Identifying Sentiment from Crowd Audio

Patrick J. Donnelly
Department of Computer Science
Oregon State University
Corvallis OR, United States
patrick.donnelly@oregonstate.edu

Alex Prestwich
Department of Computer Science
Oregon State University
Corvallis OR, United States
prestwia@oregonstate.edu

***Abstract*—Oftentimes the voices of a few drown out the voices of many. We interpret this problem literally as we investigate the ability of supervised machine learning models to predict sentiment directly from crowd audio in which multiple speakers are speaking simultaneously. Using a dataset of one second audio recordings of individuals saying "yes" or "no" (n=4753), we mixed together different voices speaking at the same time to create a large dataset (n=150k) of crowd audio responses. Each audio mixture was annotated with the specific proportion of the crowd response. On different mixture sizes ranging from 1 to 10 constituent voices, we trained BLSTM models using raw audio to predict the average sentiment of the crowd response. We were able to predict the average crowd sentiment for up to 10 speakers speaking simultaneously (Pearson's r=0.70), suggesting the feasibility of a deep-learning approach to approximate simple sentiment directly from audio recordings with overlapping speakers. We discuss our results in the context of the need for machine listening systems that make can make equitable real-time determinations about crowd sentiment, so that no voice goes unheard.**

Keywords—crowd audio, deep learning, long short-term memory, raw audio, sentiment analysis

I. Introduction

We often find ourselves in large crowd settings in which a cacophony of voices fuse into a single boisterous auditory texture. We hear the thunderous roar of the crowd in stadiums, audience cheers in concert halls, shrieks of elation on the playground, or the determined shouts of activists protesting in the streets. In these situations humans are able easily understand the overall sentiment of the crowd. And we are able to do this quickly, often triggering limbic responses such as fear or safety.

Which clues allow the brain to make such instantaneous judgements? Although evidence shows that our brains may favor visual stimuli over auditory stimuli [1], we are nevertheless able to make such judgements based on auditory information alone [2]. Perhaps our auditory system attends to a few of the closest voices in the crowd, relying on linguistic cues to make determinations. In auditory scenes with multiple sound sources, such as in the classic cocktail party problem, multiple voices fuse together to degrade the harmonicity of an individual's voice upon which the brain relies to identify and track target speakers [3]. Similarly, when multiple speakers are talking at different volume levels, we may expect to attend primarily to the loudest voices in the mixture. However, even when we are unable to make out any specific words, humans are usually rather adept at forming a general impression about the overall sentiment of a crowd response.

Although humans perform this task relatively well, how well can we expect an artificial intelligence to replicate this behavior? Would a machine listening algorithm be less affected by the inharmonicity of mixtures of voices speaking simultaneously? Would we expect an intelligent machine listener to also be biased by the loudest voices, especially if replying upon traditional speaker recognition features? Or, very importantly, can we teach a machine listener to also hear and prioritize these quieter voices? In this work we consider an intelligent machine listener that is able to estimate the rate of sentiment from overlapping speech in which not all of the voices agree.

In this work we describe an approach to automatically estimate simple sentiment from overlapping "yes" and "no" responses directly from crowd audio. This is a difficult task because the individual voices combine into a single sonic texture in which current state of the art speech recognition approaches cannot transcribe. Our study is novel in several respects. First, we consider only single channel audio streams of crowd speech, without attempting natural language processing (NLP) or source separation. Second, we train deep learning model based on raw audio with no attempt to extract acoustic, spectral, nor linguistic features. Given a trained model, this will improve the reaction time of our system for performance in real world environments. Third, we explicitly split our data into training and evaluation datasets in which no individual voice occurs in both. This strict validation procedure helps ensure our system will better generalize to new voices. Lastly we prioritize learning the true representation, not the majority sentiment or that of loudest voices. By training models to recognize the rate of sentiment, we strive to better hear, understand, and represent all voices in the crowd. This study is an initial exploration of this ideal.

This paper is organized as follows. In Section II we review literature related to the crowd sentiment. We describe our dataset of overlapping "yes" and "no" responses as Section III. We also outline the chosen model and experimental design. In Section IV we compare results of mixtures ranging from one to ten voices. Lastly, in Section V, we discuss our contributions, future research directions, and potential applications of this work.

II. Related Work

Machines are increasingly tasked to understand human speech and make determinations in real-time. From real time

978-1-6654-8157-1/22/$31.00 ©2022 IEEE

audio dictation (see review [4]) and smart personal assistants to voice controlled wheelchairs, voice control systems must quickly and correctly understand the meaning of the spoken command. This involves processing the audio signal, performing speech recognition, extracting meaningful semantic information, querying the cloud or a local database, interpreting the response, and responding with an action. Critically, these systems rely upon correct natural language understanding in order to respond appropriately.

Recognizing voice commands can be quite difficult in cases of overlapping speakers. Speech segmentation and diarization (see review [5]) remain active research areas within the field of NLP. Recent attempts have tasked recurrent deep learning models to perform speaker overlap-aware diarization [6]. Approaches that jointly perform speech recognition and speaker diarization models have also shown promise [7]. Recent research suggests that examining the harmonic structure associated with pitch in voiced speech can help segment speech with overlapping speakers [8]. Furthermore, convolutional models have been shown to count speakers in overlapped speech with lower error rates than human annotators [9].

There have been a few approaches that attempt to detect emotion from voice recordings (see review [10]) and some of these approaches do not rely on natural language understanding alone. Recent deep learning approaches to learn emotion directly from raw audio have focused on architectures stacking convolutional and recurrent layers [11]. Deep learning approaches have also been applied to learning de-identified prosodic representations from raw audio [12]. Other approaches seek to leverage different modalities of features together to predict sentiment. For example, ensemble techniques combining acoustic and linguistic features have been used to perform sentiment analysis in multiparty calls [13]. Other approaches exploring audio-visual content perform multimodal sentiment analysis on combinations of text, voice, and video (see review [14]). However, none of these studies have attempted to perform sentiment analysis on highly overlapping speech.

Although it is common practice to employ human-in-the-loop machine learning to crowd-source the labeling of sentiment analysis data, there has been relatively little research on sentiment analysis of the crowds themselves. The principle focus thus far has been natural language analysis of blog feeds on social networks. These approaches leverage crowd wisdom to predict outcomes of major events – such as predict the Super Bowl winner [15], forecast stock market trends [16], or analyze trends in video game reviews [17]. One active area of research is the detection of crisis events (see review [18]). In another unique approach, the authors performed visual sentiment analysis of disaster-related images on social media [19].

In the present work we consider single streams of crowd audio in which multiple voices are speaking "yes" or "no" simultaneously. For these reasons, and unlike most of the studies reviewed above, we do not attempt natural language understanding of the overlapping speech.

III. Data and Methods

To explore this approach we first needed to identify a dataset of audio recordings featuring overlapping speakers. We did not find such dataset that contained sentiment annotations. Instead we generated a dataset by mixing individual voices.

A. Dataset of Single Voice Audio

Given that our approach seeks to better listen to all voices in a crowd texture, we chose to generate a dataset rather than task human annotators to label existing audio of crowd responses. By controlling the mixture of voices in the crowd audio, we can train machine learning models against the actual proportion of sentiment rather than with annotations of a majority sentiment.

For our source of individual voice recordings, we selected the Google Brain simple speech command dataset. This dataset was featured in the 2018 TensorFlow speech recognition challenge[1]. It contains 65,000 one-second utterances of 30 short common words. These are recorded utterances of different human speakers, containing both male and female voices. From this dataset, we selected the commands "yes" (n=2376) and "no" (n=2377) for a total of 4753 single voice recordings, which we down-sampled to 16-bit 16 kHz wave files. We then batch normalized the individual audio files to attenuate the audio while preserving relative differences between the different speakers.

Each audio example contains a single spoken word, either yes or no. The precise timing of the attack of the word varies between examples, but each utterance is completed within the one second long sample. To remain consistent to expected real-world conditions, we made no attempt to align the attack between the different individual one second utterances.

To ensure independence between our train and test sets, we partitioned the single voice files into two source partitions, containing approximately equal numbers of yes and no examples. From one source dataset (n=2377), we generate mixtures of voices for training our models. From the other (n=2376), we generate mixtures to be used for evaluating our models. By strictly partitioning the source voices between train and test sets, we hope to encourage our trained models to better generalize to new voices.

B. Datasets of Crowd Audio

From our source partition of recordings for training, we then create a dataset of mixtures of voices. Selecting from the set of individual source recordings, randomly with replacement, we generate single-channel audio recordings of multiple individual voices mixed together. Each voice within a specific mixture is unique, but a source voice may be repeated in multiple mixtures.

Before mixing, we attenuate each individual voice proportionally to the number of voices to be mixed to ensure the mixture audio does not clip. We create different datasets of mixtures with different numbers of contributing voices, ranging from just a single voice (v=1) to mixtures of up to ten individuals (v=10) speaking simultaneously. For each number of voices v=1 to 10, we generate 10,000 different mixtures for use as a training set. For each set, we stratified our sampling

[1] https://www.kaggle.com/c/tensorflow-speech-recognition-challenge

among the different proportions to achieve uniformity by response ratio. We repeated this same process to generate a test set (n=5000), drawing from the other source partition of single voices that were held out for testing.

C. Experimental Design

Because we know the precise proportion of yes and no in each of our crowd mixtures, we model our approach as a regression problem in which we predict the proportion of positive sentiment, as measured by the proportion of yes responses in the mixture. We must note that despite our overly simplistic design choice for the purpose of this study, linguistically, yes and no are polar particles that can be used interchangeably for positive and negative sentiment [20].

In all experiments, our models were trained using raw audio. Each 16-bit WAV file is broken into a sequence of 256 consecutive samples, which are fed to the input layer of the deep learning model. We use a batch size of 64 in all experiments. For comparison across our different experiments, we train all models for 20 epochs. From our training set, we partitioned 30% for use as a validation set. We used mean squared error (MSE) as the loss function. We evaluated our trained regression models using a held-out test set in which no voice appeared in the training set.

We compared different popular deep learning models including 1D-Convolution Neural Network (Conv1D), Gated Recurrent Unit (GRU), Long short-term memory (LSTM), bidirectional LSTM (BLSTM), and a stacked architecture of a Conv1D layer followed by recurrent layers (CNN+RNN). Based on preliminary results our chosen network architecture consists of an input layer with 256 units followed by three stacked BLSTM recurrent layers, each with 128 units in both directions. Each BLSTM layer uses a dropout rate of 0.3. The output from the final layer is fed to a fully connected dense layer, with a single unit with RELU activation. We use this model architecture for all our experiments.

IV. Results

We performed several experiments to explore the feasibility our proposed approach on this problem. First we evaluated our approach on the single voice response case as a baseline for comparison. Next, we evaluated our approach on mixtures of v=1 to 10 voices. Lastly, we merged our datasets and evaluated our approach when the number of voices responding varies.

A. Single Voice Baseline

Because we are training on raw audio, we make no attempt to use natural language understanding of our sounds. To better understand the potential of our model to interpret these words as mere sounds, we first performed an experiment on our one voice dataset. Because our generated dataset contains more mixtures than our source datasets have individual voices, the v=1 dataset was sampled with replacement and contains duplicate voices.

We trained our model on a training set of 10k examples and evaluated on a test set of 5k examples, both with equal proportions of yes and no responses. We were able to predict the positive sentiment (yes) of the individual responses with a correlation of Pearson's r=0.84. For one voice, we achieved an overall accuracy 0.91. We would expect a state of the art the speech transcription system to achieve near perfect recognition of common words such as "yes" and "no". Nevertheless, we consider this a promising result for our approach which does not rely on natural language understanding in its attempt to hear.

B. Varying the Number of Voices

Next, we trained models on our datasets of voices v=2 through v=10. We show these results as Fig. 1. Although we performed best with only v=2 voices (0.76), we observed a general consistency in our results even as the number of speakers increased. We find this a very encouraging result that speaks to the potential for our approach to scale to much larger mixtures of voices, an important prerequisite for real-time deployment.

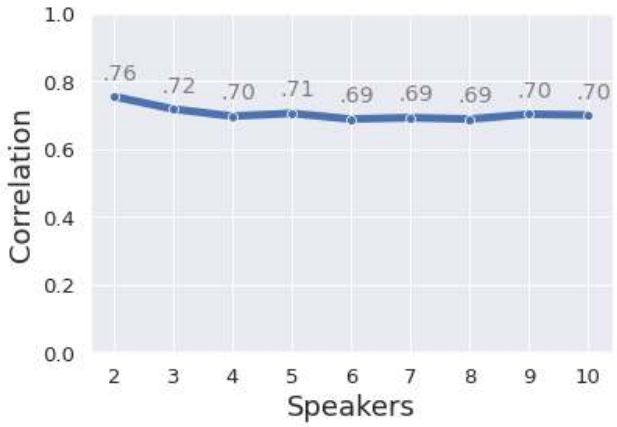


Fig. 1. Pearson's correlation by the number of speakers.

C. Mixtures up to 10 Voices

In this experiment we combined our nine datasets v=2 through v=10 together. We exclude v=1 since it contains a single voice and is not a mixture. Each of the nine individual training sets contain 10,000 example and each of test sets contain 5,000 examples. The combined set of mixtures results in n=90,000 training examples and n =45,000 testing examples.

On this combined set of mixtures with a variable number of speakers, we achieved a Pearson's correlation of r=0.70. We find it encouraging that our system yields the same correlation in this experiment as we achieved with our hardest subset of v=10 voices. This finding speaks to the scalability of our system to an unknown number of voices. We show a discretized view of this prediction distribution as Fig. 2.

Because the true rate varies depending on the number of voices in the mixture, we have discretized for better visual comparison to our v=10 voice experiment. We also include a histogram of the distribution of true rates. Since we merged datasets v=2 to 10, and unlike our previous experiments, this rate is not equally distributed.

In our mixed set of varying number of voices, there are more examples of all no responses (0.0) and all yes responses (1.0) because these examples appeared in all of the individual datasets for voices v=2 to 10. Despite the presence of outliers, we are encouraged that the interquartile ranges for these cases

of full consensus are lower than the cases of contention. Over the set of all 45,000 test examples, we note an average interquartile range of 0.48. Although this is a large range, especially if we seek to predict the exact rate of agreement, it suggests that it is feasible to perhaps predict broad categories of crowd sentiment, such as low, medium, and high.

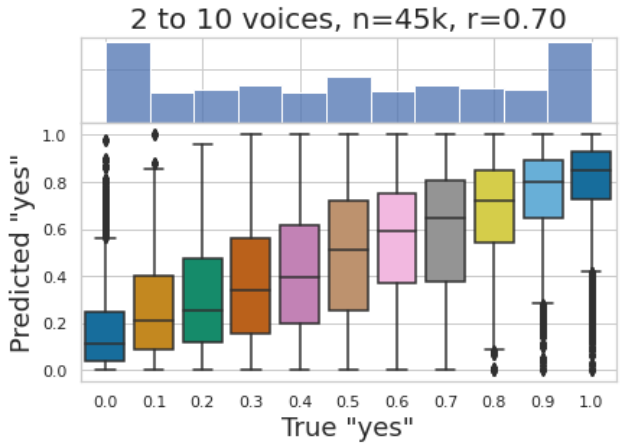


Fig. 2. Distribution of errors with a histogram of showing number of examples by sentiment rate for the mixed set.

In Fig. 3 we show a normalized confusion matrix of these results, with discretized predictions for each of the 11 possible true positive response rates. We annotated with the rate of predicted accuracy along the diagonal which represents correct predictions. This view highlights our system's success in predicting the cases in which a majority of speakers are in agreement (lower-left and upper-right), while showing our system's difficulties in precisely predicting cases in which the speakers are split in their agreement (middle).

Across this large dataset of mixtures with varying number of voices we were able to predict majority sentiment with an accuracy rate of 0.77. We observe that we are able to predict majority positive sentiment better than we predicted majority negative sentiment. Furthermore, we achieved a relative balance between correct rates of no (0.75) and yes (0.80). This finding suggests that training with examples with varying numbers of voices can improve robustness across all rates of agreement.

V. Discussion

We present a novel educational use case that seeks to determine the rate of sentiment from crowd audio with a high number of overlapping speakers. We began this investigation of this problem on the simple case of "yes" and "no" responses. We presented a machine listening approach designed to estimate the rate of sentiment across the group, rather than to classify the overall majority sentiment. We demonstrate the potential of a this approach to estimate sentiment from crowd audio with results relatively consistent as the number of speakers increased.

A. Main Finding

In summary, our machine listening approach was able to estimate the rate of positive sentiment across a large dataset of crowd audio mixtures with a varying number of overlapping speakers (Pearson's r=0.70). Furthermore, we were able to predict the majority sentiment with 77% accuracy. We observed that our system appears to scale as we increased the number of overlapping speakers from 2 to 10 voices (average r = 0.71 ± 0.02). We achieved these results using a validation paradigm in which we evaluated performance on a dataset of held-out voices. This increases our confidence in the generalizability of our approach to new speakers and differing mixtures of voices.

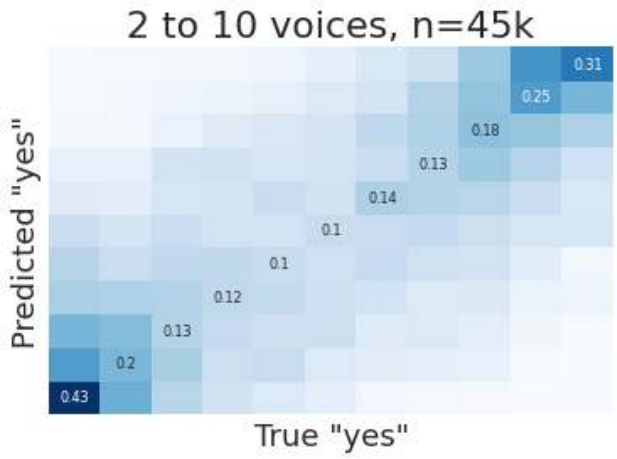


Fig. 3. Heatmap showing the discretized prediction rate where the diagonal represents the rate of perfect prediction.

Although we do not have a baseline of previous work to which we compare our results, we find these results encouraging and consider our approach worthy of further investigation and refinement. We are optimistic that our present results reflect the feasibility of a machine listening approach to determine the rate of sentiment in crowd mixtures of raw audio.

B. Contributions

This work advances understanding of the task of determining sentiment analysis from crowd audio. We present a novel research question and provide an initial approach to attempt to solve it. We consider the role of our system in a hypothetical future real-world mix-modality classroom response application, and we design our approach for efficiency in order to provide predictions with minimal lag in near real-time. We focus on a deep learning approach directly from raw audio to reduce the latency of our inference procedure by bypassing time-consuming feature extraction and engineering steps.

We train and evaluate our models on mixtures originating from two separate sets of voices. By validating our approach independently of speaker, we design our approach to better generalize to other speakers. Lastly, we design our approach as regression problem measuring the rate of sentiment, rather than as a classification problem estimating the majority sentiment. We make this choice to jump directly to this more difficult experimental question because we seek to attempt to hear all voices in the mixture, not just the loudest or closest voice.

C. Limitations and Future Work

In this work we introduce a new problem domain and propose a deep learning approach in an attempt to solve it. Because of the novelty of this work, there are a number of limitations which motivate our future work.

In this study, we only considered the binary case of yes/no responses. Given the simplicity of these two words, we consider this a good starting point to measure positive and negative sentiment of crowd audio. However, because of this limitation, we do not yet know how well our approach will scale to different responses. As future work, we need to identify our data sources that would allow us to expand our consideration from single words to sets of words carrying the same semantic meaning (e.g. *yes* and *yeah*). Importantly, we need to investigate more complex response patterns, such as *A-B-C-D* or *1-2-3-4,* to determine if our approach is able to scale beyond *yes-no*.

As future work we intend to explore the effect of distractor responses. For example, if we are measuring yes-no responses, what is the effect if the response contains a notable proportion of neutral responses that are neither yes nor no, such as a student responding "*what*?" to a question? We will also investigate effects of adding generated noise to the audio mixture to help ensure better robustness to the recording environment that might be expected in real-world conditions.

Although we consistently achieved moderate correlations of our predictions with the true rate of sentiment in the mixture across our different test sets, we also observed many instances of outliers in which our models greatly mispredicted the sentiment ratio. We will examine these specific examples, identifying and analyzing any temporal and spectral features that might be causing our system to treat them differently. Additionally, we will continue to explore different deep learning model architectures to reduce the error in these mispredictions.

Given the novelty of this approach, we do not have a baseline of comparison to inform how humans would perform on this task. As future work, we intend to design a human subject study in which listeners are presented with examples of mixed responses and are asked to judge a majority positive or negative sentiment. With such a dataset, we could begin to compare our system against humans. Lastly, we will explore approaches to count the number of individual voices in the mixture, in addition to estimating the sentiment rate.

D. Applications

There are many real-world scenarios that may benefit from a machine listening system that is able to interpret crowd sentiment. For example, attractions such as amusement parks may wish to quantify user experience by park goers' vocal reactions, or perhaps to identify individuals in discomfort or distress in order to take appropriate actions.

Sports broadcaster may desire a system that can quickly interpret the crowd's reaction after a major play. Comedians or performers may wish to measure audience reaction and engagement in real-time so they can alter their act accordingly. And lastly, movie theaters screening new films may wish to measure audience reactions in an environment in which low lighting makes computer vision infeasible.

E. Concluding Remarks

We anticipate a near future in which machine listening devices are as commonplace as computer vision systems. We hope that our presentation of a new problem area and details of our initial approach will lay a strong foundation for future work.

There is much work to be done to better understand this difficult research problem. With this initial baseline of work exploring the simple "yes" or "no" dichotomy, we look to inspire researchers to continue our emphasis on an equitable, rather than majority, understanding of crowd sentiment.

Automatically determining specific rates of sentiment from many overlapping speakers is important to ensure that every voice is heard, especially when the listener is not human.

REFERENCES

[1] S. Van Damme, G. Crombez, and C. Spence, "Is visual dominance modulated by the threat value of visual and auditory stimuli?," *Experimental Brain Research*, vol. 193, no. 2, pp. 197–204, 2009.

[2] C. Klinge, B. Röder, and C. Büchel, "Increased amygdala activation to emotional auditory stimuli in the blind," *Brain*, vol. 133, no. 6, pp. 1729–1736, 2010.

[3] S. Popham, D. Boebinger, D. P. Ellis, H. Kawahara, and J. H. McDermott, "Inharmonic speech reveals the role of harmonicity in the cocktail party problem," *Nature communications*, vol. 9, no. 1, pp. 1–13, 2018.

[4] K. B. Bhangale and K. Mohanaprasad, "A review on speech processing using machine learning paradigm," *International Journal of Speech Technology*, vol. 24, no. 2, pp. 367–388, 2021.

[5] T. J. Park, N. Kanda, D. Dimitriadis, K. J. Han, S. Watanabe, and S. Narayanan, "A review of speaker diarization: Recent advances with deep learning," *Computer Speech & Language*, vol. 72, p. 101317, 2022.

[6] L. Bullock, H. Bredin, and L. P. Garcia-Perera, "Overlap-aware diarization: Resegmentation using neural end-to-end overlapped speech detection," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7114–7118.

[7] H. H. Mao, S. Li, J. McAuley, and G. Cottrell, "Speech recognition and multi-speaker diarization of long conversations," *arXiv preprint arXiv:2005.08072*, 2020.

[8] A. O. Hogg, C. Evers, and P. A. Naylor, "Multiple hypothesis tracking for overlapping speaker segmentation," in *2019 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 2019, pp. 195–199.

[9] V. Andrei, H. Cucu, and C. Burileanu, "Overlapped Speech Detection and Competing Speaker Counting—Humans Versus Deep Learning," *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, no. 4, pp. 850–862, 2019.

[10] D. M. Schuller and B. W. Schuller, "A review on five recent and near-future developments in computational processing of emotion in the human voice," *Emotion Review*, vol. 13, no. 1, pp. 44–50, 2021.

[11] N. He and S. Ferguson, "Multi-view neural networks for raw audio-based music emotion recognition," in *2020 IEEE International Symposium on Multimedia (ISM)*, 2020, pp. 168–172.

[12] J. Weston, R. Lenain, U. Meepegama, and E. Fristed, "Learning de-identified representations of prosody from raw audio," in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, vol. 139, pp. 11134–11145. [Online]. Available: https://proceedings.mlr.press/v139/weston21a.html

[13] Y. Jia and S. SungChu, "A deep learning system for sentiment analysis of service calls," *arXiv preprint arXiv:2004.10320*, 2020.

[14] S. Angadi and R. Venkata Siva Reddy, “Survey on sentiment analysis from affective multimodal content,” in *Smart intelligent computing and applications*, Springer, 2019, pp. 599–607.

[15] R. R. Iyer, R. Zheng, Y. Li, and K. Sycara, “Event outcome prediction using sentiment analysis and crowd wisdom in microblog feeds,” *arXiv preprint arXiv:1912.05066*, 2019.

[16] X. Wu, Q. Ye, H. Hong, and Y. Li, “Stock selection model based on machine learning with wisdom of experts and crowds,” *IEEE Intelligent Systems*, vol. 35, no. 2, pp. 54–64, 2020.

[17] R. Jeffrey, P. Bian, F. Ji, and P. Sweetser, “The Wisdom of the Gaming Crowd,” in *Extended Abstracts of the 2020 Annual Symposium on Computer-Human Interaction in Play*, 2020, pp. 272–276.

[18] A. Kruspe, J. Kersten, and F. Klan, “Detection of informative tweets in crisis events,” *Natural Hazards and Earth System Sciences (NHESS)*, 2021.

[19] S. Z. Hassan, K. Ahmad, A. Al-Fuqaha, and N. Conci, “Sentiment analysis from images of natural disasters,” in *International Conference on Image Analysis and Processing*, 2019, pp. 104–113.

[20] D. Goodhue and M. Wagner, “Intonation, yes and no,” *Glossa: a journal of general linguistics*, vol. 3, no. 1, 2018.