Group: Real Wuglet Research Institute's Pet Shark Happiness Research Group

Adriana Knight, Bayliss Wagner, Jake Chanenson, Naomi Park

Proposal 1: Predicting Author Gender from Political Writing

We want to see if we can train a language model to be able to identify whether an author is male or female. We also want to use our NLP tools such as sentiment analysis to compare writing by women with writing by men; this will include creating a list of most common unique words used by men and women; compare usage of certain words by men vs. by women (removing most popular words to take Zipf's law into account); we will investigate the stereotype that women use more emotional language than men; and differences in topics that each group focuses on. We will likely have to source our own dataset from [NYT, WP, WSJ, The Atlantic, The National Review, New Yorker, BBC, Guardian].

Scholarship:

- Gender as a Variable in Natural-Language Processing: Ethical Considerations
- Gender Classification of Literary Works
- Deep Learning Network Models to Categorize Texts According to Author's Gender and to Identify Text Sentiment
- Investigating the Role of Emotion-Based Features in Author Gender Classification of Text
- Stylistic text classification using functional lexical features

Questions:

- What sorts of preexisting NLP tools can we use to complete our project?
- Possible data sources
    - We found several papers that created datasets of texts, Tweets or internet posts labeled by gender that we could use as a backup
- How many items in our data set do we need?/How many men and how many women?
    - We will be collecting these by hand, so would 50 men and 50 women items be acceptable?
    - Do we need to worry about the relative word length of each item in the data set?
    - Should we control for topic coverage (i.e. all items are covering the 2020 US election) or should we get a diverse range of items (Some US election coverage, some intl. Politics coverage, ect ect)?

Proposal 2: Comparing Privacy Policies Using Natural Language Processing

For this project, we propose an examination of the similarities of privacy policies for different online services across platforms within the same country and across different countries for the same platforms. **Using similar applications of word vectors as with Lab 3, we intend to generate measures of relative similarity across different identified groups of interest.** If there are broad similarities between all policies, we intend to perform a close reading of what privacy policies entail, and what data every single service collects. If there are significant differences between countries or platforms (or in the case of tiktok, between under-13 and 13+ uses), then we would aim to see where and how these policies differ. We would manually scrape

privacy policies from major social media platforms and services such as Facebook, Snapchat, Google Drive, TikTok, Instagram, LinkedIn, Zoom, and perhaps non-US based platforms like Weibo as well. Alternatively, we would look into some privacy policy corpuses created by the prior work listed below. **The goal of this project is to closely examine the general state of internet privacy, and open up avenues to think critically about what is and isn't being collected, as well as if there are any "safe havens" with comparably better privacy policies for their user base.**

Scholarship:

Analyzing Privacy Policies at Scale: on crowdsourced policy annotations and the effectiveness/usefulness of generalizing privacy policies using methods such as crowdsourcing and NLP. Probably most useful to us as background information.

An empirical study of natural language parsing of privacy policy rules using the SPARCLE policy workbench | Proceedings of the second symposium on Usable privacy and security: this team worked to create a tool called SPARCLE, built to help users parse privacy policies to identify policy elements and generate a machine readable (XML) version of the policy.

The Usable Privacy Policy Project: Combining Crowdsourcing, Machine Learning and Natural Language Processing to Semi-Automatically Answer Those Privacy Questions Users Care About: this team developed tools to semi-automatically extract key features for users to make informed decisions about their privacy as they interact with different websites, and also logs trends in the wording and content of privacy policies across different platforms.

The Creation and Analysis of a Website Privacy Policy Corpus: this team developed a corpus of 115 privacy policies (267K words) with manual annotations for 23K fine-grained data practices.

[2008.09159] Privacy Policies over Time: Curation and Analysis of a Million-Document Dataset: Analysis of privacy policy landscape and its changes over time. Their findings suggest a lack of reporting on tracking technology and third parties as well as policies themselves increasing in length and becoming more challenging to read.

Questions:

- How much do policies really differ across platforms?
- How similar are privacy policies for the same platforms across different countries (do rules significantly change depending on what country you are in)?
- Is there a measure we can use to determine the relative reading level of different proposals (complexity of vocabulary, length, etc.)?
- For tik tok specifically: how do we distinguish the 13+ and under-13 privacy policies, and are children actually being protected by the differences?
- Are there persistent tools to help us determine the similarity of content in our assembled corpses?