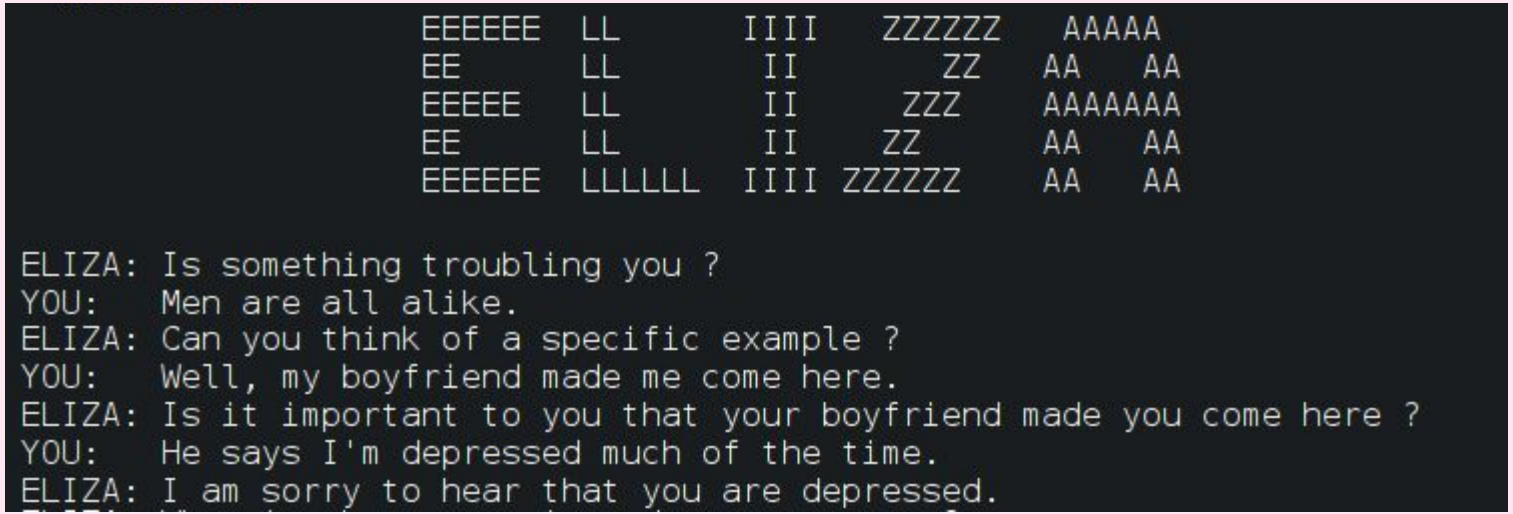
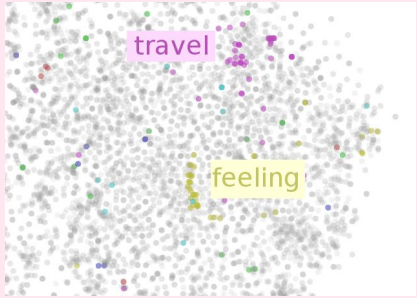
Bias in Natural Language Processing

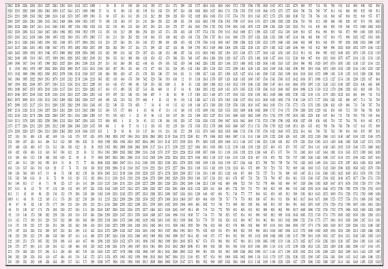
Impacts and challenges of biased NLP algorithms



What is Natural Language Processing?

- Automatic analysis and representation of human language, used to extract and categorize information, or reproduce relevant text





- Today’s contemporary techniques work off of statistical observations on large training corpora (millions, billions, or even trillions of words)

- NLP models recognize and reproduce patterns by studying many, many examples

Uses of Natural Language Processing

- Information retrieval (e.g. search)
- Sentiment analysis (automated moderation in social media)
- Information extraction
- Machine translation
- Natural language generation

What’s so Bad about a Little Language [Processing]?

- NLP models encode and replicate bias learned from training documents
- Example: A word embedding trained on a corpus of New York Times articles disproportionately associated common European-American names with pleasant terms, and common African-American names with unpleasant terms
- Bias embedded deep within a model is then reproduced in various applications, such as search results, applicant screening or automatically generated text
- Example: A GPT-3 model trained on Common Crawl frequently generated highly toxic sentences, even when prompts were non-toxic.

Why Haven’t We Debiased Our Models?

- Bias is difficult to measure, and those most affected are not always involved in creating and training models
- Contemporary NLP models require more training data than can be effectively curated
  - Example: One popular “clean” training corpus was made by discarding any page with words from a list of “Dirty, Naughty, Obscene or Otherwise Bad Words.”

- Neural Networks used in training are opaque and uninterpretable
- Language changes over time
- New models are developed and deployed faster than they can be studied

Steps, Research and Goals

- Bias in NLP models is an active area of research, hopefully allowing for better techniques for mitigating bias and recognizing problematic training documents
- More energy needs to be dedicated to documenting and curating training corpora, with more input and participation from marginalized groups
- Evaluating bias and harm in language models needs to happen during development (pre-mortem) instead of after deployment
- Research into alternatives to ever-growing language models

References and Resources

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big. Bender, McMillan-Major, Gebru, Shmitchell.

Understanding the Origins of Bias in Word Embeddings. Brunet, Alkalay-Houlihan, Anderson, Zemel.

Semantics derived automatically from language corpora contain human-like biases. Caliskan, Bryson, Narayanan.