CSC401/2511
ONE OF THE FIRST TUTORIALS EVER, EVERYBODY SAYS SO

Frank Rudzicz
THE WORLD WE LIVE IN

Jack Posobiec
@JackPosobiec

False flag. Planted Comet Pizza Gunman will be used to push for censorship of independent news sources that are not corporate owned

Michael G Flynn
@mflynnJR

Until #Pizzagate proven to be false, it'll remain a story. The left seems to forget #PodestaEmails and the many "coincidences" tied to it.
The Donald before learning the Vegas shooter was an old white man.
After

Something is not right here. This narrative doesn't make any sense, there is something big missing that we'll probably find out in the next few days.

Don't you find it odd that the shooter had no motivation? Have you seen the born identity super soliders concept? They basically program shooters to act when given a command.

This feels fishy. A 60 year old retiree manages to get a fully automatic weapon and several hundred rounds of ammo & decides to go on a rampage and a country music concert? What are we missing here?

Antifa?
This smells like a false flag, research this.

Yeah a white supremacist is gonna shoot up a country music concert. Uh huh. Either false flag or demo-commie trash.

Oh fuck, didn't Alex Jones very recently warn of an impending false flag??

So this shooter may have been ISIS. Why did the FBI already dismiss links to terrorism...something smells

Why is EVERY news outlet pushing the ISIS claim?? This is really smelling like a false flag op to me now! Soros is my main suspect!!

Bitch ass shillary using this to attack suppressors and the ebul nra. Fuck that stupid cunt. This has false flag written all over it

Something about the known facts is very fishy right now.

We don't even know for sure at the moment Paddock was the killer. They found him dead of an apparent suicide in the hotel room, but he might have been killed by whoever actually shot the guns and those people left before the police arrived. I'm not calling conspiracy yet but its probably way too early and the pieces we have so far dont really fit yet.
REPORT: Was Vegas Shooter Part of Antifa? Here's What The Media Isn't Saying (VIDEO) · Silence is Consent

SILENCEISCONSENT.NET
Five Things That Just Don’t Add Up About The Las Vegas Mass Shooting

Our hearts and prayers go out to all those killed or injured in the Las Vegas shooting, and in a nation where so many anti-Americans are kneeling in...
The mainstream media narrative about the Las Vegas shooting has been debunked by two videos proving there were multiple gunmen involved in an orchestrated attack.

Las Vegas: Video Confirms Multiple Shooters, Coordinated Attack

The mainstream media narrative about the shooting has been debunked by two videos proving there were at least two gunmen.
<table>
<thead>
<tr>
<th>Logical Fallacies</th>
</tr>
</thead>
<tbody>
<tr>
<td>strawman</td>
</tr>
<tr>
<td>false cause</td>
</tr>
<tr>
<td>appeal to emotion</td>
</tr>
<tr>
<td>the fallacy fallacy</td>
</tr>
<tr>
<td>slippery slope</td>
</tr>
<tr>
<td>ad hominem</td>
</tr>
<tr>
<td>tu quoque</td>
</tr>
<tr>
<td>personal incredulity</td>
</tr>
<tr>
<td>special pleading</td>
</tr>
<tr>
<td>loaded question</td>
</tr>
<tr>
<td>burden of proof</td>
</tr>
<tr>
<td>ambiguity</td>
</tr>
<tr>
<td>the gambler's fallacy</td>
</tr>
<tr>
<td>bandwagon</td>
</tr>
<tr>
<td>no true scotsman</td>
</tr>
<tr>
<td>genetic</td>
</tr>
<tr>
<td>black-or-white</td>
</tr>
<tr>
<td>begging the question</td>
</tr>
<tr>
<td>the texas sharpshooter</td>
</tr>
<tr>
<td>middle ground</td>
</tr>
<tr>
<td>appeal to authority</td>
</tr>
<tr>
<td>composition/division</td>
</tr>
<tr>
<td>appeal to nature</td>
</tr>
<tr>
<td>anecdotal</td>
</tr>
</tbody>
</table>
WHAT CAN BE DONE?

• There are probably many solutions, including better education and a ramping down of political zealotry from Our Glorious Leaders.

• But this is a class on natural language processing.

• Can we detect bias automatically from online texts?
LANGUAGE ANALYSIS AND LYING

“Don’t use a big word when a diminutive one would suffice.”

https://youtu.be/4ab2ZeZ-krY

HOW TO SPOT A LIAR

https://youtu.be/H0-WkpmTPrM
ASIDE: HOW CAN BIAS DETECTION HELP?

• **Social media platforms:**
  • May want to more closely monitor highly biased groups (e.g. allocate more human annotators to look for ban-able content like inciting violence or doxing).

• **Sociologists and network scientists:**
  • Better understanding of biased online communities can help us address the root causes of bias.
  • How do online communities become biased?
  • Are biased online communities uniformly biased?
**REDDIT CORPUS**

- We have curated data from Reddit by scraping subreddits, using Pushshift, by perceived political affiliation.

<table>
<thead>
<tr>
<th>Left (598,944)</th>
<th>Center (599,872)</th>
<th>Right (600,002)</th>
<th>Alt (200,272)</th>
</tr>
</thead>
<tbody>
<tr>
<td>twoXChromosomes (7,720,661)</td>
<td>news (2,782,9911)</td>
<td>theNewRight (19,466)</td>
<td>conspiracy (6,767,099)</td>
</tr>
<tr>
<td>occupyWallStreet (397,538)</td>
<td>politics (60,354,767)</td>
<td>whiteRights (118,008)</td>
<td>911truth (79,868)</td>
</tr>
<tr>
<td>lateStageCapitalism (634,962)</td>
<td>energy (416,926)</td>
<td>Libertarian (3,886,156)</td>
<td></td>
</tr>
<tr>
<td>progressive (246,435)</td>
<td>canada (7,225,005)</td>
<td>AskTrumpSupporters (1,007,590)</td>
<td></td>
</tr>
<tr>
<td>socialism (1,082,305)</td>
<td>worldnews (38,851,904)</td>
<td>The_Donald (21,792,999)</td>
<td></td>
</tr>
<tr>
<td>demsocialist (5269)</td>
<td>law (464,236)</td>
<td>new_right (25,166)</td>
<td></td>
</tr>
<tr>
<td>Liberal (151,350)</td>
<td></td>
<td>Conservative (1,929,977)</td>
<td></td>
</tr>
</tbody>
</table>

- These data are stored on the teach.cs servers under `/u/cs401/A1/data/`. These files should only be accessed from that directory (and not copied). All data are in the JSON format.
If you want to experiment a bit, there are some fields of metadata that might be interesting, but the main thing is **body**.
In order to infer whether the author of a given comment leans a certain way, politically, we use three steps:

1. **Preprocess** the data, so that we can extract meaningful information, and remove distracting ‘noise’.
2. **Extract** meaningful information.
3. **Train** classifiers, given labeled data.
1. Replace all newline characters with spaces.
2. Replace HTML character codes (i.e., &...;) with their ASCII equivalent.
3. Remove all URLs (i.e., tokens beginning with http or www).
4. Remove duplicate spaces between tokens.
5. Apply the following steps using spaCy:
   1. Tagging (dog -> dog/NN)
   2. Lemmatization (words/NNS -> word/NNS)
   3. Sentence segmentation
      1. (“I know words. I’ve got the best words” -> “I know words.
       I’ve got the best words”)

PREPROCESSING
PUTTING IT ALL TOGETHER

I know words. I’ve got the best words.

I/PRP know/VBP word/NNS ./PUNCT
I/PRP have/VB get/VBN the/DT good/JJS word/NNS ./PUNCT
Both lemmatization and stemming are often used to transform word tokens to a more base form.

- This helps to improve sparseness.
- It also helps in using various resources.
  - (e.g., funkilicious might not exist in a norm or embedding, but ‘funk’ ought to).
LEMMATIZATION VS STEMMING: DAWN OF SPARSENESS

- **lemma**: *n.* an abstract conceptual form of a word that has been mentally selected for utterance in the early stages of speech production.
  - E.g., \textit{lemma(best)} = \textit{good} (degree)
  - E.g. \textit{lemma(houses)} = \textit{house} (number/amount)
  - E.g. \textit{lemma(housing)} = \textit{housing}

- **stem**: *n.* usually, a part of a word to which affixes can be attached.
  - E.g. \textit{stem(houses)} = \textit{stem(housing)} = \textit{hous}

We use lemmatization given some of our features, but check out \texttt{nltk.stem} in the \texttt{NLTK} package.
PREPROCESSING: YOUR TASK

• Copy the template from /u/cs401/A1/code/a1_preproc.py. There are two functions you need to modify:
  • In preproc1, perform each preprocessing step above.
  • In main, replace the lines marked with TODO with the code they describe. Add a new `cat` field with the name of the class.

• The program takes three arguments:
  1. your student ID (mandatory),
  2. the output file (mandatory), and
  3. the maximum **number of lines** to sample from each category file (optional; default=10,000).

  python a1 preproc.py 999123456 -o preproc.json
import spacy

nlp = spacy.load('en_core_web_sm', disable=['parser', 'ner'])
sentencizer = nlp.create_pipe("sentencizer")
nlp.add_pipe(sentencizer)

utt = nlp(u"I know words. I have the best words")

for sent in utt.sents:
    ... print(sent.text)
    ... for token in sent:
        ... print(token.text, token.lemma_, token.pos_, token.tag_, token.dep_,
        ...         token.shape_, token.is_alpha, token.is_stop)
We provide our student IDs so we each see a different part of the available data.

- By default, you should only sample 10,000 lines from each of the Left, Centre, Right, and Alt files, for a total of 40,000 lines.

- From each file, start sampling lines at index $[ID \mod \text{len}(X)]$

Feel free to play around with more or less data, respectful of your peers on the servers, but this step guarantees it’s tractable (and that there’s no ‘desired’ level of accuracy).
The `a1_extractFeatures.py` program reads a preprocessed JSON file and extracts features for each comment therein, producing and saving a $D \times 174$ NumPy array, where the $i^{th}$ row is the features for the $i^{th}$ comment, followed by an integer for the class (0: Left, 1: Center, 2: Right, 3: Alt), as per the cat JSON.

```
{"id":"c05os7s",
"body":"wait! be you say that 9 / 11 be a * conspiracy *?! like ...
an * inside job * or something ?",
\texttt{cat:}"Alt"}
```

$i^{th}$ comment in input

$i^{th}$ row in output
1. Number of words in uppercase (≥ 3 letters long)
2. Number of first-person pronouns
3. Number of second-person pronouns
4. Number of third-person pronouns
5. Number of coordinating conjunctions
6. Number of past-tense verbs
7. Number of future-tense verbs
8. Number of commas
9. Number of multi-character punctuation tokens
10. Number of common nouns
11. Number of proper nouns
12. Number of adverbs
13. Number of wh- words
14. Number of slang acronyms
15. Average length of sentences, in tokens
16. Average length of tokens, excluding punctuation-only tokens, in characters
17. Number of sentences.
18. Average of AoA (100-700) from Bristol, Gilhooly, and Logie norms
19. Average of IMG from Bristol, Gilhooly, and Logie norms
20. Average of FAM from Bristol, Gilhooly, and Logie norms
21. Standard deviation of AoA (100-700) from Bristol, Gilhooly, and Logie norms
22. Standard deviation of IMG from Bristol, Gilhooly, and Logie norms
23. Standard deviation of FAM from Bristol, Gilhooly, and Logie norms
24. Average of V.Mean.Sum from Warringer norms
25. Average of A.Mean.Sum from Warringer norms
26. Average of D.Mean.Sum from Warringer norms
27. Standard deviation of V.Mean.Sum from Warringer norms
28. Standard deviation of A.Mean.Sum from Warringer norms
29. Standard deviation of D.Mean.Sum from Warringer norms

30-173. LIWC/Receptivity features
Warringer: These norms measure the valence (V), arousal (A), and dominance (D) of each lemma, according to the VAD model of human affect and emotion. See: Warriner, A.B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 English lemmas. Behavior Research Methods, 45:1191-1207.

Bristol et al: measure the age-of-acquisition (AoA), imageability (IMG), and familiarity (FAM) of each word, which we can use to measure lexical complexity. See: Gilhooly, KJ, Logie, RH (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words Behavior Research Methods, 12(4):394-427.
The Linguistic Inquiry & Word Count (LIWC) tool has been a standard in a variety of NLP research, especially around authorship and sentiment analysis.

- This tool provides 85 measures mostly related to word choice.

- The company Receptiviti provides a superset of these features, which also includes 59 measures of personality derived from text.

- To simplify things, we have already extracted these 144 features for you. Simply copy the pre-computed features from the appropriate uncompressed npy files stored in /u/cs401/A1/feats/.
Comment IDs are stored in _IDs.txt files (e.g., Alt_IDs.txt). When processing a comment, find the index (row) \( i \) of the ID in the appropriate ID text file, for the category, and copy the 144 elements, starting at element \( i \cdot 144 \), from the associated feats.dat.npy file.

```json
{"id":"c05os7s",
"body":"wait! be you say that 9/11 be a * conspiracy *?! like ...
... an * inside job * or something?",
"cat":"Alt"}"
```

Comment

```
... c05nn92 c05o811 c05os7s c05p5vj c05pbbg ...
```

Alt_IDs.txt

Start at 46 \( \cdot 144 \)

End at 47 \( \cdot 144 - 1 \)

Alt_feats.dat.npy
LIWC/RECEPTIVITI 3: FEATURE NAMES

liwc_sexual
liwc_shehe
liwc_social
liwc_space
liwc_swear
liwc_tentat
liwc_they
liwc_time
liwc_verb
liwc_we
liwc_work
liwc_you
receptiviti_active
receptiviti_adjustment
receptiviti_adventurous
receptiviti_aggressive
receptiviti_agreeableness
receptiviti_ambitious
receptiviti_anxious
receptiviti_artistic
receptiviti_assertive
receptiviti_body_focus
CLASSIFICATION

• Four parts:
  • Compare classifiers
  • Experiment with the amount of training data used
  • Select the best features for classification
  • Do cross-fold validation
Randomly split data into 80% training, 20% testing.

We have 5 classification methods, which you can consider to be ‘black boxes’ (input goes in, classes come out).

1. Support vector machine with linear kernel
2. Gaussian naïve Bayes classifier.
3. Random forest classifier
4. Neural network
5. Adaboost (with decision tree)
CLASSIFICATION I: COMPARE CLASSIFIERS

- **Accuracy**: the total number of correctly classified instances over all classifications: $A = \frac{\sum_i c_{i,i}}{\sum_i \sum_j c_{i,j}}$.

- **Recall**: for each class $\kappa$, the fraction of cases that are truly class $\kappa$ that were classified as $\kappa$: $R(\kappa) = \frac{c_{\kappa,\kappa}}{\sum_j c_{\kappa,j}}$.

- **Precision**: for each class $\kappa$, the fraction of cases classified as $\kappa$ that truly are $\kappa$: $P(\kappa) = \frac{c_{\kappa,\kappa}}{\sum_i c_{i,\kappa}}$.

$c_{i,j}$: number of times class $i$ was classified as class $j$
CLASSIFICATION 2: AMOUNT OF DATA

- You previously used a random $0.8 \cdot 40K = 32K$ comments to train.
- Using the classifier with the highest accuracy from Sec3.1, retrain the system using an arbitrary $1K, 5K, 10K, 15K, 20K$ samples from the original $32K$. 
**CLASSIFICATION 3: FEATURE ANALYSIS**

- Certain features may be more or less useful for classification, and too many can lead to various problems.
- Here, you will select the best $k$ features for classification for $k = \{5, 50\}$.
- Train the best classifier from Sec3.1 on just $k = 5$ features on both $1K$ and $32K$ training samples.
- Are some features always useful? Are they useful to the same degree ($p$-value)? Why are certain features chosen and not others?
CLASSIFICATION 4: CROSS-FOLD VALIDATION

• What if the ‘best’ classifier from Sec3.1 only appeared to be the best because of a random accident of sampling?
• Test your claims more rigorously.

<table>
<thead>
<tr>
<th>Iteration 1</th>
<th>Iteration 2</th>
<th>Iteration 3</th>
<th>Iteration 4</th>
<th>Iteration 5</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Part 1</td>
<td>Part 2</td>
<td>Part 3</td>
<td>Part 4</td>
<td>Part 5</td>
</tr>
</tbody>
</table>

: Err1 %     : Err2 %     : Err3 %     : Err4 %     : Err5 %
• You have **complete freedom** to expand on this assignment in any way you choose.

• You should have no expectation to the *value* of such an exploration – **check with us** (privately if you want) about the appropriateness of your idea.

• Bonus marks can make up for marks lost in other sections of the assignment, but your overall mark **cannot exceed 100%**.
• If things go well, we would love to run a special ‘workshop’ where:
  1. students who did interesting **bonuses** could describe their work
  2. grad students (working around the theme) could present their **projects**
  3. we could hold a **competition** for best systems in A1, A2, A3
• Problem: the instructors and TAs already have a lot on their plates.
• Solution (?): If any of you are interested in spearheading such a get-together at the end of the term (and getting bonus marks), we’d be glad to support.
RULES OF LOGARITHMS

• You may need these in later assignments:
  • **Definition:** \( \log_a x = N \leftrightarrow a^N = x \)
  • **Product:** \( \log_a (xy) = \log_a x + \log_a y \)
  • **Quotient:** \( \log_a \left( \frac{x}{y} \right) = \log_a x - \log_a y \)
  • **Power:** \( \log_a (x^p) = p \log_a x \)
  • **Base change:** \( \log_a x = \frac{\log_b x}{\log_b a} \)

• Reminder: avoid common logarithmotechnic errors:
  • \( \log_a (x + y) \neq \log_a x + \log_a y \)
  • \( \log_a (x - y) \neq \log_a x - \log_a y \)