Studying the Impact of Multimodality in Sentiment Analysis

Ahmad Elshenawy
Steele Carter

Goals/Motivation

- How are judgments influenced by different modalities?
- Compare sentiment contributions of different modalities
- Use Interannotator agreement to measure objectivity of sentiment and ease of judgment
- Observe how results change for fine grained judgments of review chunks
- Assess effectiveness of crowdsourcing sentiment from video transcriptions rather than actual videos

Background/prior work

- Very little work on sentiment in speech and video
- Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web (Morency et al)
  - Built sentiment classifiers using features from 3 different modalities:
    - Text
    - Audio
    - Video
  - Created YouTube corpus of video reviews
  - Found that integrating all 3 modalities yields best performance

More Background/prior work

- Sentiment analysis of online spoken reviews (Pérez-Rosas and Mihalcea 2013)
    - Compared classifiers built on manual vs automatic transcriptions of video reviews from ExpoTv
        - Found that it is possible to predict sentiment of videos using only transcriptions
        - Found that manual transcriptions (MTurk) yield better classification results than automatic
    - Also compared video reviews to written Amazon reviews
        - Written reviews easier to classify

Corpus

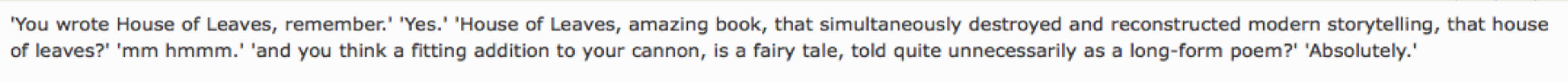
- We created our own corpus of Youtube video reviews, consisting of 3-5 minute long book reviews.
- Originally 35 videos were found and analyzed, but the experiment uses only 20 videos.
    - corpus reduced primarily due to cost concerns
    - 6 positive, 6 negative, 8 neutral
- Originally video transcriptions were obtained via crowdsourcing
    - was way too slow, and way too expensive

Annotation

- Transcribed each video by hand
  - Labeled disfluencies (um, er, etc.)
- Also labeled our own evaluations of sentiment for comparison and spam filtering
- Added timestamps dividing transcriptions into chunks

Modalities - Text

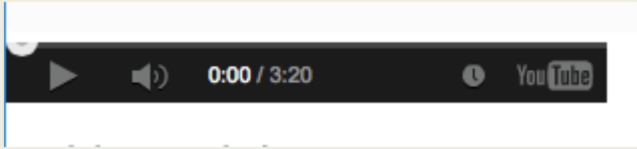
- **Text only**: typical in sentiment analysis, workers are given only a piece of text.
    - Text is the most frequently studied modality in the field of sentiment analysis.
    - Problem with text-only analysis is that words can be ambiguous, and sometimes writing cannot convey tones/expressions effectively.



'You wrote House of Leaves, remember.' 'Yes.' 'House of Leaves, amazing book, that simultaneously destroyed and reconstructed modern storytelling, that house of leaves?' 'mm hmmm.' 'and you think a fitting addition to your cannon, is a fairy tale, told quite unnecessarily as a long-form poem?' 'Absolutely.'

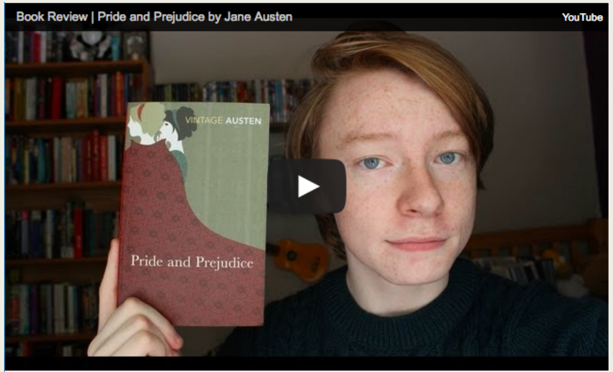
Modalities - Audio

- **Audio only**: workers are given an audio-only piece of the review.
  - Next to text, audio is the next most commonly studied aspect of sentiment analysis.
  - Audio can help resolve some of the ambiguities and misunderstandings that text-only approaches present.



Modalities - Video

- **Video only**: workers are given a video piece of the review where the video is muted, and they are given no option to increase the volume.
  - Very little work has been done on video modalities.
  - We included this experiment just to see how facial expressions and physical gestures can convey sentiment.



Modalities - Audio/Video

- **Audio/Video**: workers are given unaltered videos to watch.
  - Videos contain both audio and video.
  - As mentioned in the previous slide, video in sentiment analysis is a novel field.
  - Our presumptive nominee for the strongest category out of all modalities examined.

Experiments

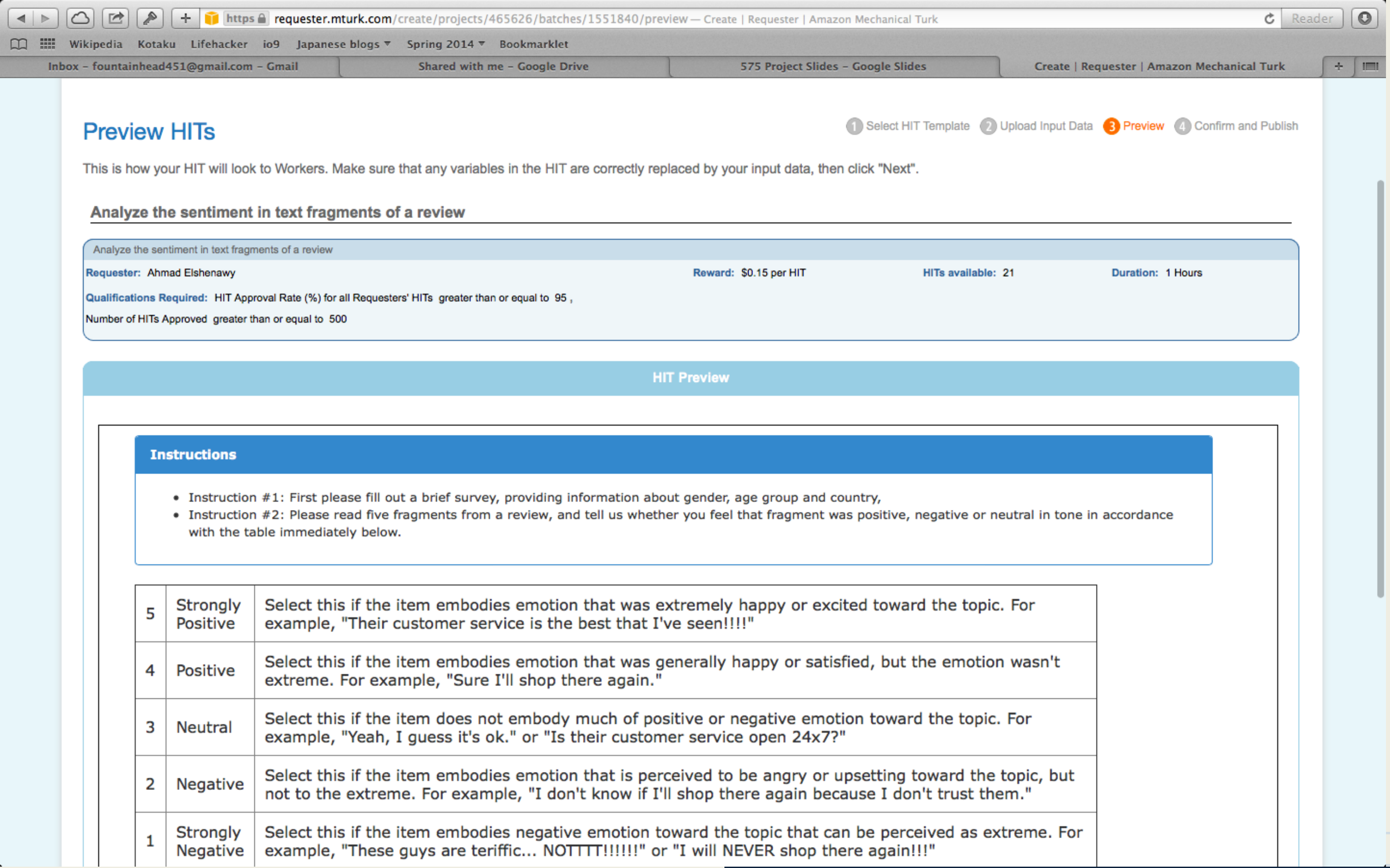
- Two experiments were conducted:
- Video Chunks
  - Videos were annotated with timestamps, breaking up videos into ~20-30 second chunks, typically also demarcating new topics within the review.
  - A HIT was designed where workers are presented with 5 of these chunks, and asked to judge the sentiment of that chunk.
  - Our 20-video corpus resulted in about 110 chunks.

Experiments - Cont’d

- Full video
  - Videos were annotated with timestamps, but this time they were left whole.
  - A HIT was designed where workers are presented with 1 whole review (an entire video, or the transcription of an entire video)
  - 20 videos were included

HIT Design

- Experiment ended up needing 8 Mechanical Turk HITs.
  - One set of HITs for each modality.
    - Text only, audio only, video only, audio/video
  - One set of HITs for chunks vs whole reviews
- Required **a lot** of javascript and HTML coding
- Collected 10 judgments per video/fragment, paying about $0.15 per task*.
  - 20 video HITs per modality
  - 21 5-chunk HITs per modality


https requester.mturk.com/create/projects/465626/batches/1551840/preview — Create | Requester | Amazon Mechanical Turk
Reader
Wikipedia
Kotaku
Lifehacker
io9
Japanese blogs
Spring 2014
Bookmarklet
Inbox – fountainhead451@gmail.com – Gmail
Shared with me – Google Drive
575 Project Slides – Google Slides
Create | Requester | Amazon Mechanical Turk
1 Select HIT Template
2 Upload Input Data
3 Preview
4 Confirm and Publish
HIT Preview

Preview HITs

This is how your HIT will look to Workers. Make sure that any variables in the HIT are correctly replaced by your input data, then click "Next".

Analyze the sentiment in text fragments of a review

Analyze the sentiment in text fragments of a review

Requester: Ahmad Elshenawy **Reward:** $0.15 per HIT **HITs available:** 21 **Duration:** 1 Hours

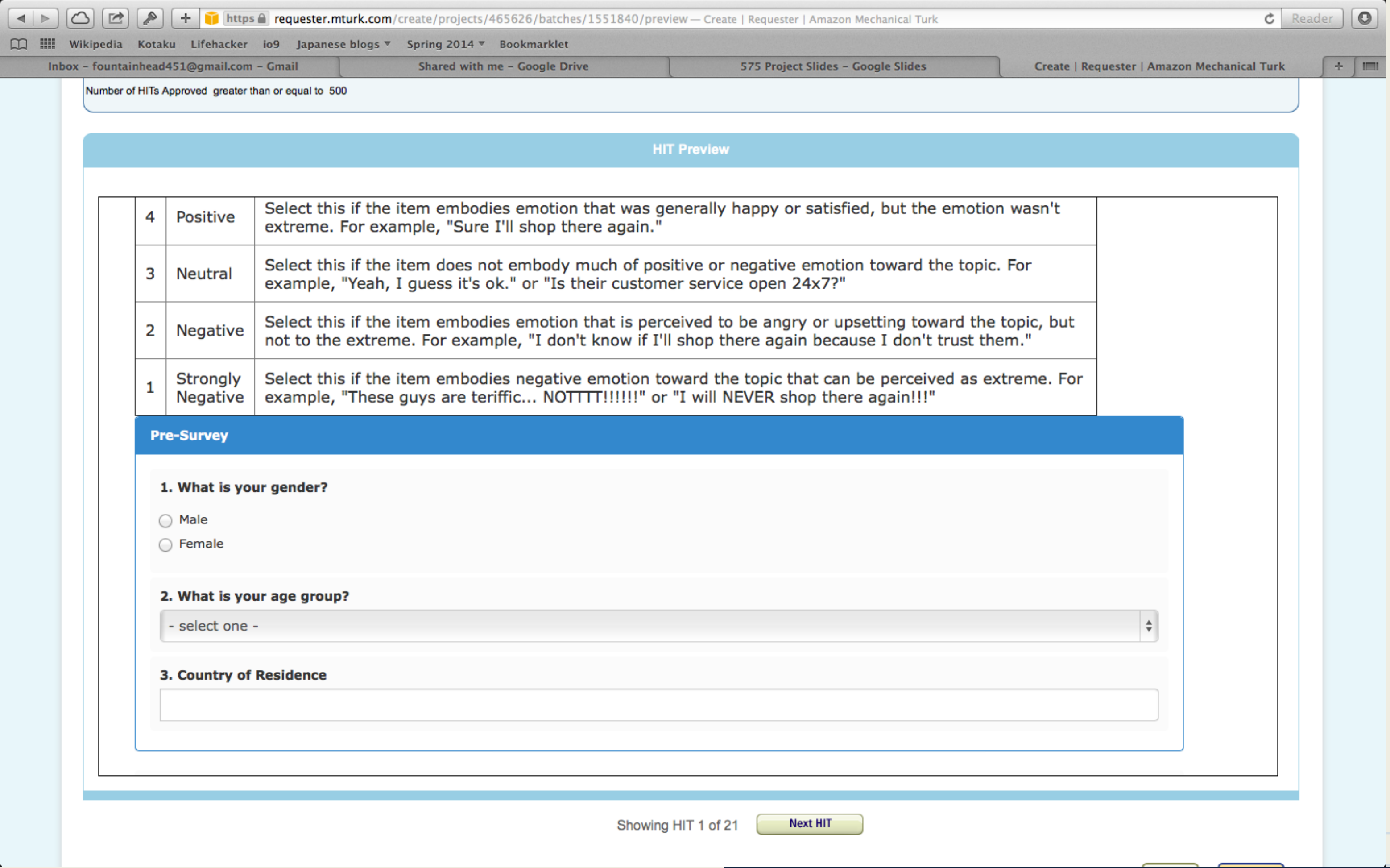
Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 ,
Number of HITs Approved greater than or equal to 500

Instructions

- Instruction #1: First please fill out a brief survey, providing information about gender, age group and country,
- Instruction #2: Please read five fragments from a review, and tell us whether you feel that fragment was positive, negative or neutral in tone in accordance with the table immediately below.

5	Strongly Positive	Select this if the item embodies emotion that was extremely happy or excited toward the topic. For example, "Their customer service is the best that I've seen!!!!"
4	Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
3	Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
2	Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
1	Strongly Negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are teriffic... NOTTTT!!!!!!" or "I will NEVER shop there again!!!"

Instructions


https requester.mturk.com/create/projects/465626/batches/1551840/preview — Create | Requester | Amazon Mechanical Turk
Reader
Wikipedia
Kotaku
Lifehacker
io9
Japanese blogs
Spring 2014
Bookmarklet
Inbox – fountainhead451@gmail.com – Gmail
Shared with me – Google Drive
575 Project Slides – Google Slides
Create | Requester | Amazon Mechanical Turk

Number of HITs Approved greater than or equal to 500

HIT Preview

4	Positive	Select this if the item embodies emotion that was generally happy or satisfied, but the emotion wasn't extreme. For example, "Sure I'll shop there again."
3	Neutral	Select this if the item does not embody much of positive or negative emotion toward the topic. For example, "Yeah, I guess it's ok." or "Is their customer service open 24x7?"
2	Negative	Select this if the item embodies emotion that is perceived to be angry or upsetting toward the topic, but not to the extreme. For example, "I don't know if I'll shop there again because I don't trust them."
1	Strongly Negative	Select this if the item embodies negative emotion toward the topic that can be perceived as extreme. For example, "These guys are teriffic... NOTTTT!!!!!!" or "I will NEVER shop there again!!!"

Pre-Survey

1. What is your gender?

○ Male
○ Female

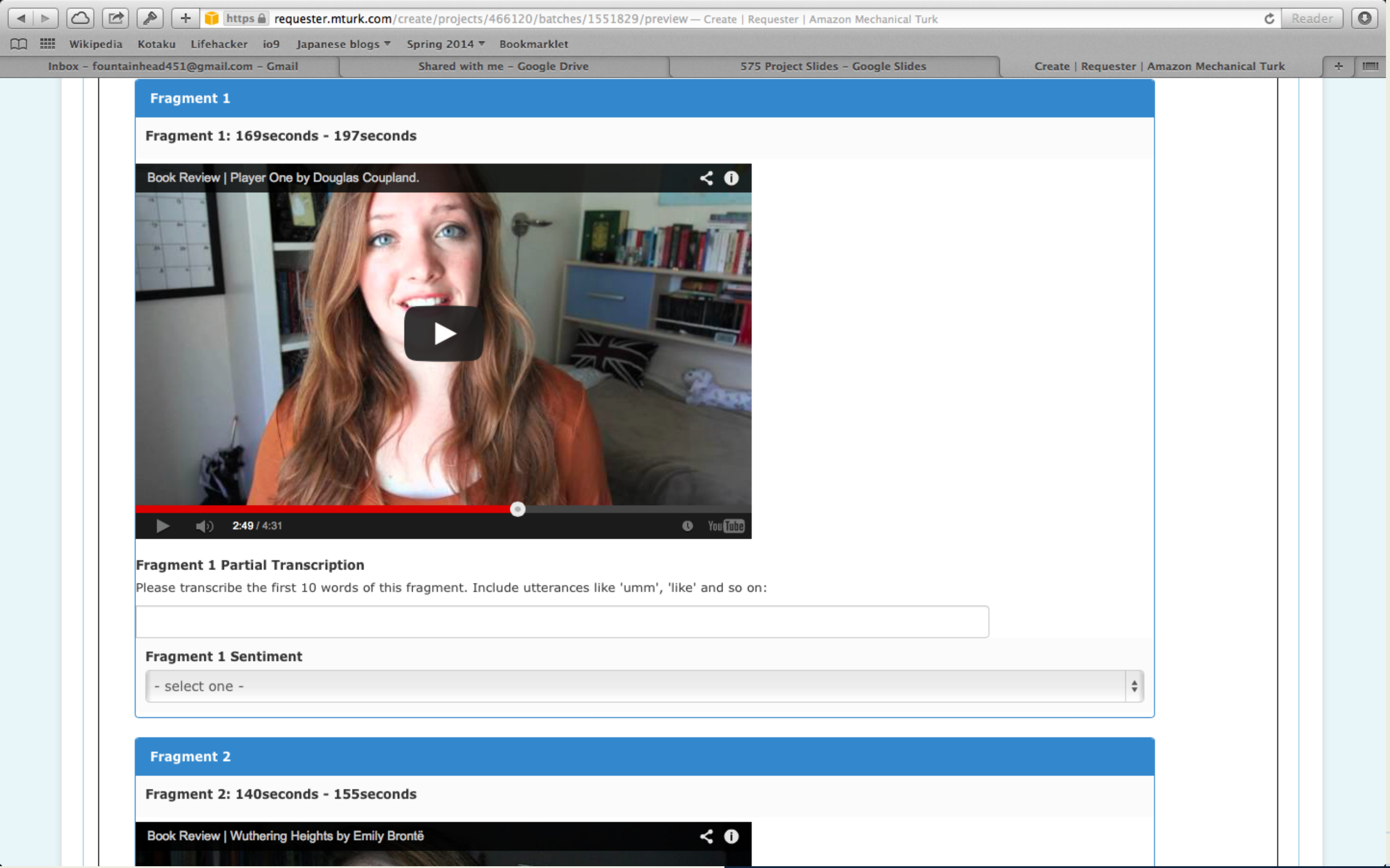
2. What is your age group?

- select one -

3. Country of Residence

Showing HIT 1 of 21

Pre-survey


https requester.mturk.com/create/projects/466120/batches/1551829/preview — Create | Requester | Amazon Mechanical Turk
Reader
Wikipedia
Kotaku
Lifehacker
io9
Japanese blogs
Spring 2014
Bookmarklet
Inbox – fountainhead451@gmail.com – Gmail
Shared with me – Google Drive
575 Project Slides – Google Slides
Create | Requester | Amazon Mechanical Turk
Book Review | Player One by Douglas Coupland.
2:49 / 4:31
YouTube
Book Review | Wuthering Heights by Emily Brontë

Fragment 1

Fragment 1: 169seconds - 197seconds

Fragment 1 Partial Transcription

Please transcribe the first 10 words of this fragment. Include utterances like 'umm', 'like' and so on:

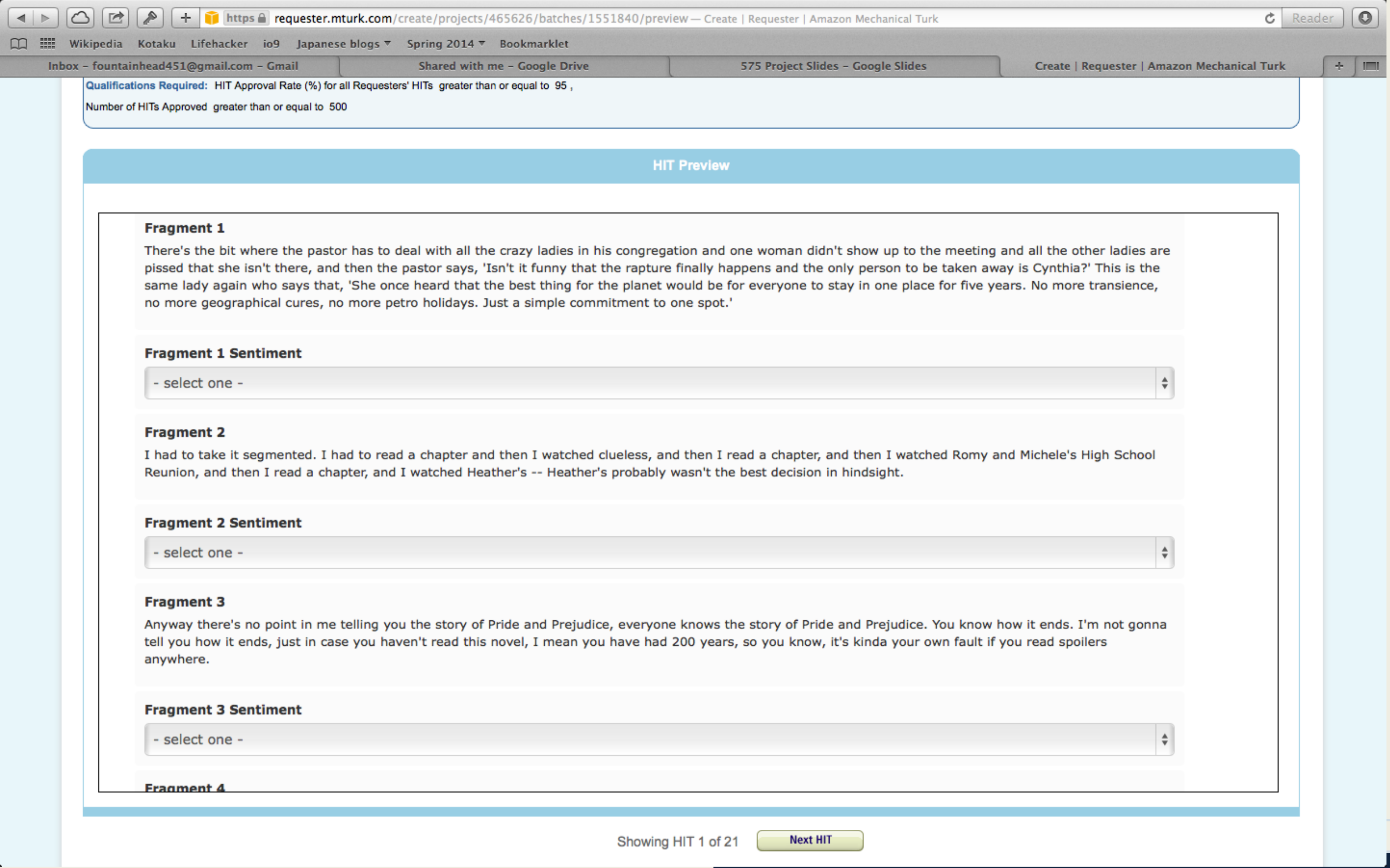
Fragment 1 Sentiment

- select one -

Fragment 2

Fragment 2: 140seconds - 155seconds

Example of an Audio/Video Chunk HIT



https requester.mturk.com/create/projects/465626/batches/1551840/preview — Create | Requester | Amazon Mechanical Turk
Reader
Wikipedia
Kotaku
Lifehacker
io9
Japanese blogs
Spring 2014
Bookmarklet
Inbox – fountainhead451@gmail.com – Gmail
Shared with me – Google Drive
575 Project Slides – Google Slides
Create | Requester | Amazon Mechanical Turk

Qualifications Required: HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95 ,

Number of HITs Approved greater than or equal to 500

HIT Preview

Fragment 1

There's the bit where the pastor has to deal with all the crazy ladies in his congregation and one woman didn't show up to the meeting and all the other ladies are pissed that she isn't there, and then the pastor says, 'Isn't it funny that the rapture finally happens and the only person to be taken away is Cynthia?' This is the same lady again who says that, 'She once heard that the best thing for the planet would be for everyone to stay in one place for five years. No more transience, no more geographical cures, no more petro holidays. Just a simple commitment to one spot.'

Fragment 1 Sentiment

- select one -

Fragment 2

I had to take it segmented. I had to read a chapter and then I watched clueless, and then I read a chapter, and then I watched Romy and Michele's High School Reunion, and then I read a chapter, and I watched Heather's -- Heather's probably wasn't the best decision in hindsight.

Fragment 2 Sentiment

- select one -

Fragment 3

Anyway there's no point in me telling you the story of Pride and Prejudice, everyone knows the story of Pride and Prejudice. You know how it ends. I'm not gonna tell you how it ends, just in case you haven't read this novel, I mean you have had 200 years, so you know, it's kinda your own fault if you read spoilers anywhere.

Fragment 3 Sentiment

- select one -

Fragment 4

Showing HIT 1 of 21 Next HIT

Example of a Text Chunk HIT

Why MTurk?

- HIT Design
    - As mentioned earlier, our HITs required a considerable amount of HTML and Javacript coding.
        - Had to utilize Youtube’s javascript API to modify all of our videos in the ways that we needed.
    - Crowdflower does not allow us the control and flexibility needed to develop such code.
- Spam Detection
    - We were overwhelmed with waves of spam.
        - People analyzing transcriptions of entire videos in only 10 seconds...
        - People watching and analyzing 210 seconds of video in 80 seconds...
        - In our first phase of spam filtering, 175+ HITs were flagged as spam out of ~470 total HITs submitted.
    - MTurk allows us control over who gets paid and who doesn’t.
    - MTurk allows requesters to submit a CSV with rejection notes, instead of rejecting one-by-one, facilitating ease of large-scale rejection
    - With CrowdFlower, we would have lost a large sum of money, time and results to spammers.
- **We would not have been able to conduct this same set of experiments on CrowdFlower.**

Spam detection/prevention

- Created a python script to automate the task of spam filtering.
    - With an expectation of ~1600 submissions for the whole project, we needed an efficient way of checking spam.
- Spam was filtered in 4 phases

Spam Filtering - Phase 1

- Checking Work Time (in seconds):
    - Proved to be the most effective method
    - For non-Text modalities:
        - For each Whole Video task, we calculated the length of the video in seconds.
        - For each Video Chunk task, we calculated the cumulative length of time in seconds for all 5 chunks in the HIT.
        - If the Work Time of the submission was less than the above calculated length, the HIT was rejected as spam.

Spam Filtering - Phase 1

- Checking Work Time (in seconds) - cont’d:
    - For Text modality:
        - Since some people can be legitimately fast readers, we couldn’t rely on time as heavily as the other modalities.
        - So, we arbitrarily decided on a threshold of 20 seconds.
            - If a text task (Full or Chunk) was completed in a time less than this, we decided to flag it as spam.
            - With further spam detection downstream, we felt this was the best we could do without being unreasonable.

Spam Filtering - Phase 2

- Check Transcriptions:
  - For HITs with audio, we asked workers to transcribe first 10 words of the video.
    - Video Chunks:
      - Check for five transcriptions.
        - If one or more transcriptions are left blank, flagged as spam.
      - Since no video fragment is repeated in a HIT, we checked for 5 unique transcriptions from the worker.
        - If less than 5 are found, flagged as spam.
      - Lastly, check to make sure the transcriptions are valid
        - If one or more transcriptions are shorter than 20 characters, flag as spam.

Spam Filtering - Phase 2

- Check Transcriptions - cont’d:
  - For HITs with audio, we asked workers to transcribe first 10 words of the video.
    - Full Video:
      - If the transcription was left blank, flagged as spam.
      - Lastly, check to make sure the transcription is valid
        - If the transcription is less than 20 characters, flagged as spam.
  - After running the script we also check, by hand, all the transcriptions and compare them to our hand-written transcriptions.
    - If they are incoherent or wildly different, flagged as spam.

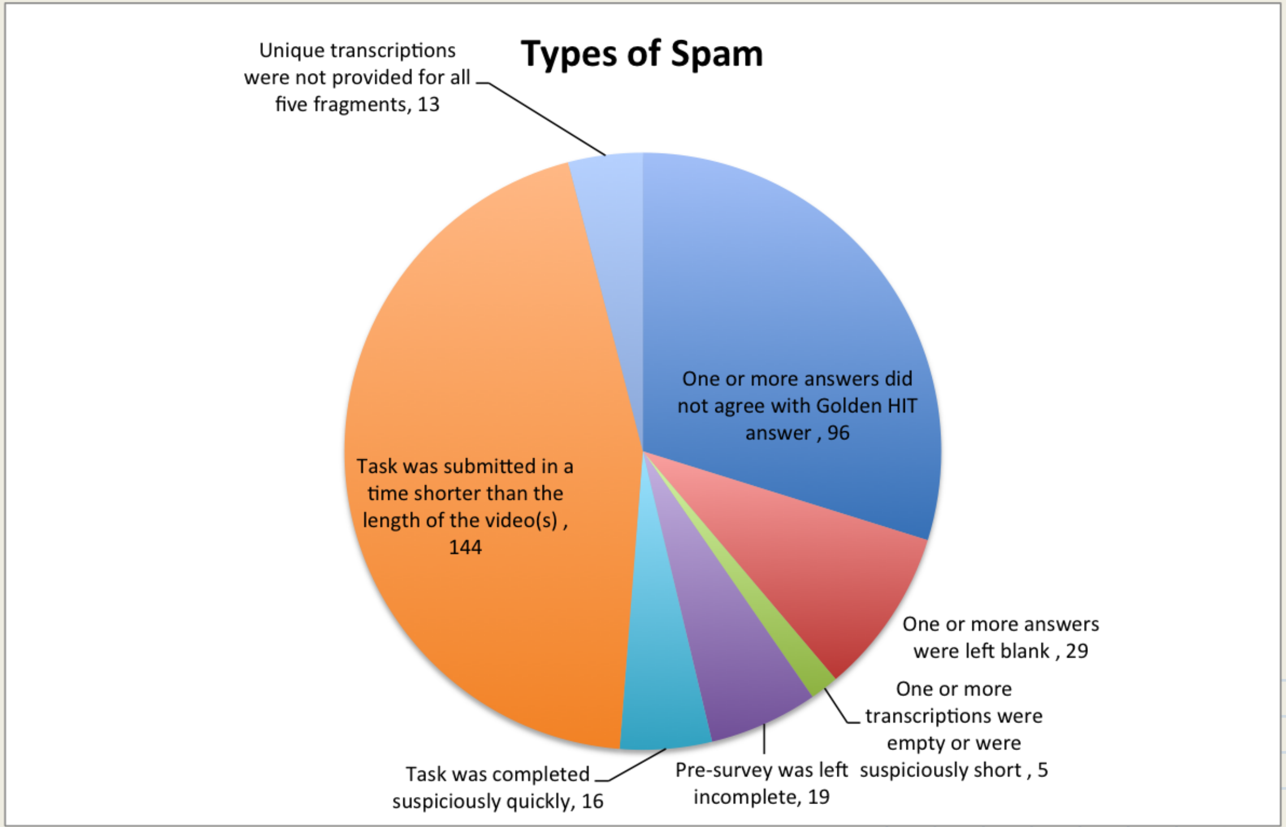
Spam Filtering - Phase 3

- Golden HITs:
    - A number of videos and video chunks that were obviously positive/negative in sentiment were selected to use as Golden HITs.
        - e.g. a blatantly positive video is given a score of 5.
            - if a submitted score is <=3 for this Golden HIT, flagged as spam
            - We specifically chose “less than or equal to” instead of just “less than” to catch people who may have arbitrarily chosen “3” as a score for all submissions.

Spam Filtering - Phase 4

- Comparing to average worker:
    - The average score for a video/fragment was calculated from all submissions that have so far passed spam filtering.
        - If a worker’s submission was significantly different from the average for all workers, it was flagged as spam.
            - e.g. if the average for a video was 4.3, and a worker submits a score of 1, we mark it as spam.
    - Seemed to be the least helpful of all four phases.

Spam Breakdown



Spam Issues

- We decided that the only spam filtering method that we could use on the Video-only HITs was Checking Time.
  - We knew that the Video-only tasks were exceptionally difficult, and decided Golden HITs and checking averages would not be fair ways to check for this task.
    - May have influenced results for this modality.

Issues

- Experiment required 8 different jobs
- Very expensive
- Some jobs were ignored by workers
    - accept/reject rates might be to blame
    - Lowered requirements and increased pay to compensate
    - Titles may have been confusing: (sentiment of audio/video clip vs just video clip)

Results

- ***Still in progress***

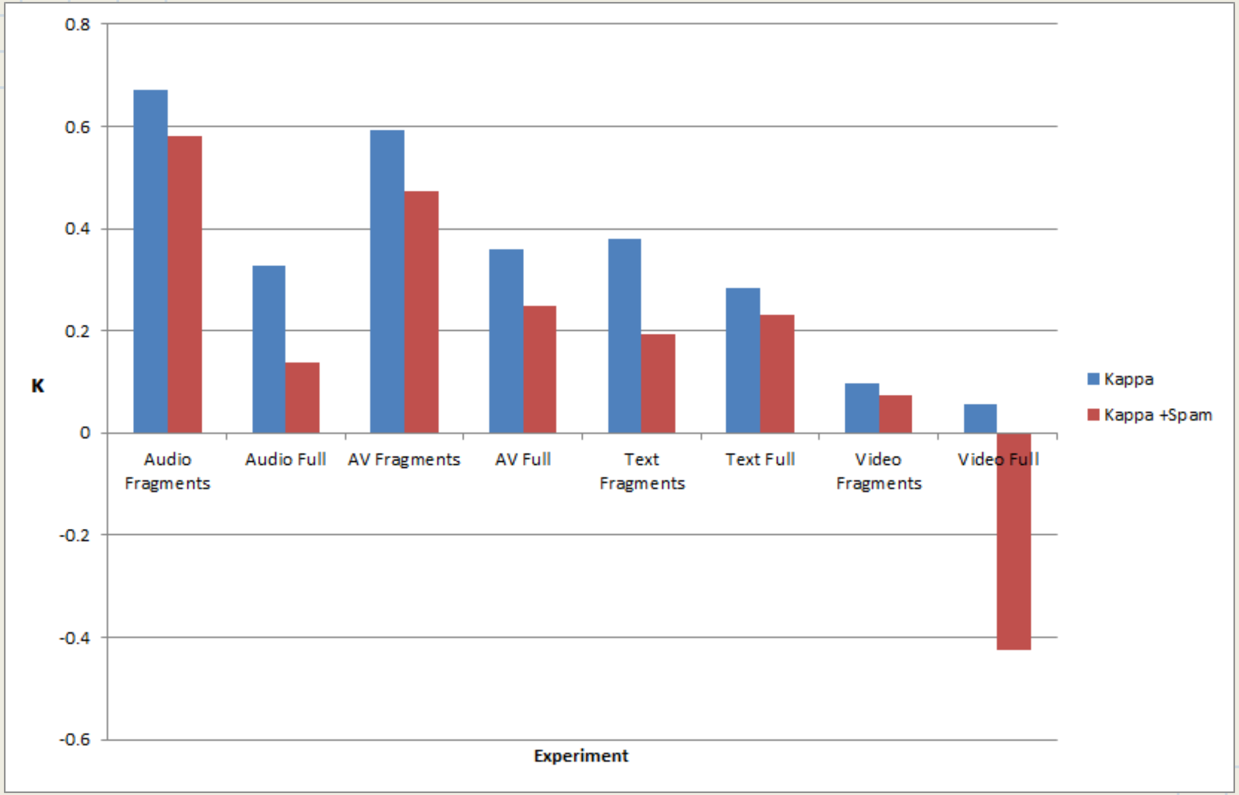
experiment	Audio Fragments	Audio Full	AV Fragments	AV Full	Text Fragments	Text Full	Video Fragments	Video Full
Fleiss Kappa*	0.6585229	0.32718761	0.59219439	0.35777352	0.38837062	0.31190628	0.09787617	0.05682245
Kappa+spam	0.5818058	0.13806713	0.47368365	0.24906424	0.19287932	0.23110002	0.07276448	-0.42413806
Avg. Sigma	0.3433623	0.61234192	0.42548755	0.58129920	0.50113399	0.56583090	0.88603217	0.80918976
Sigma+spam	0.4302887	0.70151700	0.54313308	0.63393201	0.62201485	0.65499547	0.88479075	0.83677014
Interfragment Sigma	0.5524046	N/A	0.57162278	N/A	0.56395743	N/A	0.34367873	N/A

* (Fleiss 1971)

Kappa by Polarity

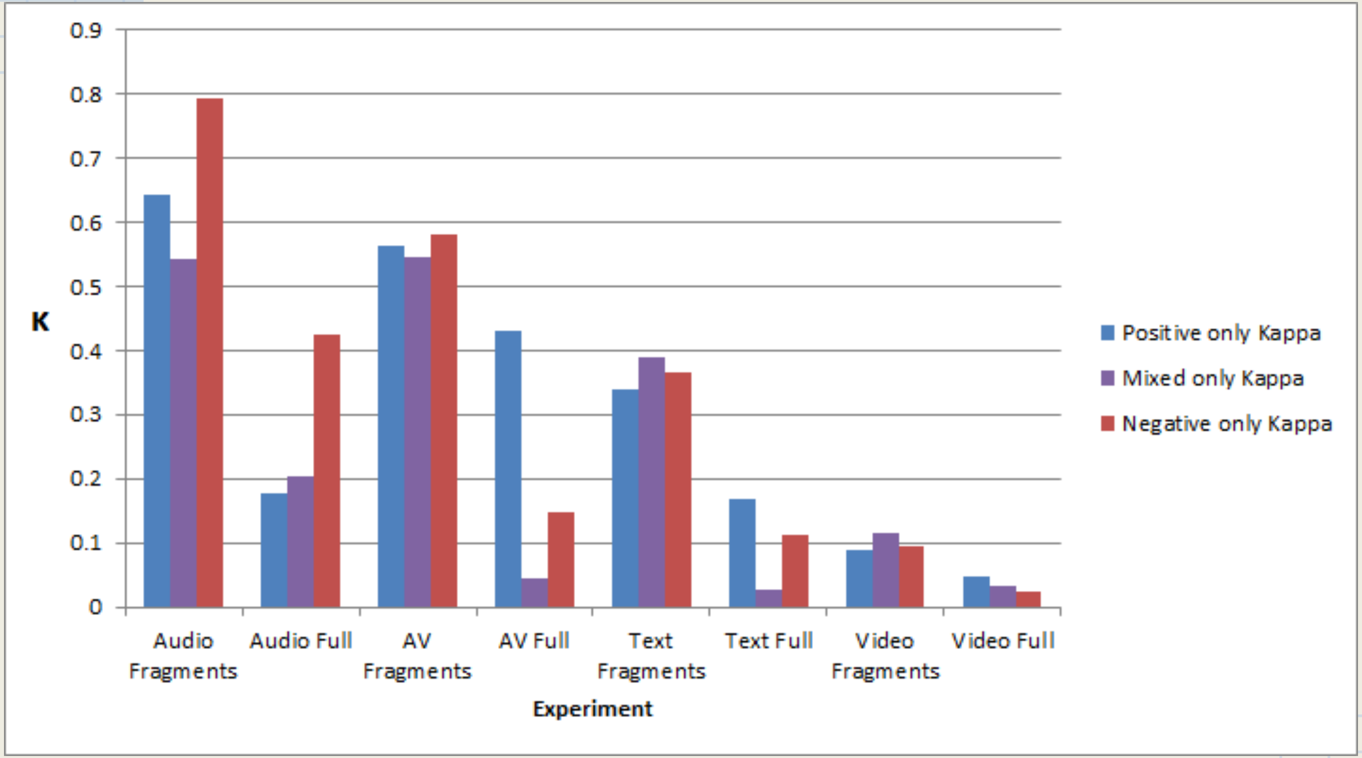
experiment	Audio Fragments	Audio Full	AV Fragments	AV Full	Text Fragments	Text Full	Video Fragments	Video Full
Positive only Kappa	0.6489054	0.17715364	0.59693590	0.43155888	0.36078757	0.26873225	0.08649687	0.04355832
Mixed only Kappa	0.5073978	0.20278203	0.51781328	0.04376120	0.38218785	0.02749375	0.11599278	0.03381642
Negative only Kappa	0.7918547	0.42564102	0.59446659	0.14628956	0.37527369	0.11280101	0.09395510	0.02311707

Fleiss Kappa



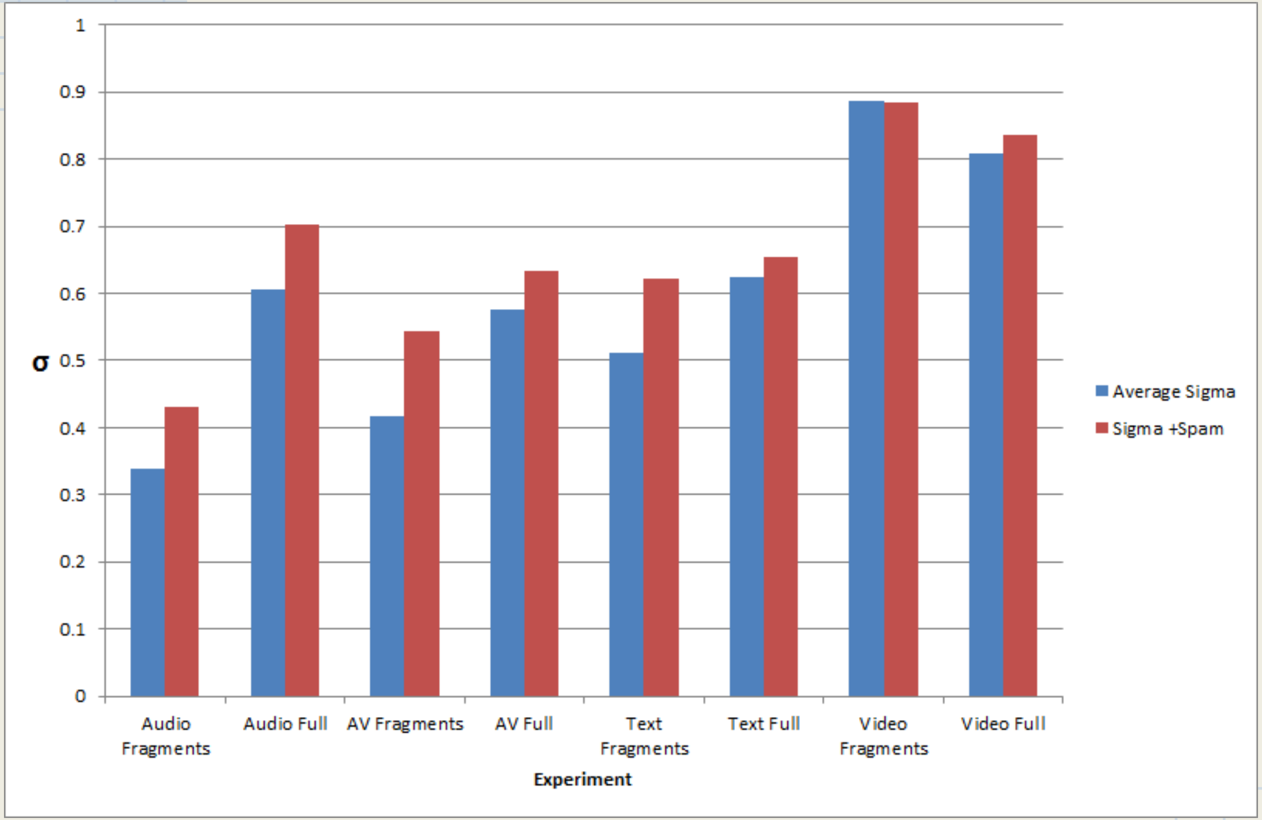
- Fragments show better agreement
- Audio and AV best agreement
- Video only very low agreement
- Spam filtering increases agreement
- Lower AV than Audio may be due to increased pay and lowered worker restrictions

Kappa by Polarity



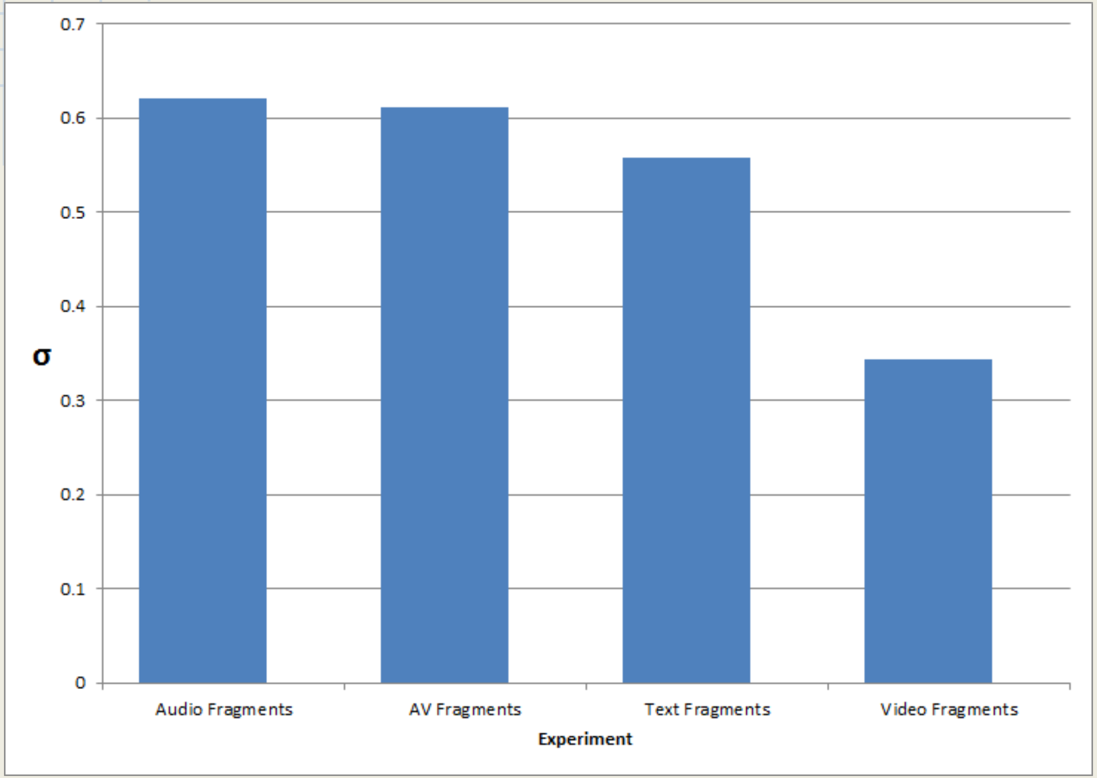
- Mixed tends to be lower
- High variability between experiments
- Negative audio less ambiguous

Standard Deviation



- Not adjusted for chance
- Same overall pattern as Kappa (inverted though)
- Note Kappa better reflects spam on Video only

Interfragment Deviation

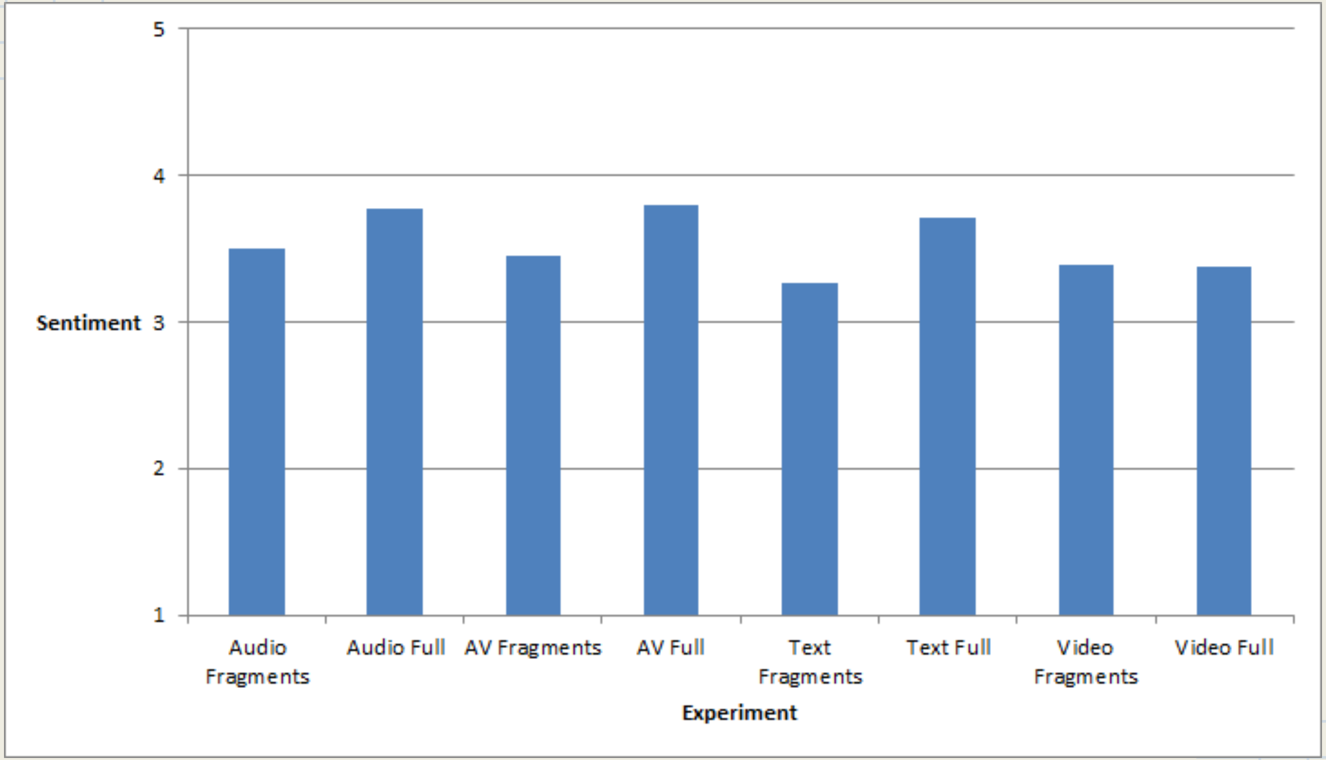


- Video only fragments agree more across one video
- Likely meaningless, (video only tends to always average out to 3)
- How to normalize for chance agreement with continuous numbers?

Average Sentiment

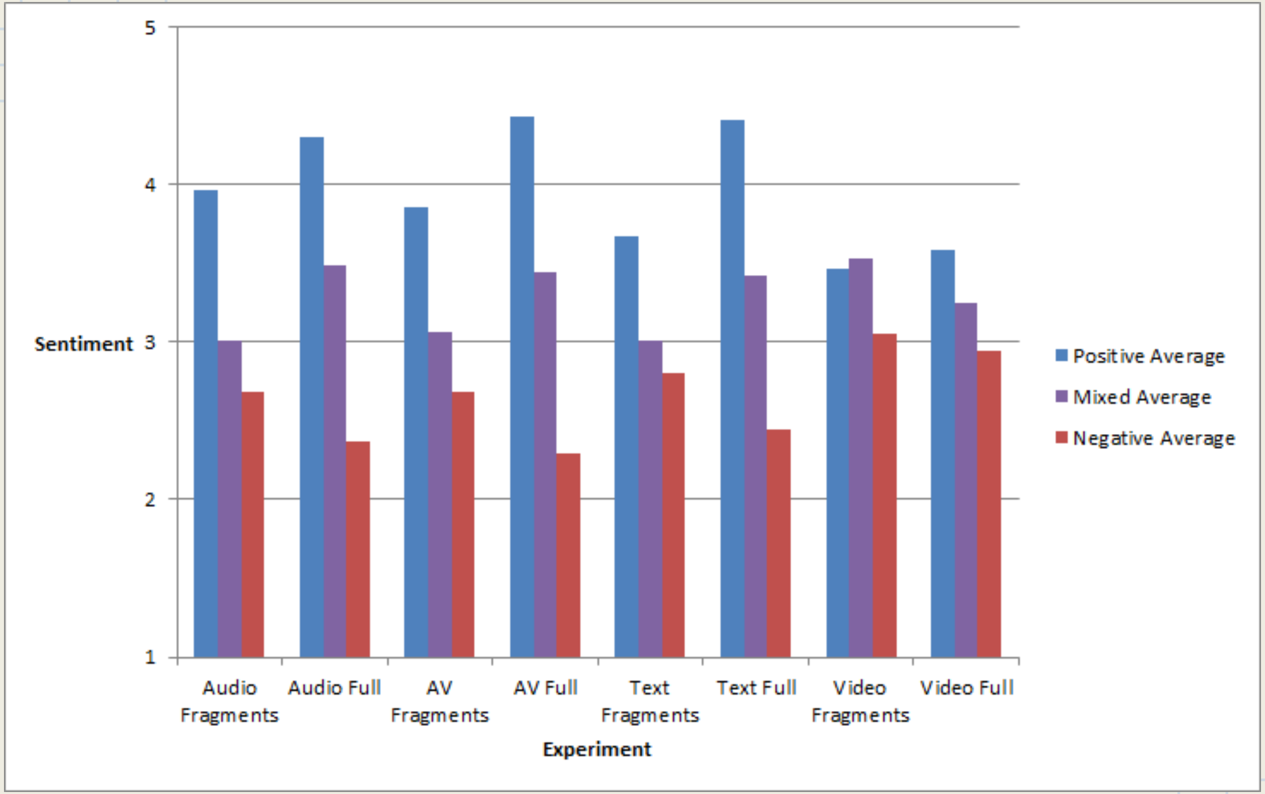
experiment	Audio Fragments	Audio Full	AV Fragments	AV Full	Text Fragments	Text Full	Video Fragments	Video Full
Average	3.4909774	3.72504177	3.43903091	3.77362155	3.25512650	3.6625	3.38677248	3.38013888
Positive Average	3.9000410	4.30436507	3.80468664	4.43387445	3.49980042	4.14583333	3.47170138	3.56689814
Mixed Average	3.0798412	3.48819444	3.12047619	3.43928571	3.00614973	3.425	3.53529411	3.25
Negative Average	2.6928571	2.36875	2.69222582	2.29226190	2.79941077	2.45	3.05509259	2.95

Overall Average Sentiment



- Full higher average
- Due to fragments having more 3’s (summary/neutral snippets)

Avg. Sentiment By Polarity

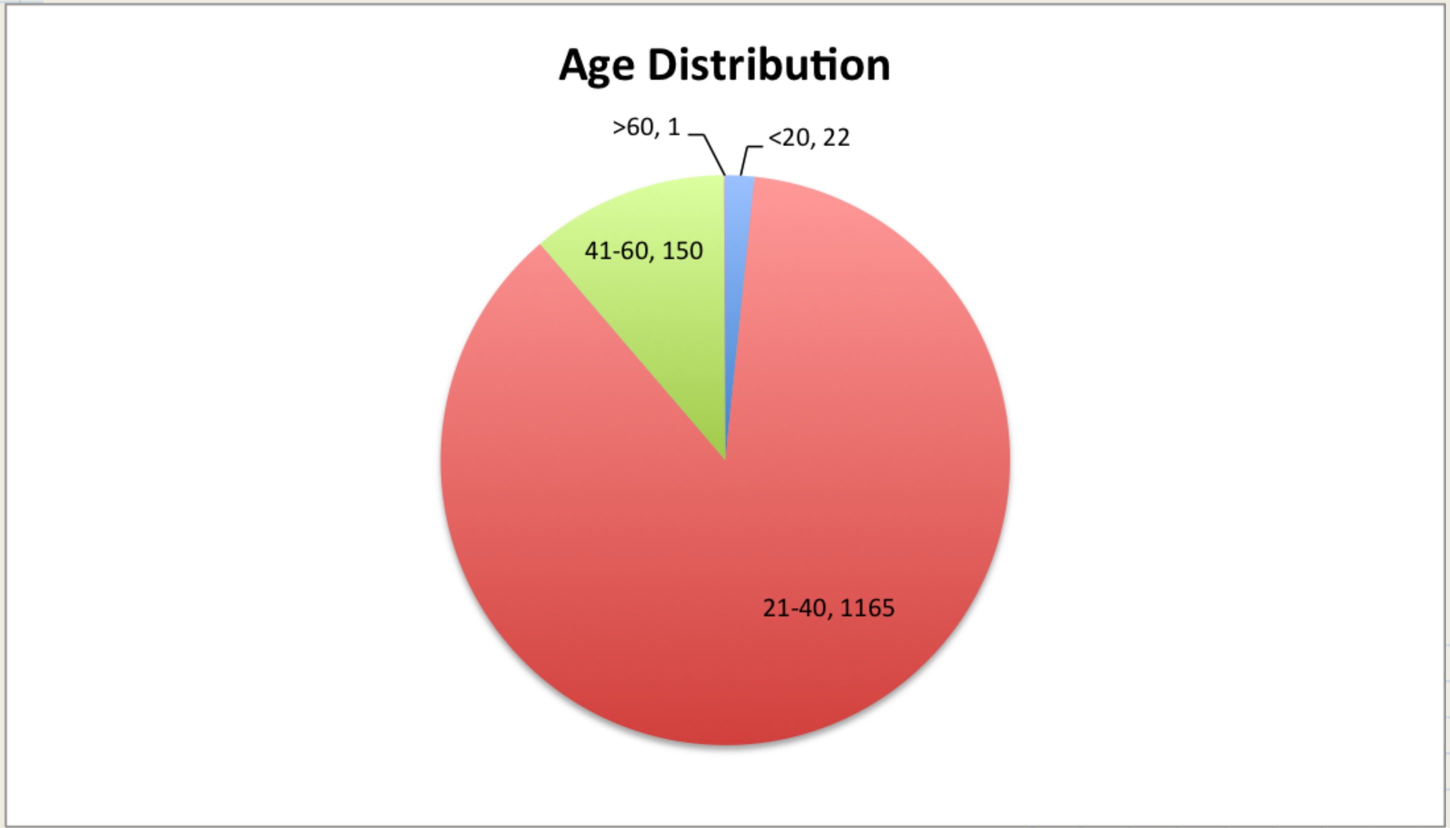


- Full more polarized
- AV most polarized
- Video only fragments can (kind of) distinguish negative, but not mixed vs positive
- Polarity possibly indicative of accuracy

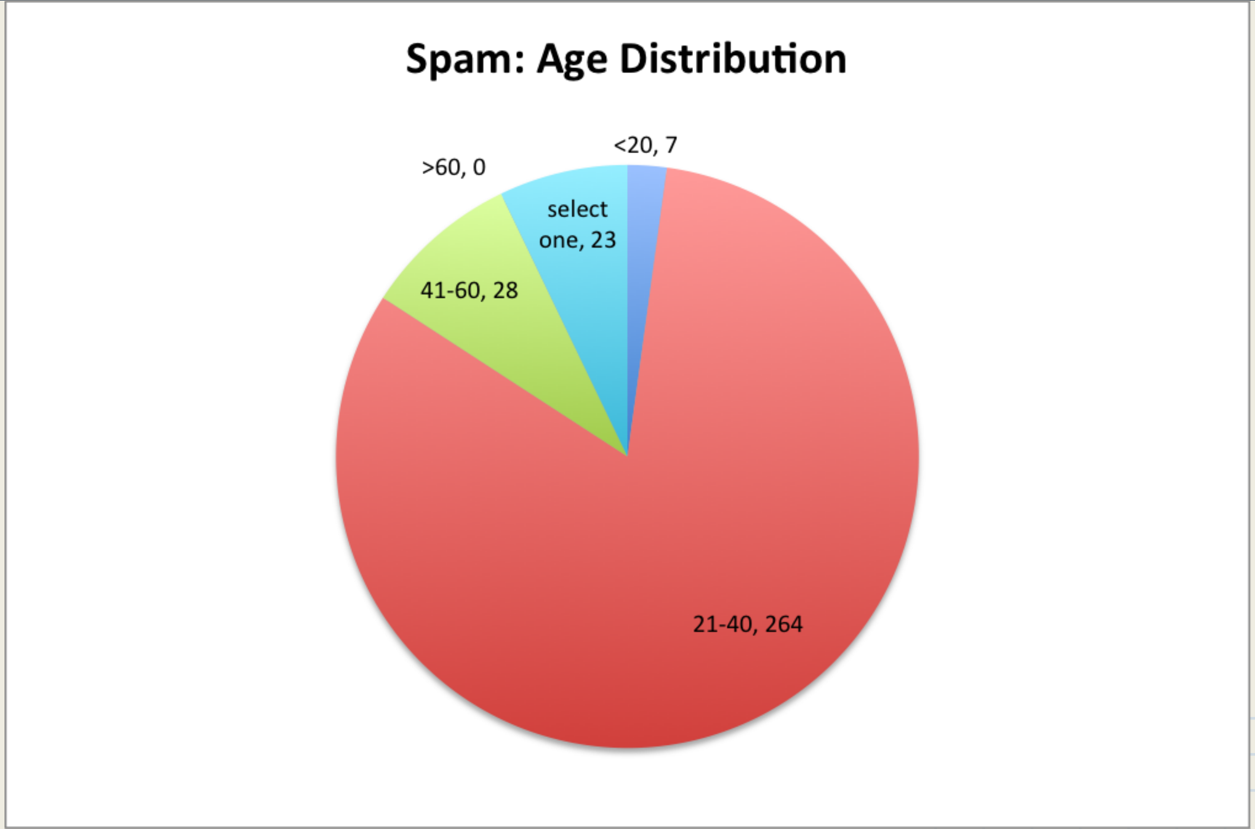
Analysis to Come

- Scatterplots comparing average by video/by fragment
- Box and whisker plot
- Annotate all data with “gold”, compare accuracy
- Any ideas?

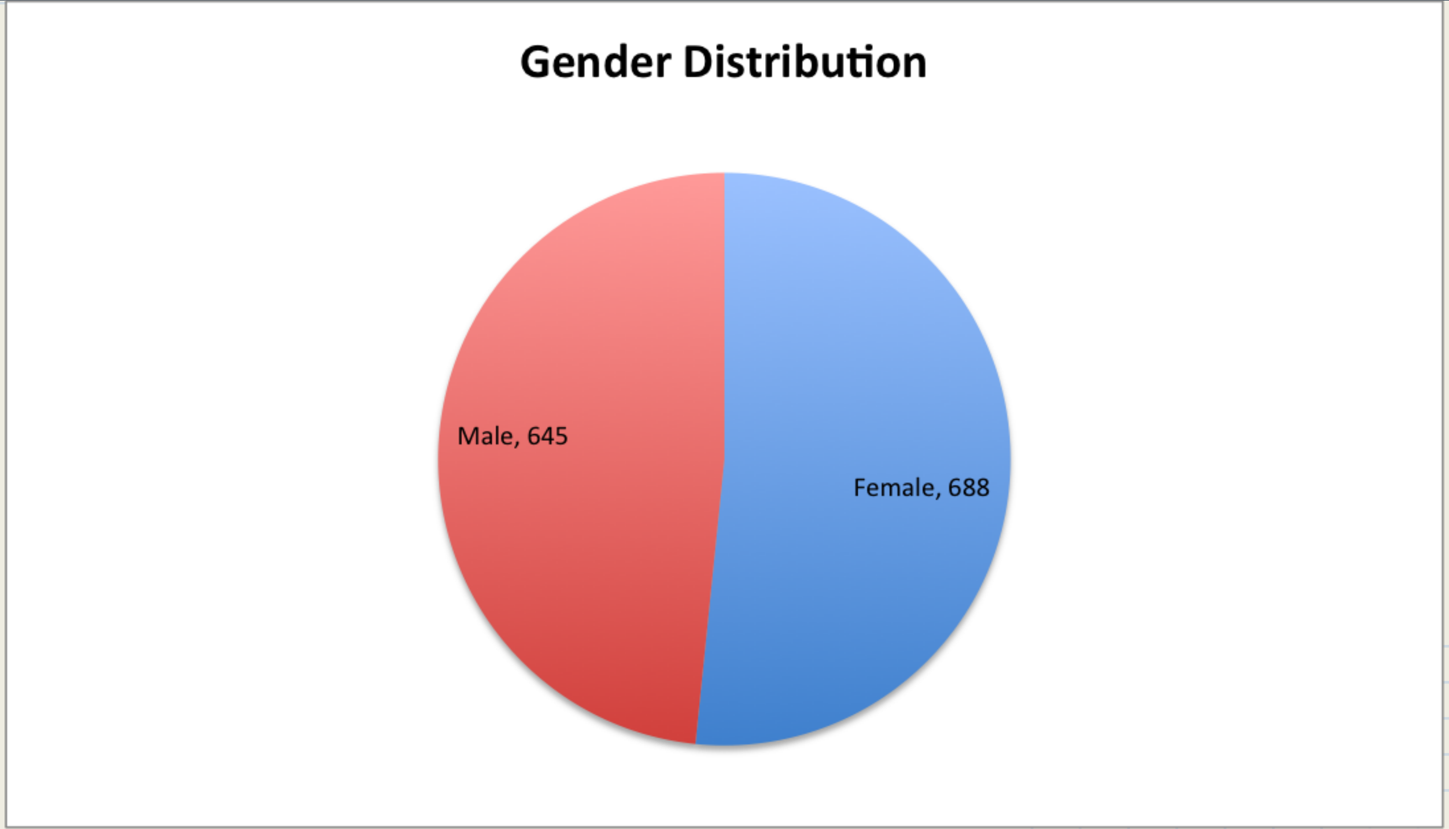
Demographics - Filtered Data



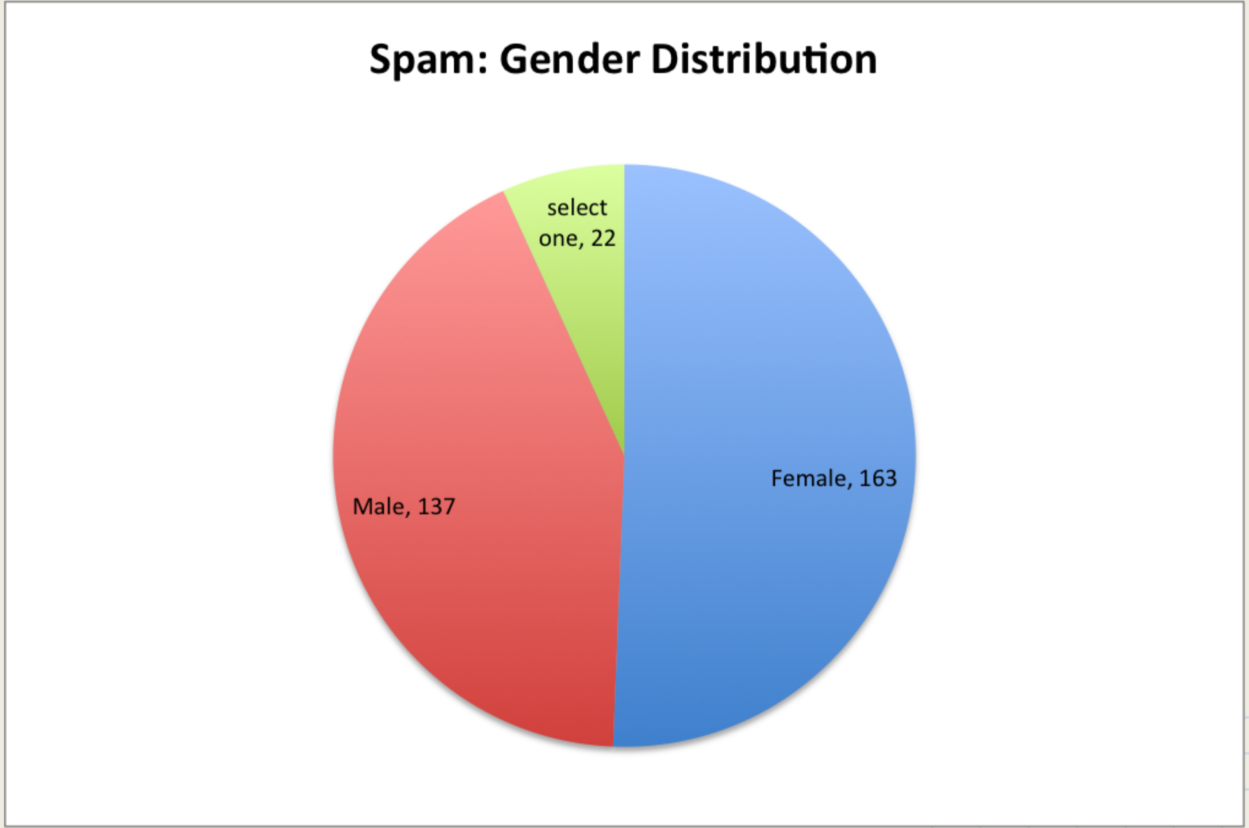
Demographics - Spam



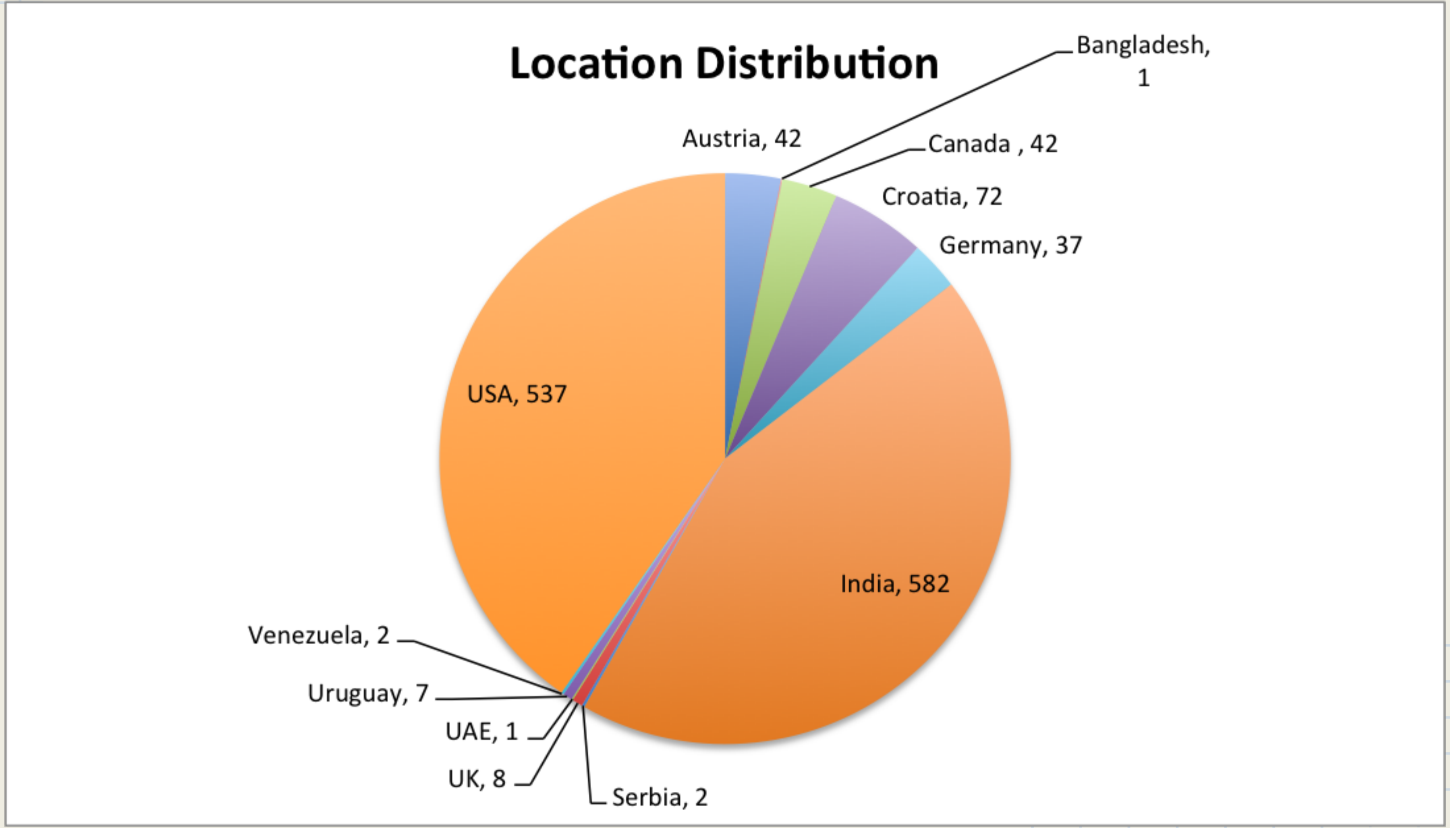
Demographics - Filtered Data



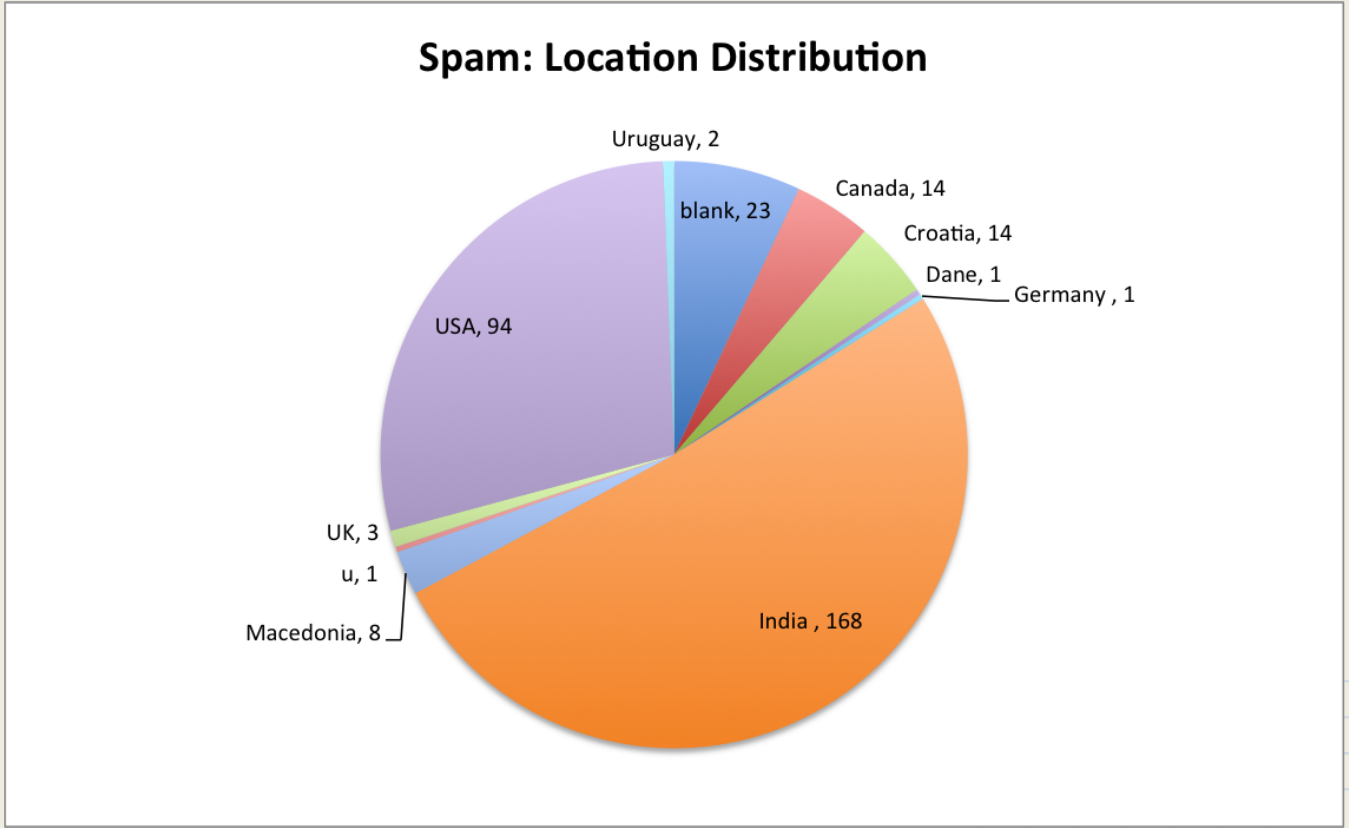
Demographics - Spam



Demographics - Filtered Data



Demographics - Spam



Conclusion

- Audio and Audio/Video had consistently higher agreement than text only
  - Using transcriptions of videos for sentiment analysis less reliable
- Audio/Video less agreement than Audio only
  - Video introducing noise?
  - For human sentiment judgment video not necessary?
  - More likely due to worker restrictions/pay
- Higher Interfragment agreement for video only
  - Pattern probably doesn’t stand if normalized for chance agreement
  - Still hopeful that sentiment of whole video can be deduced from video only fragments with **neutral content** (if not by untrained human annotators then by trained human annotators or video feature based classifiers)

Future Work

- Attempt to classify full videos using visual features based only on neutral fragments
- Add better spam filtering for video only
- Increase sample size
- Increase number of annotations per fragment/video
- Try different domains (debates vs reviews)
- Measure sentiment for muted video alongside transcription
- Compare sentiment evaluations for manual transcriptions against automatic transcriptions
- Explore more fine grained sentiment categorizations (continuous scales; mixed option in addition to neutral)
- Assess emotional states (angry, happy, sad, etc.)
- Assess judgments of individual words

Reference

- Morency, Louis-Phillipe and Mihalcea, Rada and Doshi, Payal. Towards Multimodal Sentiment Analysis: Harvesting Opinions from the Web, Proceedings of ICMI '11 Proceedings of the 13th international conference on multimodal interfaces, p. 169-176.
- Verónica Pérez-Rosas, and Rada Mihalcea. (2013). Sentiment analysis of online spoken reviews. INTERSPEECH, page 862-866.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. Psychological Bulletin, Vol. 76, No. 5 pp. 378–382