Crowdsourcing Breakthroughs for Language Technology Applications
AAAI Technical Report WS-15-24

# It's Not Just What You Say, But How You Say It: Muiltimodal Sentiment Analysis Via Crowdsourcing

**Ahmad Elsehnawy,[1] Steele Carter,[1] Daniele Braga[2]**
University of Washington, Seattle, WA
Voicebox Technologies, Bellevue, WA
aelshen@uw.edu, steele42@uw.edu, danielab@voicebox.com

## Abstract

This paper examines the effect of various modalities of expression on the reliability of crowdsourced sentiment polarity judgments. A novel corpus of YouTube video reviews was created, and sentiment judgments were obtained via Amazon Mechanical Turk. We created a system for isolating text, video, and audio modalities from YouTube videos to ensure that annotators could only see the particular modality or modalities being evaluated. Reliability of judgments was assessed using Fleiss Kappa inter-annotator agreement values. We found that the audio only modality produced the most reliable judgments for video fragments and that across modalities video fragments are less ambiguous than full videos.

***Index Terms*** — sentiment analysis, crowdsourcing, multimodal databases, inter-annotator agreement techniques

## Introduction

The last six years have seen the rise of social media giants such as Facebook, Twitter, Reddit and a number of other growing-yet-more-popular social media platforms, and with the proliferation of internet access to a young and tech-savvy public, people are relying on these platforms for almost all of their day-to-day information and decisions. It is for this reason that researchers are focusing energy into studying blogs, microblogs and written reviews.

However, this focus is growing counter-intuitive, as outlets like YouTube, Skype and Vine have become household names. People are beginning to use more than just their keyboards to express themselves to one another, but research has not been keeping abreast of the changes.

YouTube reports that 100 hours of video are uploaded to YouTube every minute[1]. It is fast becoming a bottomless wellspring of content and information, accessed by millions of people everyday. These videos contain valuable sentiment information, however very little research has been done towards discovering how to extract this information.

Video reviews offer a number of advantages over their written relatives, with users able to see smiles, frowns, furrowed brows, or to hear sarcasm, shouting, sighs or impressions. There are a number of paralinguistic events present in video reviews that consumers pick up on to inform themselves, and these are features that would be invaluable for computational models, if only for the mountain of ambiguities in human communication that cannot be teased apart with only text. Or so we believe, and this is what we set out to prove with the following experiment.

In this paper we set out to evaluate reliability of sentiment polarity judgments for the video, text, and audio modalities of YouTube video reviews. We did this by crowdsourcing judgments of video reviews that had been restricted to a particular modality or modalities and evaluating agreement between annotators.

## Prior Work

In truth, when it comes to analyzing video in the field of sentiment analysis, there hasn't been much work. Typical work in the field centers on text, with recent years bringing more and more attention to audio analysis as well. There are a number of reasons why video has yet to receive much attention. For one, video analysis relies on technologies that are still underdeveloped. The state of the art technologies for facial recognition and eye tracking are not as reliable as comparably advanced systems for textual and audio analyses. Furthermore, visual features belong to a set

[1] YouTube statistics: http://www.youtube.com/yt/press/statistics.html

of paralinguistic features that many computational linguists have thus far neglected in research.

However, video analysis has not been entirely neglected by researchers. One of the more comprehensive studies on the matter was conducted by Morency et al (2011), wherein the authors worked to create a multimodal classifier to analyze the sentiment of a corpus of YouTube video reviews, building a joint model of text, audio and video features. Morency et al. demonstrate in the paper that different features correlate well with different types of sentiment polarity.

Lexical features were found to be good at predicting if a word will be polar, but too many words used in reviews are neutral or sentimentally vacuous. Two audio features, pitch variation and pause duration, indicate strong polarity and neutrality respectively. Video features, smiles and look away, were suggestive of positive and non-positive respectively.

Morency et al then created a classifier utilizing all of the above features, and they found that a trimodal feature set, accounting for lexicon, pitch, pause, smile and look away achieved the greatest results by a significant margin. All unimodal approaches were bested by the trimodal classifier, proving that the next big steps in sentiment analysis will be in the inclusion, combination and analysis of paralinguistic features.

There is one main shortcoming from Morency et al (2011), and that is that they do not successfully demonstrate that a trimodal approach is significantly more accurate than any particular bimodal approaches. It is no small feat to build a feature set of the audio or video features they discuss in the paper, and we are not shown whether or not the use of both of those modalities is the most valuable of all the combinations. Perhaps the inclusion of video features does not notably improve the results of a text-audio classifier, or that audio truly does enhance the results of a text-video classifier. This is one of the goals we set out with when conducting this experiment.

Another important contribution to the topic of sentiment analysis of video content comes from Pérez-Rosas and Mihalcea (2013), who conducted an experiment to see how crowdsourced transcriptions of video product reviews gathered from ExpoTv.com compared against automatic transcriptions. Our takeaways from Pérez-Rosas and Mihalcea (2013) are twofold.

One, Pérez-Rosas and Mihalcea demonstrate that it is possible to predict sentiment in video reviews using only transcriptions. This demonstrates that, yes, text-only analyses are valid for visual media. But more importantly: two, in their experiment, Pérez-Rosas and Mihalcea compare how their classifiers fared with ExpoTv.com video reviews and Amazon.com written reviews.

They found that the system they had built performed notably worse for video reviews than for written reviews, showing that to successfully analyze the sentiment of expressions from visual media, novel means of analysis and experimentation are needed beyond the state-of-the-art that is in place for textual analysis.

## Dataset

In order to carry out this experiment, we would need access to a corpus of YouTube videos. Morency et al (2007) have a corpus of YouTube video reviews available, but that dataset is insufficient for the needs of this experiment. The dataset we needed had to have the following: a reviewer speaking to and facing the camera, transcriptions, annotations for sentiment, and timestamp annotations breaking the videos down into a number of smaller fragments. To our knowledge, no such dataset is readily available, and so we created it ourselves. A list of the URLs for videos used along with all of our spam filtering code can be found on GitHub[2].

Our research found that the most standard format of video reviews on YouTube came from user book reviews, wherein the reviewers were constantly on-screen, facing and directly addressing the camera, and featured as single reviewers without others visible on-screen. Since it was not our goal to build a fully featured automatic classifier, we were not concerned about studying multiple domains, as the purview of this experiment was only to study how different modalities can influence a human observer's interpretation of information.

The final dataset compiled for this experiment was composed of 20 YouTube videos, ranging from 3 to 5 minutes in length. 12 of the videos featured female reviewers, and 8 were male. The videos all featured relatively young reviewers (i.e. teens to mid-thirties), typically of Caucasian ethnicity, and all videos were entirely in English. The majority of the videos were spoken with American English accents, but there were a small number of UK English speakers in the dataset. We had originally set out to have all videos be of the same dialect of English, but we found that within the domain of YouTube book reviews, there was a disproportionately high number of female reviewers. To try and keep a gender-balanced dataset, we had to include English speakers of another dialect region. Finally, the 20 videos resulted in 110 video fragments, ranging in duration from 10 to 40 seconds.

[2] https://github.com/aelshen/575-Project.git

## Methodology

For this experiment, we chose to crowdsource the data collection. It was a central goal of this experiment to utilize as large a sample size as we could manage in order to address the larger questions at hand. In order to see if additional modalities truly do contribute significant features when it comes to sentiment analysis, we need to see how they impact the decisions of human judges. This would require access to a large subject pool, only obtainable from crowdsourcing platforms.

### Crowdsourcing Platform

Crowdsourcing has quickly been gaining attention from speech and NLP researchers over the last few years. It is an affordable way for researchers to engage with a multitude of subjects or to process large amounts of difficult data.

Much research has gone into determining the validity of crowdsourcing applications in speech and NLP research: Parent & Eskenazi (2011) examine the value of crowdsourcing in speech research, and ultimately conclude that crowdsourcing is indeed a valuable tool to the speech community; Parson et al (2013) demonstrate that crowdsourcing is a valuable means of collecting meaningful and relevant data from workers; Pérez-Rosas and Mihalcea (2013) successfully argue that crowdsourced transcriptions are comparable, if not superior to automatic transcriptions; and Mellebeek et al (2010) show that crowdsourced worker annotations have a high inter-annotator agreement with expert annotations.

For this experiment, Amazon Mechanical Turk (hereby referred to as MTurk) was used to gather data, chosen over other competing crowdsourcing platforms for flexibility in design and ease of quality control management.

### HIT Design

Amazon defines a HIT (Human Intelligence Task) as a single, self-contained task that a worker can work on, submit an answer, and collect a reward for completing[3].
As discussed at the head of the paper, this experiment seeks to study how multiple different modalities can influence human sentiment judgments, and to examine how a given modality will perform on utterance-level clips vs. full video reviews.

The four modalities in question are Text, Audio, Video, and Audio/Video. Text is the modality of the written word, where a worker is given a piece of writing to analyze. Audio covers spoken language, represented by a sound-only clip from a review. Video refers exclusively to visual data, to the exclusion of any and all text and sound. The final modality is Audio/Video, a combination of spoken-language and visual data.

With the above in mind, we needed to design and deploy 8 unique HITs to the platform. One set of HITs for each of the four modalities we wanted to study to examine its usefulness in judging sentiment at the utterance, i.e. fragment, level, and another set of HITs to do so at the full-length, i.e. video, level.

For the first experiment, hereby referred to as the Fragments experiment, a worker was presented with a set of 5 fragments, randomly selected from our corpus of 110 review fragments, of the current modality. For each fragment, the worker was asked to read/listen/watch, and then assign a sentiment score of 1 to 5, with 1 being the most negative and 5 the most positive. For text, this meant 5 pieces of text presented on the page. For Audio, we designed the HIT to play the audio of the fragment. This was accomplished exclusively through the use of YouTube's JavaScript API. For Video, workers were given five YouTube videos with automatically muted volume. YouTube player controls were selectively disabled for this task so that workers would be unable to increase the volume. The final modality, Audio/Video, had 5 unaltered video clips embedded into the task.

For all of the Fragment HITs, JavaScript code was implemented specifically so that workers would be unable to watch the video beyond the specific fragment presented to them. If a worker attempted to scrub the video beyond the timestamp for the given fragment, the video would reset to the start point and pause automatically. Upon completion, each fragment video would reset to the beginning and pause allowing the worker to easily replay the clip.

For the video-level experiment, hereby referred to as the Full experiment, workers were instead given one review, but consisting of the full length of the review. For text, this meant reading the entire transcription of the review. For Audio, Video and Audio/Video, workers were given the entire video to consume, consistent with the formats explained above.

For all HITs, workers were given clear instructions as to the rules and expectations of the task, and they were asked to complete a very brief survey, asking for basic demographics like gender, age group and country of residence.

For both the Fragment and Full experiments, workers were paid $0.15 per assignment, and before they were able to do our work, they must have had an MTurk approval rating greater than or equal to 95% (i.e. their work on MTurk having been accepted at least 95% of the time) and they must have submitted at least 500 HITs through the platform. These settings were insufficient for the Audio/Video Fragment experiment. Due to time concerns, we were forced to increase the pay to $0.25 per assignment

[3] Mechanical Turk FAQ: https://www.mturk.com/mturk/help?helpPage=overview

and remove the aforementioned worker restrictions in order to attract enough annotators. This change likely impacted our results due to an effect shown by Gneezy and Rustichini (2000) where increased pay can attract greater amounts of spammers. With our extensive quality control, we were not particularly concerned about spam problems for this experiment, and our work found that all of the tasks were beset by significant amounts of spam, no matter the worker restrictions or the pay.

We collected 10 judgments per Fragment/Video, taking advantage of crowdsourcing's broad and affordable work pool to collect as many judgments as possible per video item. Typical sentiment analysis experiments usually only take the majority of three judgments when annotating for sentiment. On a final note, the HITs were available to workers from any country with a large English-speaking population.

## Quality Control

A lot of the research on crowdsourcing has focused specifically on the idea of quality control. Buchholz and Latorre (2011) examine how crowdsourced data can be vulnerable to spammers, and they outline numerous ways that researchers can conduct quality control to weed out this illegitimate data. Their results show that spam-filtered crowdsourced data produces high quality results, but they caution that researchers must find a balance between rejecting spammers and accidentally rejecting legitimate workers. Any filtering measures we are to apply would need to find this balance between lenience and strictness. Similarly, Parent and Eskenazi (2010) write that crowdsourced data with proper quality control provides the highest-quality transcriptions.

The quality control measure we chose to implement was actually a system with four phases of spam detection.

The first phase of spam detection involved checking how much time a worker spent on a task. Because we are working with YouTube videos that have quantifiable durations, we could easily check to make sure that the time a worker spent on a task was greater than or equal to the duration of the video(s) they were asked to watch. For the Text experiments, this was not the case. Since people are capable of reading at a variety of speeds, it was not our place to say how quickly a worker should be able to read a given text sample. So, we arbitrarily chose a threshold of 20 seconds for all Text experiments (Fragment and Full), figuring that 20 seconds was short enough a time that speed-readers would not be rejected but still catch a number of spammers.

The second phase of spam detection was for comparing worker transcriptions of a video. For the modalities that involved audio (e.g. Audio and Audio/video), we asked users to provide a partial transcription of the first 10 words of the video(s) in the task. For the Fragment HITs, since no set of 5 fragments ever includes the same video more than once, we check to see that 5 unique transcriptions are provided. We also check to make sure that no given transcriptions are left blank, and that all transcriptions are no shorter than 20 characters in length. Lastly, we manually check to make sure that all partial transcriptions are legitimate by comparing them to our own hand-transcriptions for the videos. The process was similar for the Full experiment, but we only checked for one transcription, since there is only one video per task.

The third phase involved comparing against Golden HITs. A Golden HIT is a gold standard answer that, if a user gets incorrect, lets us know that the submission is suspicious. An example of a Golden HIT for the Full experiment is a video where the reviewer assigns a score of 5/5 stars. It would be difficult to argue a situation where 5/5 could be construed as a neutral or negative score, so we check to make sure that a worker submits a score greater than 3. An example for the Fragment experiment would be a fragment where the reviewer says something like, "I absolutely hated the author's writing. It was bland to the point of being offensive." Such an utterance is unambiguously negative, so if a worker assigned a score greater than or equal to 3, we rejected the work. All Golden HITs were selected by hand from all the videos and video fragments, with care given to ensure that selected Golden HITs were unambiguous and did not express sentiment of more than one polarity. On a final note, we specifically check for "greater than or equal to 3" or "less than or equal to 3" so that a spammer who arbitrarily selects 3 for every video/fragment does not get accepted as legitimate data.

The last phase involved checking a worker's submission to the average score of other crowdsource workers. For instance, if the worker average score for a video was 4.3, suggesting a positive video, and the current worker submitted a score of 1, we chose to reject the data. We arbitrarily chose that if a worker's score deviated from the average by a margin of 3, the worker be flagged as a spammer.

For all HITs, a submission was flagged as spam if the assignment was left incomplete, meaning the worker did not complete the pre-survey, or they did not leave a value judgment for all videos in the assignment.

We decided that for the Video-only experiments, applying the last two phases of spam detection (comparing against Golden HITs and comparing against average score) was unreasonable due to the difficulty of deriving objective sentiment without linguistic content. Therefore the method of spam detection for the Video-only experiments was checking the duration of time spent on the task. This decision may have had an unwanted impact on our results.

## Demographics

As far as worker demographics go, we reached a fair gender balance, where 51% of workers were female and 49% male. 40% of non-spam workers said they were from the US, 44% from India, and the remaining 16% scattered across a number of other countries like Canada, the UK and Germany.

Interestingly, 51% of spam was found as coming from India, 29% from the US, and 10% of spammers reported no location at all, with the remainder once again distributed among a number of other countries.

These location demographics show that including India in one's crowdsourced research is a double-edged sword. India accounts for a large amount of workers, and any researcher hoping to complete a task in any timely sort of manner needs to include it. If we had not included India as a worker-candidate, the task would have taken an unacceptably long time to complete. However, India also accounts for a significant percentage of spam, meaning researchers will need to be careful when including it in experiments.

## Results

The main goal of this experiment is to compare reliability of sentiment judgments across different modalities. In order to evaluate and compare consistency of judgments we used Fleiss Kappa inter-annotator agreement (Fleiss 1973). Fleiss Kappa is a statistical measure of reliability of judgments between multiple annotators, which attempts to correct for chance agreement. A Kappa value of 1 indicates perfect agreement and a value of 0 indicates agreement no better than chance.

Table 1 and Figure 1 show the Fleiss Kappa values for each of the experiments both before and after spam filtering. These results show a consistent improvement in agreement after spam filtering. This suggests that our spam filtering methods were effective for enhancing the quality of judgments. Also, fragment experiments consistently yield higher Kappa values than their full video counterparts. This suggests greater ambiguity is introduced when evaluating an entire video.

The text only experiments yielded results on par with the audio/video experiments. This contradicts our initial intuition that reducing video reviews to their transcriptions introduces ambiguity. This may be due to increased spam for the audio/video task as a result of the difference in pay and worker restrictions. The increase in spam is shown by a much greater disparity in Kappa values before and after spam filtering for the audio/video task. It is possible that this increase in spam had an effect even after filtering.

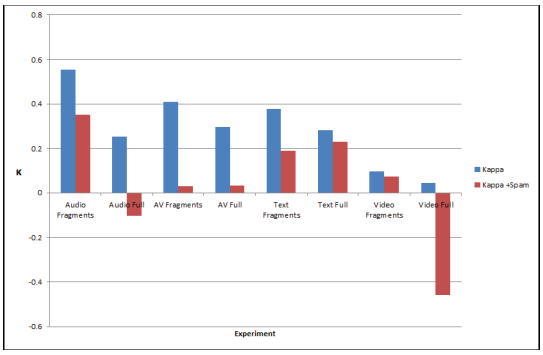


*Figure 1: Kappa Values*

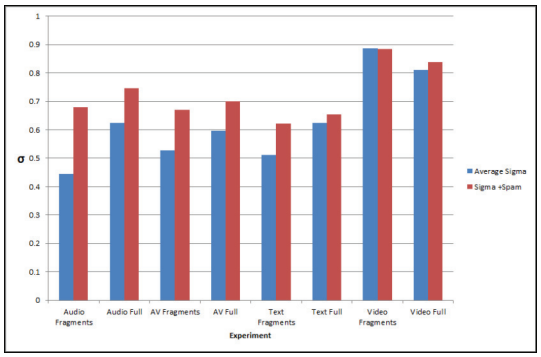


*Figure 2:Sigma Values*

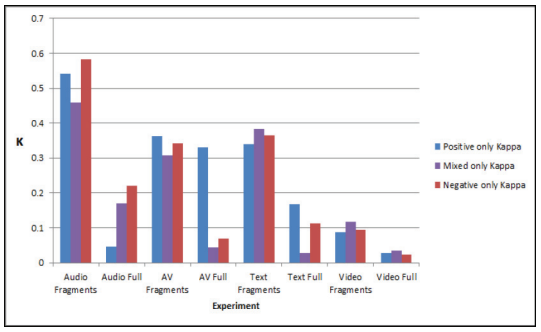


*Figure 3:Kappa by Polarity*

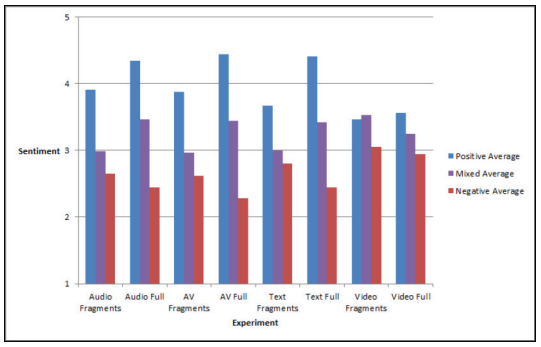


*Figure 4:Sentiment Average by Polarity*

| Spam Technique | Percentage of Spam |
|---|---|
| Time | 49.7 |
| Transcription | 16.7 |
| Golden HITs | 17.7 |
| Incomplete Assignment | 15.7 |
| Worker Average | 0.2 |

*Table 1: Quality Control results, reporting the percentage of spam submissions that each method of spam detection caught. As of writing, a total of 396 spam submissions were caught by our system. To put that in context, experimental results were compiled using ~1600 non-spam worker submissions.*

| Experiment | Audio Fragments | Audio Full | AV Fragments | AV Full | Text Fragments | Text Full | Video Fragments | Video Full |
|---|---|---|---|---|---|---|---|---|
| Fleiss Kappa | 0.554 | 0.254 | 0.408 | 0.299 | 0.376 | 0.283 | 0.098 | 0.045 |
| Kappa with Spam | 0.352 | -0.102 | 0.029 | 0.034 | 0.190 | 0.231 | 0.073 | -0.456 |

*Table 2: Fleiss Kappa evaluations*

As expected, the Video only experiment produced the lowest Kappa values, suggesting that it is very difficult for human annotators to derive objective sentiment using only gestures and facial expressions. It is important to note that this result may have been reinforced by our inability to apply the same spam filtering for the video experiment.

We were surprised to that the audio only tasks yielded the highest Kappa values, surpassing the audio/video experiment. We assumed that removing video would introduce more ambiguity, however it is possible that removing the video creates less ambiguity. Again, it is also possible that this difference is a result of the increased pay and lowered restrictions we were forced to use for the audio/video task. It is also important to note that the audio-only experiment for full videos had less agreement than its audio/video and text counterparts. This weakens the conclusion that the audio-only modality is the most reliable.

For comparison to Kappa values, we have also included standard deviation values for each experiment in Figure 2, denoted by Sigma. The Sigma values mirror the relationships shown by the Kappa values, but are inverted. There is greater deviation when Kappa values are lower and there is less agreement. However there is a notable difference for a number of the experiments before spam filtering. The standard deviation values for the full video-only task in particular remain very close before and after spam filtering, however the corresponding Kappa values increase dramatically after spam filtering. This is due to the fact that Kappa corrects for chance agreement whereas Sigma does not. Also, across all experiments there tends to be a greater change in Kappa values proportionally after spam filtering than for Sigma. This shows that Kappa values are a better indicator of the presence or absence of spam than standard deviation.

## Discussion and Conclusion

In this paper we assessed and compared the reliability of crowdsourced sentiment judgments for video reviews across different modalities. Using a small YouTube video review dataset we used crowdsourced sentiment judgments to evaluate and compare the ability of human annotators to objectively discern sentiment. Contrary to expectation our results suggest that reducing video reviews to only text transcriptions does not reduce the reliability of judgments, however removing the video modality and leaving only audio increases reliability. We do not feel particularly confident in these findings given the small size of the dataset and difficulties controlling variables such as spam filtering and worker pay between experiments.

Something we did not consider was how audio and video cues can be distinct cross-culturally. We tried to maintain a homogeneous dataset, but we were not as strict in selecting our workers. If we filter out workers that are not from the US, we find that the text only modality in fact yields the most reliable judgments. However this result is unreliable due to inconsistent demographics between tasks and due to a significantly reduced sample size. Going forward, it would be wise to select workers from within the same cultural group/region. An examination on how different cultures and languages may be sensitive to different modalities, would make for very interesting research in the future.

We also introduced a novel method for isolating modalities of video reviews so that they can easily be evaluated by crowdsourced workers. We implemented a thorough spam detection system to ensure reliability of evaluations. Our experiments confirmed the necessity of good quality control, filtering out 396 spam submissions and our results showed that our spam filtering techniques were very effective at increasing reliability of judgments in all modalities.

As a trade off for these spam filtering techniques, we were unable to use consistent spam prevention techniques across modalities. In the future we hope to control for this difference by using a single spam prevention method that works for all modalities while attempting to maintain high quality evaluations. This may be possible through the use of a screening test to ensure quality workers with a native proficiency for English that can consistently match Gold judgments. We may even find that workers can be trained to more effectively detect sentiment from the video only modality if we only allow workers into the task that are able to correctly evaluate sentiment in a pretest.

In the future we hope to find more reliable patterns distinguishing the different experiments by increasing the size of the corpus. We believe that our data set was too small and noisy to support many reliable conclusions. While we found that Audio was particularly effective at identifying negative sentiment, we suspect that there are more patterns such as this, which will emerge with the use of a larger corpus with more judgments.

We also plan to implement another type of experiment, which gives evaluators text transcriptions alongside muted video. This experiment would isolate out the audio channel while preserving the linguistic channel.

Future work may also include analysis of how different genders perceive sentiment; for instance, examining if audio features are more important for the sentiment judgments of female viewers, if males rely more heavily on visual cues, if female viewers pick up sentiment more accurately when presented with a female speaker or a male speaker, and vice-versa.

## Acknowledgments

We would like to thank Gina Levow for her advice and guidance throughout the development of our experiment. We would also like to thank Jeff Cairns for pointing us in the right direction with regard to the HTML and JavaScript involved in HIT design.

## References

Bucholz, S., & Latorre, J. 2011. Crowdsourcing Preference Tests and How to Detect Cheating. In Proceedings of the 12th Annual Conference of the International Speech Communication Association, 3053-3056. Florence, Italy: Interspeech.

Fleiss, J. L., and Cohen J. 1973. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educational and psychological measurement* 33(3): 613-619.

Gneezy, U., & Rustichini, A. 2010. Pay Enough or Don't Pay at All. *The Quarterly Journal of Economics* 115(3): 791-810.

Mellebeek, B., Benavent, F., Grivolla, J., Codina, J., Costa-Jussa, M. R., & Banchs, R. 2010. Opinion Mining of Spanish Customer Comments with Non-Expert Annotations on Mechanical Turk. In Proceedings of the NAACL HLT 2010 Workshop on Creating speech and language data with Amazon's mechanical turk, 114-121. Los Angeles, California: ACL.

Morency, L. P., Mihalcea, R., & Doshi, P. 2011. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web. In Proceedings of the 13th ACM International Conference on Multimodal Interaction, 169-176. Alicante, Spain: ICMI.

Parent, G., & Eskenazi, M. 2010. Toward Better Crowdsourced Transcription: Transcription of a Year of the Let's Go Bus Information System Data. In Spoken Language Technology Workshop of the Institute of Electrical and Electronics Engineers, 312-317. San Diego, California: IEEE.

Parson, J., Braga, D., Tjalve, M., & Oh, J. 2013. Evaluating Voice Quality and Speech Synthesis Using Crowdsourcing. In Proceedings of the 16th International Conference of Text, Speech, and Dialogue, 233-240. Pilsen, Czech Republic: TSD.

Pérez-Rosas, V., & Mihalcea, R. 2013. Sentiment Analysis of . In Proceedings of the 14th Annual Conference of the International Speech Communication Association, 862-866. Lyon, France: Interspeech.