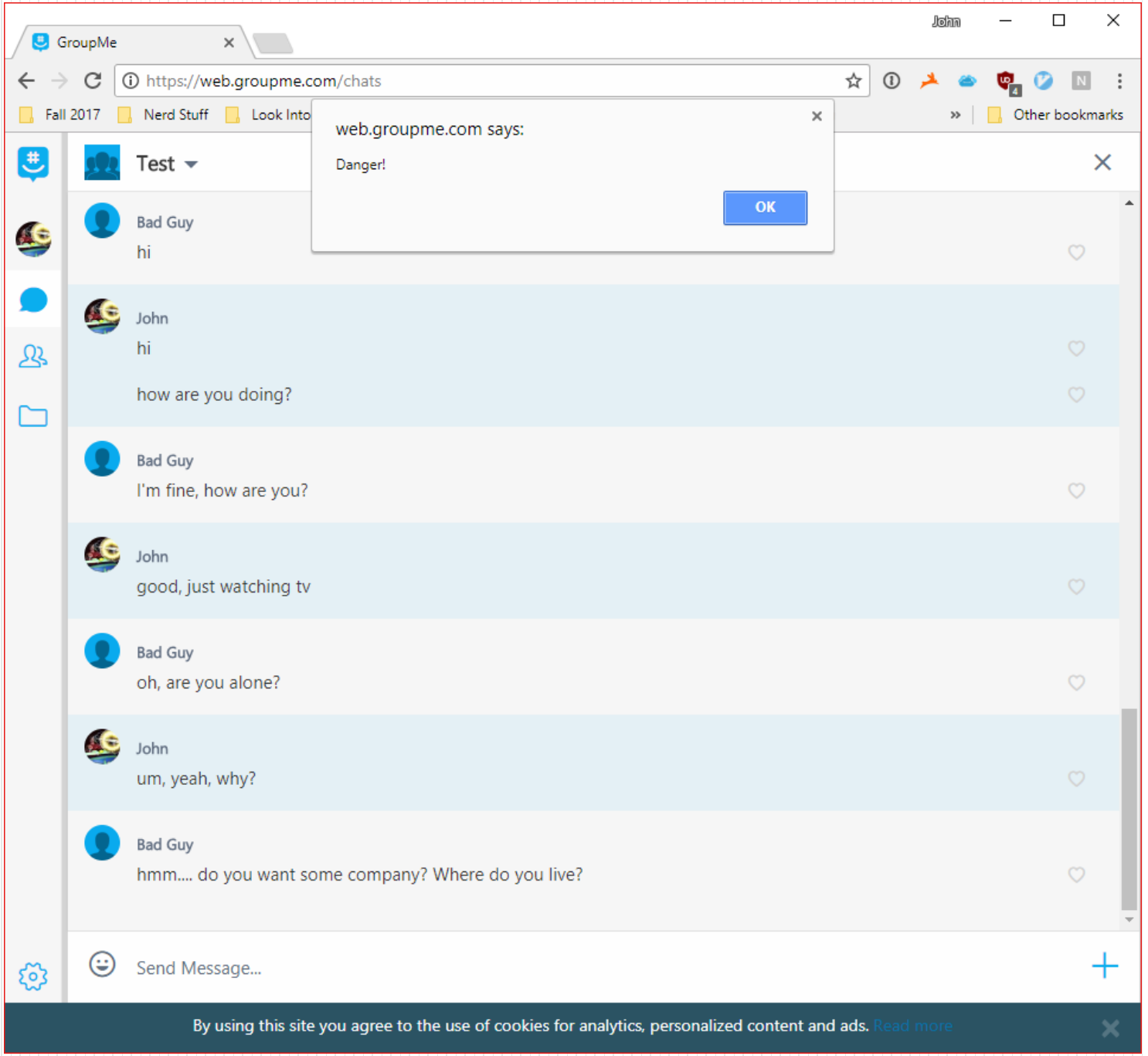
Trying to Make the Internet Safer

John Westhoff

Goal

- What I want: To make browsing the internet more safe from phishing and online predators
- Why: Some people don’t know any better
  - Service Projects – Logan Center
- How: A Google Chrome plugin that warns users about dangerous situations



*"artist"'s rendering

Method

- Getting input:
  - Plugin periodically takes all of the text on the webpage and preprocesses it
    - That includes tokenizing, forcing everything to lowercase, and removing punctuation
- Processing:
  - Two models: one for detecting predatory or not, and one for detecting phishing or not
  - Currently working on using RNNs as classifiers
- Training:
  - I have been training my predatory models on a corpus of roughly one million chat messages, 4% of which were sent by online predators
  - I have been training my phishing models on a corpus of 500 phishing emails plus 9,000 of my own non-phishing emails – I am also looking for a larger dataset

Experiments

- So far only tested a simple LSTM model for predator/not-predator
  - In the interest of development time I am training on the first 60K messages and evaluating on the next 40K, rather than all of them
  - ~30% false positive rate
  - ~12% false negative rate

Conclusions

- Phishing vs non-phishing seems easier to detect than predatory vs non-predatory, but I need more data for that classifier
- Having a big imbalance in classes is a little hard to deal with

Questions?