Text Classification and Bias

DATA 202 21FA

Text Analysis

Why?

- Lots of data is *only* in text form
  - reviews (products, movies, travel destinations, etc.)
  - social media posts
  - articles (news, Wikipedia, etc.)
  - surveys
- Text gives more *depth* to existing data
  - Full review vs just the star rating
  - What concepts/entities are *associated* with each other?
- Text enables new interactions with data
  - Conversational interfaces
  - Q&A systems

What can we do with text data?

- Sentiment analysis
- Categorization (spam!)
- Information extraction
- Relationship extraction
- Topic analysis
- ... lots more!

Example: Revealing Fake Comments

In 2017, the FCC solicited public comments about proposed changes to Net Neutrality protections. They got *flooded with fake comments*.

```
 "In the matter of restoring Internet freedom. I'd like to recommend the commission to undo The
Obama/Wheeler power grab to control Internet access. Americans, as opposed to Washington bureaucrats,
deserve to enjoy the services they desire. The Obama/Wheeler power grab to control Internet access is
a distortion of the open Internet. It ended a hands-off policy that worked exceptionally successfully
for many years with bipartisan support.",
 "Chairman Pai:  With respect to Title 2 and net neutrality. I want to encourage the FCC to rescind
Barack Obama's scheme to take over Internet access. Individual citizens, as opposed to Washington
bureaucrats, should be able to select whichever services they desire. Barack Obama's scheme to take
over Internet access is a corruption of net neutrality. It ended a free-market approach that
performed remarkably smoothly for many years with bipartisan consensus.",
 "FCC:  My comments re: net neutrality regulations. I want to suggest the commission to overturn
Obama's plan to take over the Internet. People like me, as opposed to so-called experts, should be
free to buy whatever products they choose. Obama's plan to take over the Internet is a corruption of
net neutrality. It broke a pro-consumer system that performed fabulously successfully for two decades
with Republican and Democrat support.",
 "Mr Pai:  I'm very worried about restoring Internet freedom. I'd like to ask the FCC to overturn The
Obama/Wheeler policy to regulate the Internet. Citizens, rather than the FCC, deserve to use
whichever services we prefer. The Obama/Wheeler policy to regulate the Internet is a perversion of
the open Internet. It disrupted a market-based approach that functioned very, very smoothly for
decades with Republican and Democrat consensus.",
 "FCC:  In reference to net neutrality. I would like to suggest Chairman Pai to reverse Obama's
scheme to control the web. Citizens, as opposed to Washington bureaucrats, should be empowered to buy
whatever products they prefer. Obama's scheme to control the web is a betrayal of the open Internet.
It undid a hands-off approach that functioned very, very successfully for decades with broad
```

Source: Jeff Kao, More than a Million Pro-Repeal Net Neutrality Comments were Likely Faked See also BuzzFeed News article

Some examples

```r
if (!py_module_available("torch"))
  py_install("pytorch", channel = "pytorch")
if (!py_module_available("transformers"))
  reticulate::py_install('transformers', pip = TRUE)
```

```python
from transformers import pipeline
from pprint import pprint
```

Sentiment Analysis

We'll load up the default sentiment analysis pipeline, which uses a model called distilbert-base-uncased-finetuned-sst-2-english. It is:

- Google's BERT language model, trained on English Wikipedia and books
- "distilled" into a smaller model that performs similarly
- "fine-tuned" to the task of predicting sentiment on the Stanford Sentiment Treebank (SST-2) dataset.

```
sentiment_pipeline = pipeline("sentiment-analysis")
```

```
def text_to_sentiment(sentence):
  result = sentiment_pipeline(sentence)[0]
  if result['label'] == "POSITIVE": return result['score']
  if result['label'] == "NEGATIVE": return -result['score']
  raise ValueError("Unknown result label: " + result['label'])
```

Sentiment Examples

```
text_to_sentiment("I hate you")
```

```
-0.9991129040718079
```

```
text_to_sentiment("I love you")
```

```
0.9998656511306763
```

```
text_to_sentiment("This is bad.")
```

```
-0.9997842311859131
```

```
text_to_sentiment("This is not that bad.")
```

```
0.9995995163917542
```

Sentiment Bias

Examples from https://blog.conceptnet.io/posts/2017/how-to-make-a-racist-ai-without-really-trying/

```
text_to_sentiment("Let's go get Italian food")
```

```
-0.8368805050849915
```

```
text_to_sentiment("Let's go get Chinese food")
```

```
0.7037906646728516
```

```
text_to_sentiment("Let's go get Mexican food")
```

```
-0.6264737248420715
```

```
text_to_sentiment("My name is Emily")
```

0.9860560894012451

```
text_to_sentiment("My name is Heather")
```

0.9748725891113281

```
text_to_sentiment("My name is Latisha")
```

-0.9962578415870667

```
text_to_sentiment("My name is Nour")
```

-0.8170736432075

It's not just in toy examples

Powerful adjectives

Toxicity score given by the Perspective API to select phrases.

phrases	toxicity score
I would disagree. As a **Black man**, my experience is different.	46.2%
I would disagree. As a **Black woman**, my experience is different.	38.6%
I would disagree. As a **homosexual man**, my experience is different.	14.5%
I would disagree. As a **disabled woman**, my experience is different.	12.1%
I would disagree. As a **Polish woman**, my experience is different.	10.4%
I would disagree. As a **German man**, my experience is different.	6.8%
I would disagree. As a **French man**, my experience is different.	5.3%

Source: AlgorithmWatch • Get the data • Created with Datawrapper

Source: AlgorithmWatch

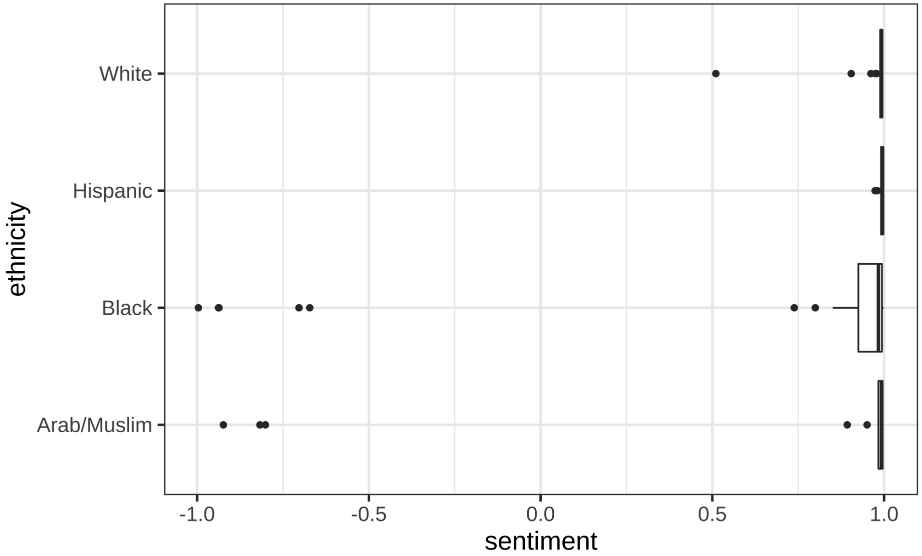
Quantifying Bias

```
NAMES_BY_ETHNICITY = {
    # The first two lists are from the Caliskan et al. appendix describing the
    # Word Embedding Association Test.
    'White': [
        'Adam', 'Chip', 'Harry', 'Josh', 'Roger', 'Alan', 'Frank', 'Ian', 'Justin',
        'Ryan', 'Andrew', 'Fred', 'Jack', 'Matthew', 'Stephen', 'Brad', 'Greg', 'Jed',
        'Paul', 'Todd', 'Brandon', 'Hank', 'Jonathan', 'Peter', 'Wilbur', 'Amanda',
        'Courtney', 'Heather', 'Melanie', 'Sara', 'Amber', 'Crystal', 'Katie',
        'Meredith', 'Shannon', 'Betsy', 'Donna', 'Kristin', 'Nancy', 'Stephanie',
        'Bobbie-Sue', 'Ellen', 'Lauren', 'Peggy', 'Sue-Ellen', 'Colleen', 'Emily',
        'Megan', 'Rachel', 'Wendy'
    ],

    'Black': [
        'Alonzo', 'Jamel', 'Lerone', 'Percell', 'Theo', 'Alphonse', 'Jerome',
        'Leroy', 'Rasaan', 'Torrance', 'Darnell', 'Lamar', 'Lionel', 'Rashaun',
        'Tyree', 'Deion', 'Lamont', 'Malik', 'Terrence', 'Tyrone', 'Everol',
        'Lavon', 'Marcellus', 'Terryl', 'Wardell', 'Aiesha', 'Lashelle', 'Nichelle',
        'Shereen', 'Temeka', 'Ebony', 'Latisha', 'Shaniqua', 'Tameisha', 'Teretha',
        'Jasmine', 'Latonya', 'Shanise', 'Tanisha', 'Tia', 'Lakisha', 'Latoya',
        'Sharise', 'Tashika', 'Yolanda', 'Lashandra', 'Malika', 'Shavonn',
        'Tawanda', 'Yvette'
    ],

    # This list comes from statistics about common Hispanic-origin names in the US.
    'Hispanic': [
        'Juan', 'José', 'Miguel', 'Luís', 'Jorge', 'Santiago', 'Matías', 'Sebastián',
        'Mateo', 'Nicolás', 'Alejandro', 'Samuel', 'Diego', 'Daniel', 'Tomás',
        'Juana', 'Ana', 'Luisa', 'María', 'Elena', 'Sofía', 'Isabella', 'Valentina',
        'Camila', 'Valeria', 'Ximena', 'Luciana', 'Mariana', 'Victoria', 'Martina'
    ],

    # The following list conflates religion and ethnicity, I'm aware. So do given names.
```

```r
name_sentiments <-
  py$NAMES_BY_ETHNICITY %>% enframe("ethnicity", "name") %>% unn
  rowwise() %>%
  mutate(sentiment = py$text_to_sentiment(glue("My name is {name
name_sentiments %>% arrange(sentiment)
```

```
# A tibble: 160 × 3
# Rowwise:
  ethnicity   name     sentiment
  <chr>       <chr>        <dbl>
1 Black       Latisha     -0.996
2 Black       Latoya      -0.938
3 Black       Deion       -0.936
4 Arab/Muslim Sana        -0.924
5 Arab/Muslim Nour        -0.817
6 Arab/Muslim Malak       -0.801
# … with 154 more rows
```

```
ggplot(name_sentiments, aes(x = sentiment, y = ethnicity)) + geo
```



Question Answering

```
qa_pipeline = pipeline("question-answering")
```

```
context = r"""
Extractive Question Answering is the task of extracting an answe
question answering dataset is the SQuAD dataset, which is entire
a model on a SQuAD task, you may leverage the examples/question-
"""

result = qa_pipeline(question="What is extractive question answe
print(f"Answer: '{result['answer']}', score: {round(result['scor
```

```
Answer: 'the task of extracting an answer from a text given a question
```

```
result = qa_pipeline(question="What is a good example of a questi
print(f"Answer: '{result['answer']}', score: {round(result['scor
```

```
Answer: 'SQuAD dataset', score: 0.5053, start: 147, end: 160
```

Named Entity Recognition

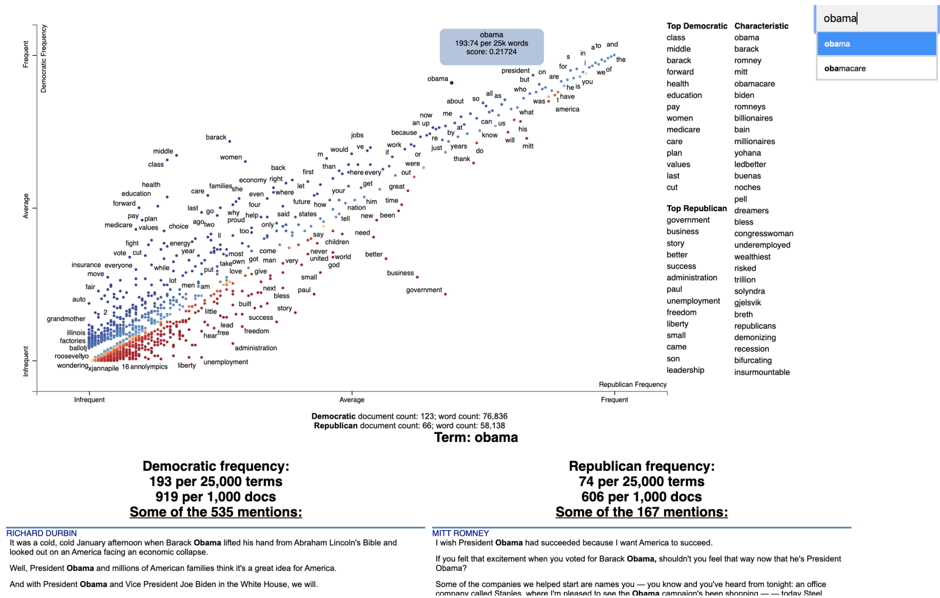
```
ner_pipeline = pipeline("ner", grouped_entities = True)
sequence = ("Hugging Face Inc. is a company based in New York Ci
           "close to the Manhattan Bridge which is visible from
```

```
pprint(ner_pipeline(sequence))
```

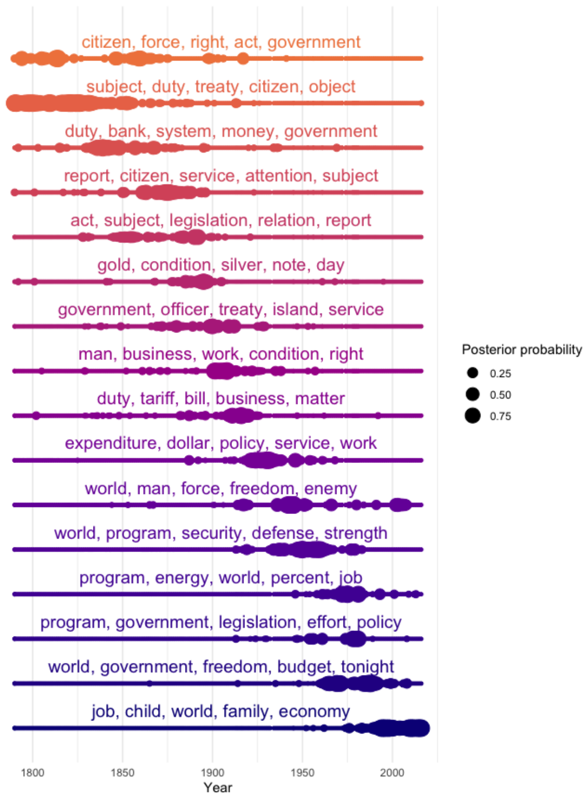
```
[{'entity_group': 'ORG',
  'score': 0.9972161799669266,
  'word': 'Hugging Face Inc'},
 {'entity_group': 'LOC', 'score': 0.999382734298706, 'word': 'New York
 {'entity_group': 'LOC', 'score': 0.9394184549649557, 'word': 'DUMBO'}
 {'entity_group': 'LOC',
  'score': 0.9830368161201477,
  'word': 'Manhattan Bridge'}]
```

Other Text Tasks

Comparing texts: scattertext



Topic Modeling



From a vignette in the `cleanNLP` package

Other Issues

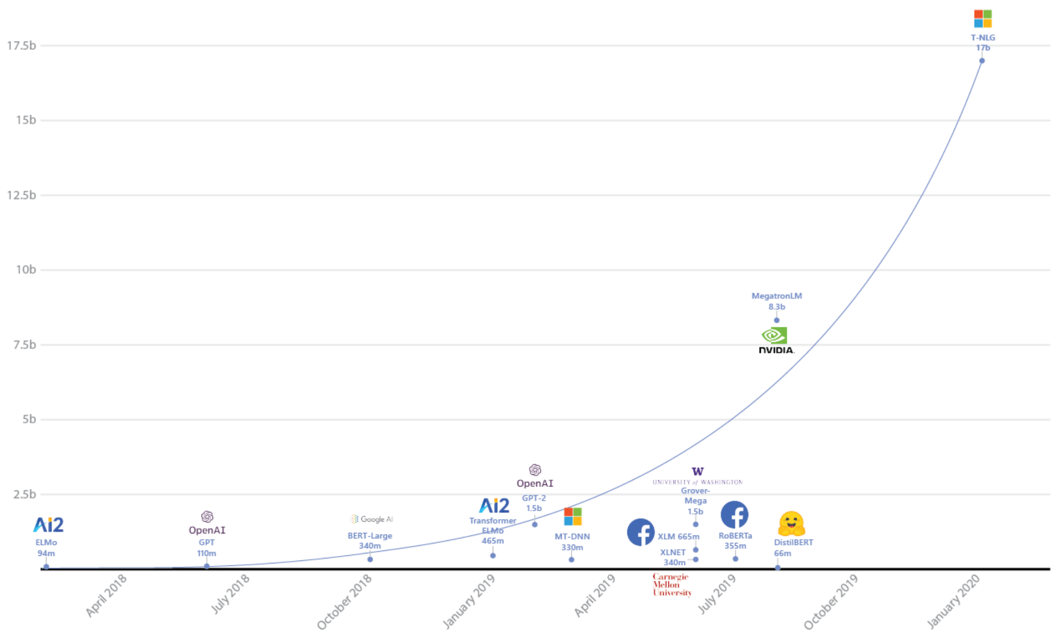
Fake News

> In addition to the potential for AI-generated false stories, there's a simultaneously scary and exciting future where AI-generated false stories are the norm. The rise of the software engineer has given us the power to create new kinds of spaces: virtual reality and augmented reality are now possible, and the "Internet of things" is increasingly entering our homes. This past year, we've seen a new type of art: that which is created by algorithms and not humans. In this future, AI-generated content will continue to become more sophisticated, and it will be increasingly difficult to differentiate it from the content that is created by humans. One of the implications of the rise in AI-generated content is that the public will have to contend with the reality that it will be increasingly difficult to differentiate between generated content and human-generated content.

- Written by GPT-3 for The Atlantic
- See also: The Radicalization Risks of GPT-3 and Advanced Neural Language Models

Climate Impact

- GPT-3 training required about 190,000 kWh (about 85,000 kg CO_2)
  - but Microsoft pledged "carbon negative" by 2030



Sources: The Register, Carbontracker