Evaluating Computer-Generated Transcripts

Lubov McKone
Summer 2022

1 Background & research questions

2 Data curation

3 Analysis & visualization

4 Conclusions & what’s next

Background

- The Industry Documents Library is a trove of data (metadata, textual data)

- Researchers should receive guidance on potential gaps or mistranslations in the data

- Text analysis is only as reliable as the quality of the transcription

An archive of 14 million documents created by tobacco companies about their advertising, manufacturing, marketing, scientific research and political activities, hosted by the UCSF Library.

Background

- Tobacco video collection:
    - 5,249 videos
    - Interviews, commercials, court proceedings, press conferences, news

- We narrowed our scope to **commercials** and **court proceedings**

Research questions

1. Taking into account factors such as year and runtime, how does computer transcription accuracy differ between television commercials and court proceedings?

2. How might transcription accuracy impact the conclusions drawn from the data?

3. What guidance can we give to researchers to prevent false conclusions?

Video selection

- Still > 1,000 commercial & court videos
- Fellows each selected 10 videos each per category
- Range of year, quality, and runtime
- 40 videos total

Data curation

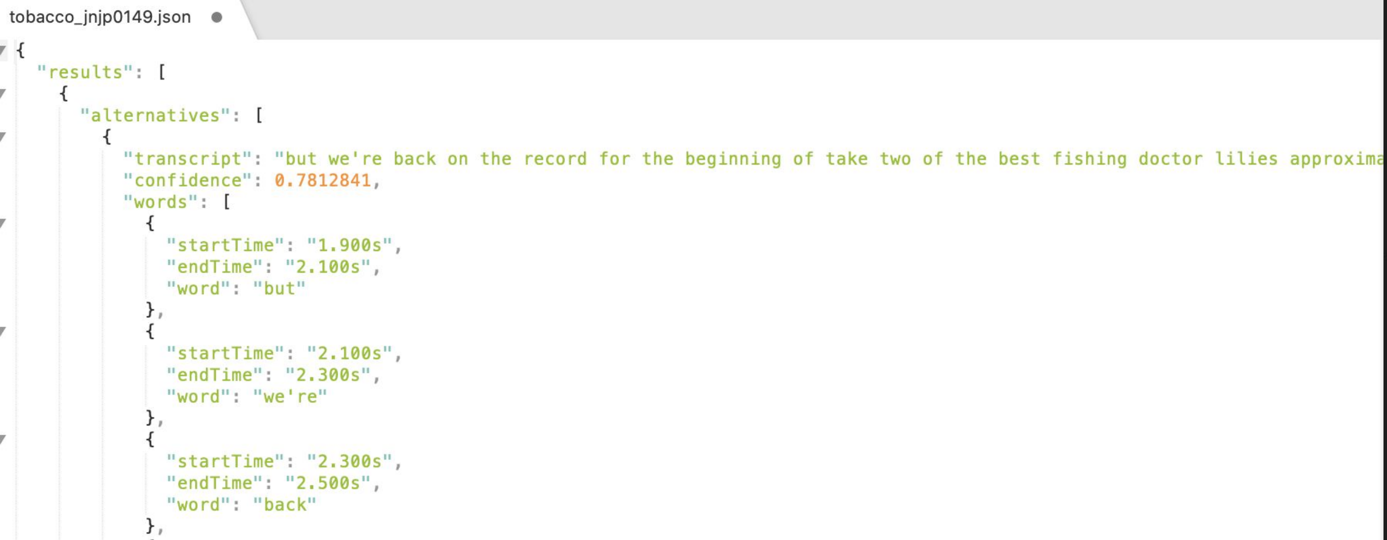
- Gather:
  - Video metadata
  - Test transcripts & reference transcripts
  - Measures of comparison

Video metadata

- Feed video urls to python `internetarchive` module to retrieve:
  - Runtime
  - Year

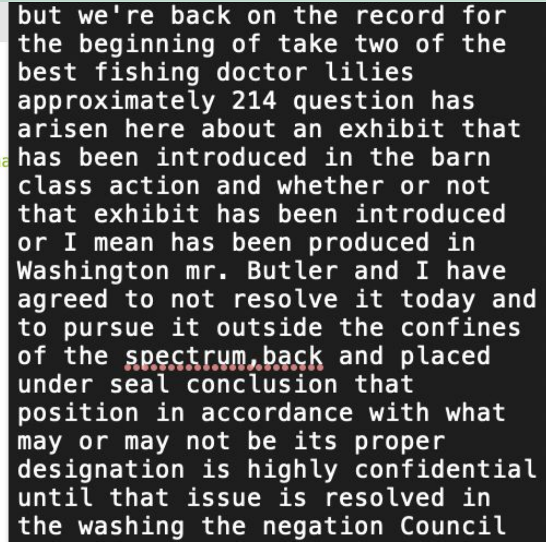
Test transcripts

- Generated transcripts from Internet Archive videos using Google AutoML
- Converted JSON into more readable .txt files



tobacco_jnjp0149.json

```
{
  "results": [
    {
      "alternatives": [
        {
          "transcript": "but we're back on the record for the beginning of take two of the best fishing doctor lilies approxima
          "confidence": 0.7812841,
          "words": [
            {
              "startTime": "1.900s",
              "endTime": "2.100s",
              "word": "but"
            },
            {
              "startTime": "2.100s",
              "endTime": "2.300s",
              "word": "we're"
            },
            {
              "startTime": "2.300s",
              "endTime": "2.500s",
              "word": "back"
            },
```



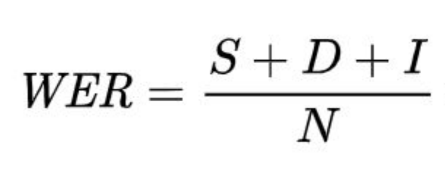
```
but we're back on the record for
the beginning of take two of the
best fishing doctor lilies
approximately 214 question has
arisen here about an exhibit that
has been introduced in the barn
class action and whether or not
that exhibit has been introduced
or I mean has been produced in
Washington mr. Butler and I have
agreed to not resolve it today and
to pursue it outside the confines
of the spectrum,back and placed
under seal conclusion that
position in accordance with what
may or may not be its proper
designation is highly confidential
until that issue is resolved in
the washing the negation Council
```

Reference transcripts

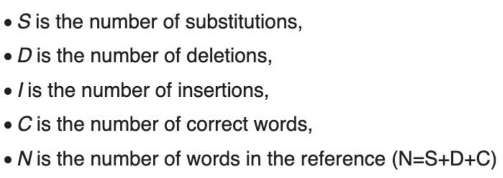
- Our Junior Fellows edited 20 transcripts each for us to use as a “correct” version of each computer-generated transcript

Measures of comparison

- Word error rate: Edit distance for corpi



$$WER = \frac{S + D + I}{N}$$



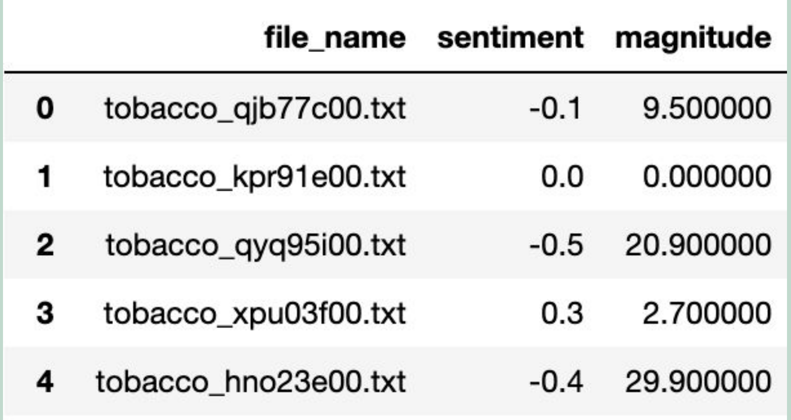
- S is the number of substitutions,
- D is the number of deletions,
- I is the number of insertions,
- C is the number of correct words,
- N is the number of words in the reference (N=S+D+C)

* We subtracted from 1 to get word accuracy rate

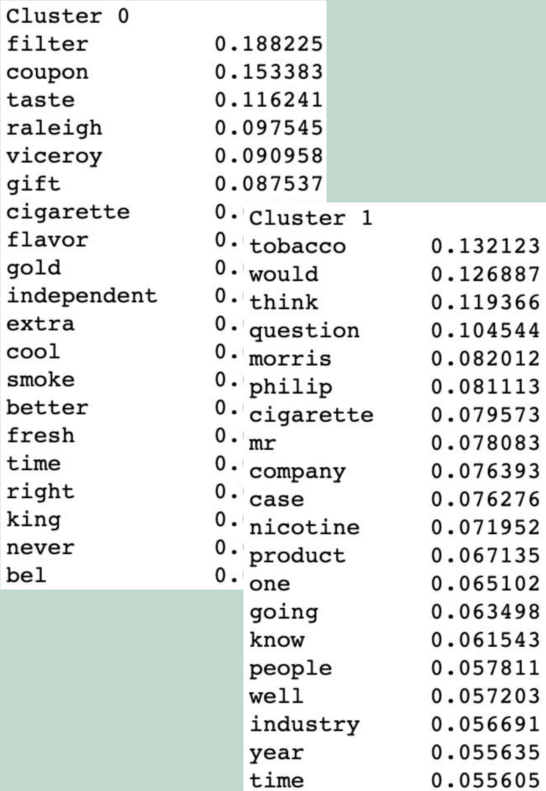
- BLEU score: Algorithm measuring n-gram matches between corpi, normalized for n-gram frequency
- Human-evaluated accuracy
- Google AutoML confidence score

Measures of comparison

- Sentiment
- Topic modeling



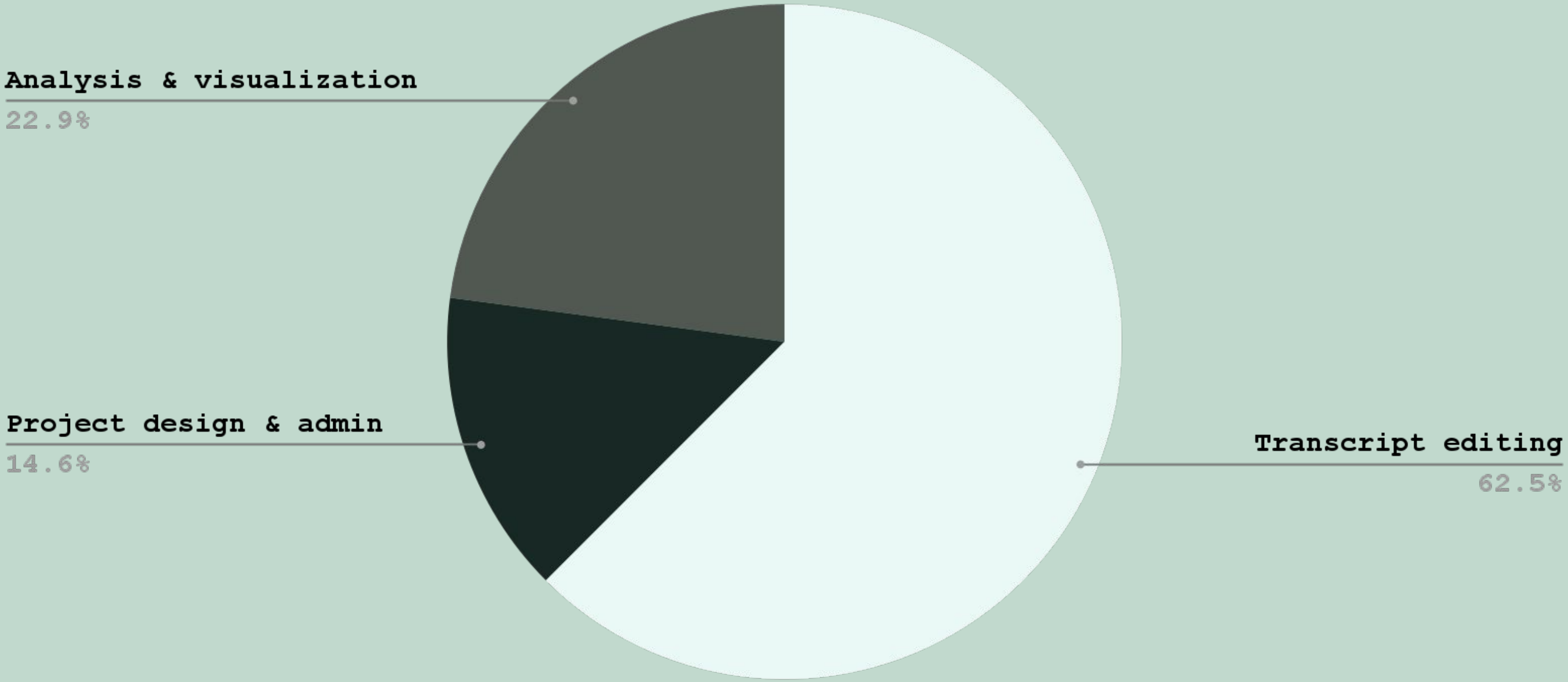
	file_name	sentiment	magnitude
0	tobacco_qjb77c00.txt	-0.1	9.500000
1	tobacco_kpr91e00.txt	0.0	0.000000
2	tobacco_qyq95i00.txt	-0.5	20.900000
3	tobacco_xpu03f00.txt	0.3	2.700000
4	tobacco_hno23e00.txt	-0.4	29.900000



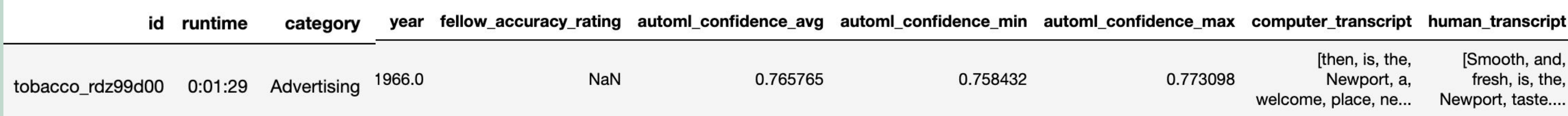
```
Cluster 0
filter          0.188225
coupon          0.153383
taste           0.116241
raleigh         0.097545
viceroy         0.090958
gift            0.087537
cigarette       0.
flavor          0.
gold            0.
independent     0.
extra           0.
cool            0.
smoke           0.
better          0.
fresh           0.
time            0.
right           0.
king            0.
never           0.
bel             0.
```

```
Cluster 1
tobacco        0.132123
would          0.126887
think          0.119366
question       0.104544
morris         0.082012
philip         0.081113
cigarette      0.079573
mr             0.078083
company        0.076393
case           0.076276
nicotine       0.071952
product        0.067135
one            0.065102
going          0.063498
know           0.061543
people         0.057811
well           0.057203
industry       0.056691
year           0.055635
time           0.055605
```

Project task breakdown



Final dataset



id	runtime	category	year	fellow_accuracy_rating	automl_confidence_avg	automl_confidence_min	automl_confidence_max	computer_transcript	human_transcript
tobacco_rdz99d00	0:01:29	Advertising	1966.0	NaN	0.765765	0.758432	0.773098	[then, is, the, Newport, a, welcome, place, ne...	[Smooth, and, fresh, is, the, Newport, taste....

- Runtime
- Category
- Year
- Fellow accuracy rating
- AutoML confidence score

- Computer & human transcripts
- Word error rate
- BLEU score
- Sentiment & magnitude for test & reference
- Topic cluster for test & reference

Research questions

1. Taking into account factors such as year and runtime, how does computer transcription accuracy differ between television commercials and court proceedings?

2. How might transcription accuracy impact the conclusions drawn from the data?

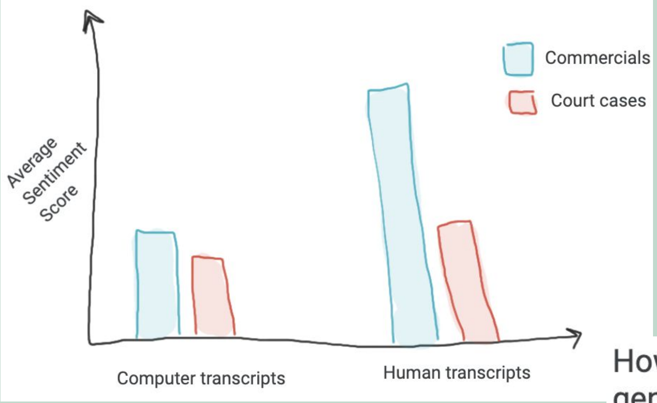
3. What guidance can we give to researchers to prevent false conclusions?

Analysis planning session

- We brainstormed questions and visualizations that could help answer our research questions

Does the Word Error Rate impact the sentiment score of the transcriptions?

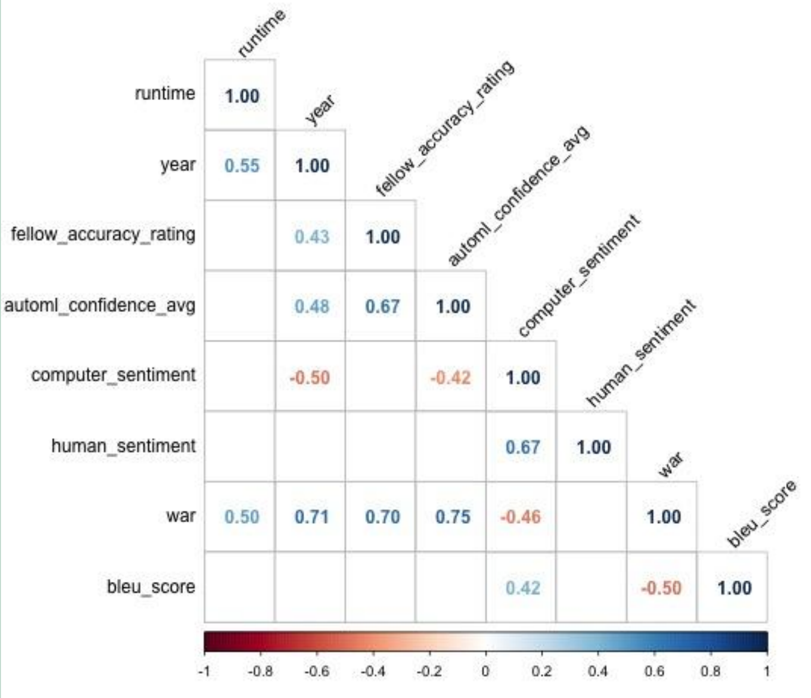
What technological improvement has universaly bettered the quality of audio, resulting in better transcriptions after its invention? What year?



2. Is there a higher correlation between google autoML confidence score and the type of videos or between the auto ML confidence score and the length of the videos (ie.e amount of text extracted)

How does sentiment score between different categories differ between computer-generated and human-generated transcripts?

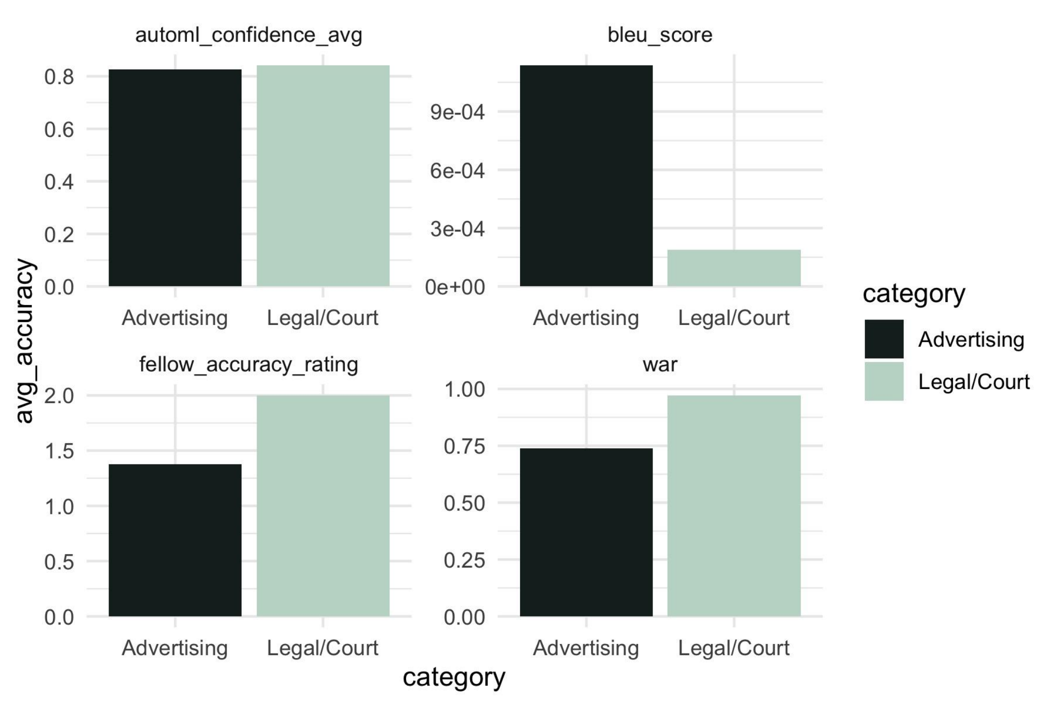
Exploring the Data



	runtime	year	fellow_accuracy_rating	automl_confidence_avg	computer_sentiment	human_sentiment	war	bleu_score
runtime	1.00							
year	0.55	1.00						
fellow_accuracy_rating		0.43	1.00					
automl_confidence_avg		0.48	0.67	1.00				
computer_sentiment		-0.50		-0.42	1.00			
human_sentiment					0.67	1.00		
war	0.50	0.71	0.70	0.75	-0.46		1.00	
bleu_score					0.42		-0.50	1.00

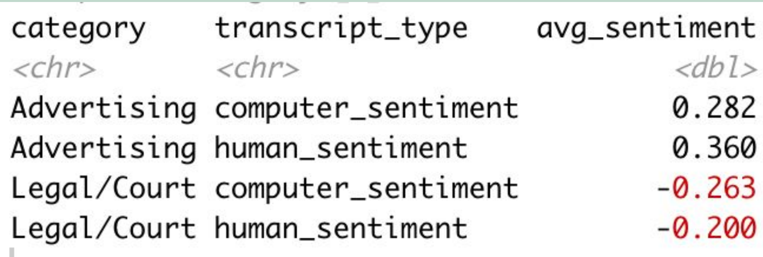
- The more recent the video, the better the transcription (fellow rating, Google AutoML confidence, WER)
- AutoML confidence, fellow accuracy rating, and Word Error Rate are all significantly positively correlated

Exploring the Data

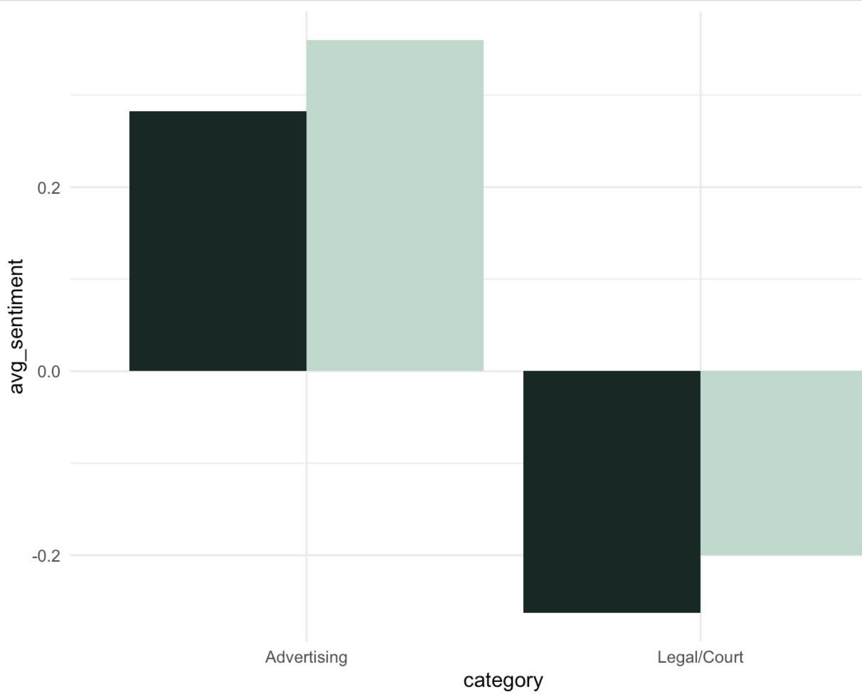


- Overall, transcript accuracy seems higher in the legal/court category than in advertising
- BLEU score may not be the most compatible metric for Google AutoML transcripts

Exploring the Data



```
category    transcript_type   avg_sentiment
<chr>       <chr>                     <dbl>
Advertising computer_sentiment        0.282
Advertising human_sentiment           0.360
Legal/Court computer_sentiment       -0.263
Legal/Court human_sentiment          -0.200
```



Next steps

- Investigate sentiment and other factors impacted by accuracy
- Investigate statistical significance and multi-variable relationships of preliminary findings
- Document this project as a reproducible case study
- Compile instructional materials

Oddities

- Excel character limit of 32,767 in one cell
- Google AutoML <-> Internet Archive videos - file not always named the same, can’t retrieve metadata
- Had a video partially in Spanish - Google AutoML can’t handle multiple languages in a video
- And more!

Thank You!