM) with sequences of word embeddings (GloVe) pretrained on the large-scale review corpus mentioned in Section 4.2, multi-layer Perceptions (MLP) with acoustic feature sets, and the combinations thereof. We use

Bayesian optimization to tune the hyperparameters (the number of hidden layers of MLP, the number of hidden units per layer, optimizers and associated parameters, dropout rate, and batch size), and concatenate embeddings learned from acoustic features passed through an MLP, embeddings passed through a BiLSTM, and a vector of additional linguistic features (detailed in Section 4.2) for the last softmax layer. The combined model structure follows [13], and we end up using 4 hidden layers, each with 560 hidden units followed by ReLU, and the IS 2009 Emotion Challenge feature set for MLP. Batch Normalization, dropout, and Adam were used in training with a cross-entropy loss. Results are denoted as Multimodal in Table 4.

5.2. ALBERT and Multimodal ALBERT

Recent breakthroughs in language representation learning proved self-supervised pre-trained language models [19] powerful in a range of NLP tasks. We adopt the memory-efficient version of BERT — ALBERT [20], augment it with an audio segment encoder, and pre-train on masked multimodal modeling and multimodal alignment prediction tasks jointly with text and speech, using a corpus of 104 videos (audios with English subtitles) from the streaming service SommTV. The first task randomly masks 15% of both words and audio segments and reconstructs them given the remaining inputs. The second task predicts whether a speech segment corresponds to a sentence.
To augment it with an audio segment encoder, we apply a BiLSTM as the encoder and an LSTM the decoder, padded with fully-connected layers. For each pair of the audio segment and corresponding word, we add L1 loss between their embeddings to force the autoencoder to extract semantic features while retaining acoustic features for reconstruction.
After pre-training, we concatenate audio embeddings with text embeddings from transcriptions after self-attention and co-attention mechanisms, for the downstream classification task fine-tuning. The overall architecture is similar to SpeechBERT [21] except that ours is based on ALBERT and additional feature vectors of individual differences are also concatenated. Results are denoted as MALBERT in Table 4.

5.3. Multi-task Learning

Based on the combined model in Section 5.2, we explore multi-task learning by adding three more tasks that either share the same training set or use additional datasets of (1) 81 information or technical sheets that correspond to 81 sessions in the Sales Pitch Dataset; (2) 96 videos of 106.4 hours with English subtitles as transcriptions from the streaming service SommTV; (3) study guides and expert guides from The GuildSomm website; and (4) the introductory and advanced course materials from The Course of Master Sommelier. The three tasks are (1) multi-class classification of the topic of the document or speech; (2) binary classification of whether the study materials are of the introductory level or advanced level; and (3) multimodal entailment: given speech transcriptions and corresponding information sheets, judge the correctness or predict their semantic relationship, where a manually annotated set of 1543 multiple turn segments was used as groundtruths. Due to different dataset sizes and potentially different difficulties associated with different tasks, catastrophic forgetting and overfitting of simpler tasks could result. Thus we experiment with a dynamic training schedule that cycles through each task and implements training if and only if the task per epoch validation loss is improved more than 0.1%. Results are included in Table 4 as the last row MTL-MALBERT.

5.4. Measuring Human Performance

We calculate human performance by providing audio segments with corresponding transcriptions to 2 non-expert subjects and asking them to classify into credible or not, and expert or not, without revealing the identity of the speaker. The Cohen's Kappa between them for credibility turns out 0.23 and for expertise level 0.25, suggesting that it is not an easy task for non-experts and there was some but very little consensus between them. We calculate the average F1 scores against the groundtruch labels and use them as the non-expert human performance results as is shown in Table 4.

5.5. Results

Table 4 demonstrates the F1-scores of different model architectures detailed in Section 5, against human performance detailed in Section 5.4. Besides multimodal models, we also include text-only model ALBERT pre-trained on a large corpus combining reviews, study materials, and transcriptions, all detailed in earlier sections. Across all the models, multiple turn segmentation yields better F1-scores compared to single turn segmentation. It is sensible because multiple turn segments contain more information especially sequential information, contributing to better pattern learning and classification. Consistent with [13, 7], we find multimodal models outperform unimodal models for credibility detection by a large margin, and that carefully designed multi-task models further improve the F1-score to be 6.2% higher than human performance. For expertise detection, however, we find an efficient model based on only textual information achieves an F1-score just below human performance, upon which multimodal models do not appear to improve. Major performance boosts are due to carefully designed and executed multi-task models instead, achieving a 3.8% margin over human performance.

Table 4: *F1-scores of Experiments vs. Human Performance*

F1-score	Credibility		Expertise Level	
Model	single turn	multi turn	single turn	multi turn
Human	NA	57.34	NA	70.52
ALBERT - Text	56.20	57.33	66.85	68.89
Multimodal	57.46	59.81	66.26	67.72
MALBERT	58.75	61.62	66.78	68.91
MTL-MALBERT	60.36	63.51	69.42	74.27

6. Conclusions, Limitations, Future Work

We presented a multimodal study of domain-specific perceived credibility and expertise in text and speech. Our statistical analyses of acoustic-prosodic and linguistic characteristics of credibility and expertise shed light on subtle cues, comparing with and contrasting previous literature in perceived deception [14], and untruthful speech and text [36, 7]. We built multimodal and multi-task deep learning models that obtain an F1 score of 63.5% in predicting credibility and 74.1% in expertise level, outperforming non-experts' performance by 6.2% in credibility and 3.8% in expertise level. One caveat and limitation lies in its potential unethical applications regarding credibility prediction if productized. Even though there is an even split between sexes in training data, there is not enough racial diversity in the subjects from whom training data was generated, which could lead to unsatisfying performance in analyzing speech of minority groups. We look forward to future research work such as debiasing, incorporating ASR, integrating other modalities, etc.

7. References

[1] A. Stolcke, C. Wooters, N. Mirghafori, T. Pirinen, I. Bulyko, D. Gelbart, M. Graciarena, S. Otterson, B. Peskin, and M. Ostendorf, "Progress in meeting recognition: The icsi-sri-uw spring 2004 evaluation system," in *Proceedings NIST ICASSP 2004 Meeting Recognition Workshop, Montreal, National Institute of Standards and Technology*, 2004.

[2] T. Schultz, Q. Jin, K. Laskowski, Y. Pan, F. Metze, and C. Fügen, "Issues in meeting transcription-the isl meeting transcription system," in *Eighth International Conference on Spoken Language Processing*, 2004.

[3] M. Galley, K. McKeown, E. Fosler-Lussier, and H. Jing, "Discourse segmentation of multi-party conversation," in *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 2003, pp. 562–569.

[4] S. Banerjee and A. Rudnicky, "You are what you say: Using meeting participants' speech to detect their roles and expertise," in *Proceedings of the Analyzing Conversations in Text and Speech*, 2006, pp. 23–30.

[5] B. Bigot, I. Ferrané, J. Pinquier, and R. André-Obrecht, "Detecting individual role using features extracted from speaker diarization results," *Multimedia Tools and Applications*, vol. 60, no. 2, pp. 347–369, 2012.

[6] P. Nakov, T. Mihaylova, L. Màrquez, Y. Shiroya, and I. Koychev, "Do not trust the trolls: Predicting credibility in community question answering forums," in *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, 2017, pp. 551–560.

[7] S. Hu, "Detecting concealed information in text and speech," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 402–412.

[8] V. Pérez-Rosas and R. Mihalcea, "Experiments in open domain deception detection," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 1120–1125.

[9] A. Heydari, M. ali Tavakoli, N. Salim, and Z. Heydari, "Detection of review spam: A survey," *Expert Systems with Applications*, vol. 42, no. 7, pp. 3634–3642, 2015.

[10] J. T. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," *Discourse Processes*, vol. 45, no. 1, pp. 1–23, 2007.

[11] S. I. Levitan, G. An, M. Ma, R. Levitan, A. Rosenberg, and J. Hirschberg, "Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection." in *INTERSPEECH*, 2016, pp. 2006–2010.

[12] S. I. Levitan, A. Maredia, and J. Hirschberg, "Acoustic-prosodic indicators of deception and trust in interview dialogues," *Proc. Interspeech 2018*, pp. 416–420, 2018.

[13] G. Mendels, S. I. Levitan, K.-Z. Lee, and J. Hirschberg, "Hybrid acoustic-lexical deep learning approach for deception detection." in *INTERSPEECH*, 2017, pp. 1472–1476.

[14] S. I. Levitan, A. Maredia, and J. Hirschberg, "Linguistic cues to deception and perceived deception in interview dialogues," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, vol. 1, 2018, pp. 1941–1950.

[15] J. Thorne and A. Vlachos, "Automated fact checking: Task formulations, methods and future directions," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 3346–3359.

[16] N. Naderi and G. Hirst, "Automated fact-checking of claims in argumentative parliamentary debates," in *Proceedings of the First Workshop on Fact Extraction and VERification (FEVER)*, 2018, pp. 60–65.

[17] B. Goodrich, V. Rao, P. J. Liu, and M. Saleh, "Assessing the factual accuracy of generated text," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 166–175.

[18] A. Wang, K. Cho, and M. Lewis, "Asking and answering questions to evaluate the factual consistency of summaries," *arXiv preprint arXiv:2004.04228*, 2020.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[20] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[21] Y.-S. Chuang, C.-L. Liu, and H.-Y. Lee, "Speechbert: Cross-modal pre-trained language model for end-to-end spoken question answering," *arXiv preprint arXiv:1910.11559*, 2019.

[22] J. Hugh and R. Jancis, "The world atlas of wine," *Mitchell Beazle, London*, 2013.

[23] J. Robinson and J. Harding, *The Oxford companion to wine*. American Chemical Society, 2015.

[24] 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.

[25] B. Schuller, S. Steidl, and A. Batliner, "The interspeech 2009 emotion challenge," in *Tenth Annual Conference of the International Speech Communication Association*, 2009.

[26] S. I. Levitan, M. Levine, J. Hirschberg, N. Cestero, G. An, and A. Rosenberg, "Individual differences in deception and deception detection," *Proceedings of Cognitive*, 2015.

[27] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[28] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[29] A. Prokofieva and J. Hirschberg, "Hedging and speaker commitment," in *Proceedings of the 5th International Workshop on Emotion, Social Signals, Sentiment & Linked Open Data, Reykjavik, Iceland*, 2014, pp. 10–13.

[30] J. J. Li and A. Nenkova, "Fast and accurate prediction of sentence specificity," in *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[31] S. Hu, "Somm: Into the model," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 1153–1159.

[32] D. Hall, D. Jurafsky, and C. D. Manning, "Studying the history of ideas using topic models," in *Proceedings of the 2008 conference on empirical methods in natural language processing*, 2008, pp. 363–371.

[33] J. Jagarlamudi, H. Daumé III, and R. Udupa, "Incorporating lexical priors into topic models," in *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2012, pp. 204–213.

[34] X. Du, J. Shao, and C. Cardie, "Learning to ask: Neural question generation for reading comprehension," in *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2017, pp. 1342–1352.

[35] D. Chen, A. Fisch, J. Weston, and A. Bordes, "Reading wikipedia to answer open-domain questions," *arXiv preprint arXiv:1704.00051*, 2017.

[36] S. I. Levitan, "Deception in spoken dialogue: Classification and individual differences," Ph.D. dissertation, Columbia University, 2019.