Twitter Trolling

•••

Gowtham Ashok

Introduction

During this summer, I worked with Professor Jennifer Golbeck on her “Trolling” Project.

I designed an interface for coding tweets as to whether they are hateful towards someone, trolling, threatening someone, or just an innocent tweet. I coded some tweets myself (around 5000). I then performed some analysis on the dataset I had collected so far, and wrote a primitive agent that detected potentially hateful conversation.

In order to learn more about the project, I took a social media analysis class as well as an AI class during the summer.

Codebook used

Trolling

If the tweet is posted with the *intent* of upsetting or offending readers.

Hate

Hate speech threatens or insults a group of people, based on religion, sexual orientation, race, gender, or any other trait.

Threat

If it contains a specific threat either at an individual or a group. Implied threats or endorsements of violence are not labeled as threats here.

Potentially Offensive

“Potentially offensive” content simply to allows coders to express displeasure at text.

Who did I work with?

I worked with Professor Jen Golbeck on her

“Trolling” Project

The people who classified tweets were

UMD Graduate students

Overview

Created an interface for classifying tweets

Learned about Social media research by:

- Designing and implementing an agent
- Working with tweets and visualization tools
- Reading papers on Social Media Research

Learned about Visualization techniques

Worked with Natural Language Processing

Design

Interface

I wrote an interface and made sure it was

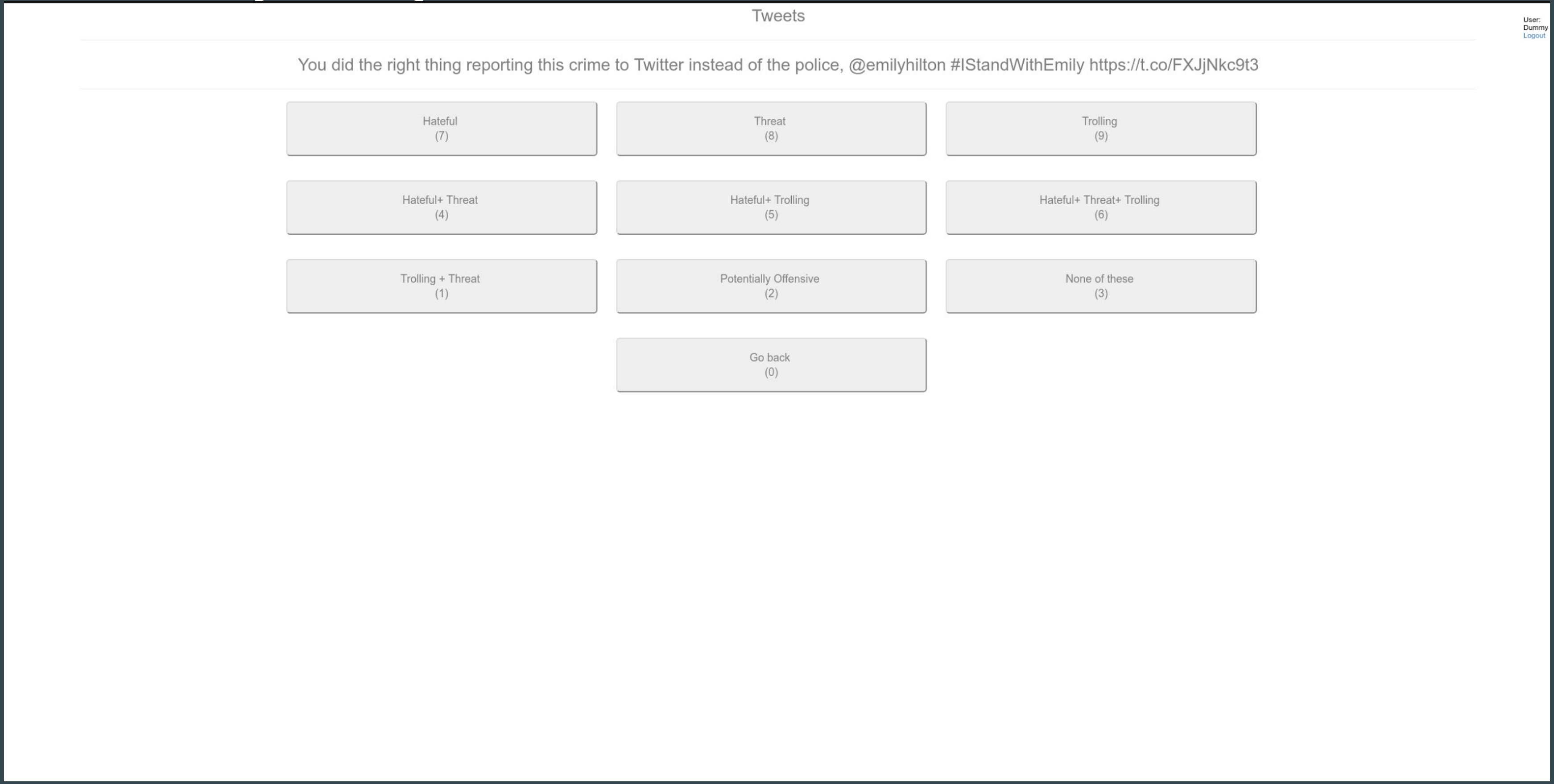
- Simple
- Easy-to-use
- Followed Responsive Design

The Graduate students helped me improve the

Interface by providing their feedback. I also used time taken to classify tweets by them, to quantitatively analyze how efficient the interface was.

I learned a lot about responsive design, PHP, Javascript and HTML coding.

Interface (contd)



You did the right thing reporting this crime to Twitter instead of the police, @emilyhilton #IStandWithEmily https://t.co/FXJjNkc9t3

Research

Coding Tweets

I coded around 5,000 tweets based on the codebook, for the internship.

Visualizations

I created visualizations by processing data from the coded tweets and uncoded tweets.

Currently the number of coded tweets are too low to get a really meaningful result. In the coming slides, you can see some results, but I find them to be too limited to have a really meaningful result.

I explored various visualization options and learned about how to use them effectively for different types of datasets, to convey information (I am only showing one type here)

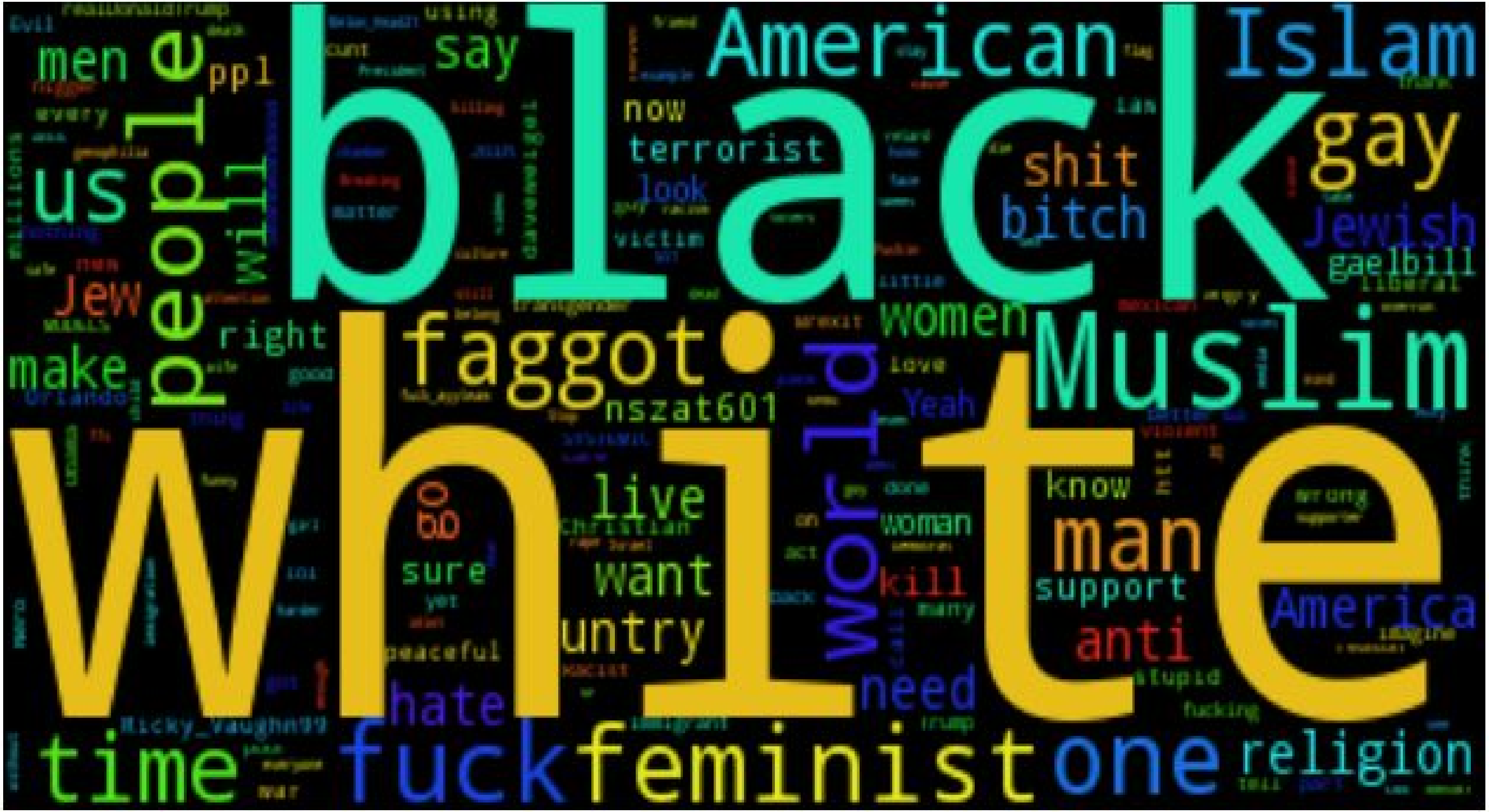
Visualizations (Contd)

These are the results of the visualizations I got so far.

I separated the tweets according to the dataset; performed some basic text processing, and then made it into a word cloud, that displays frequently occurring words.

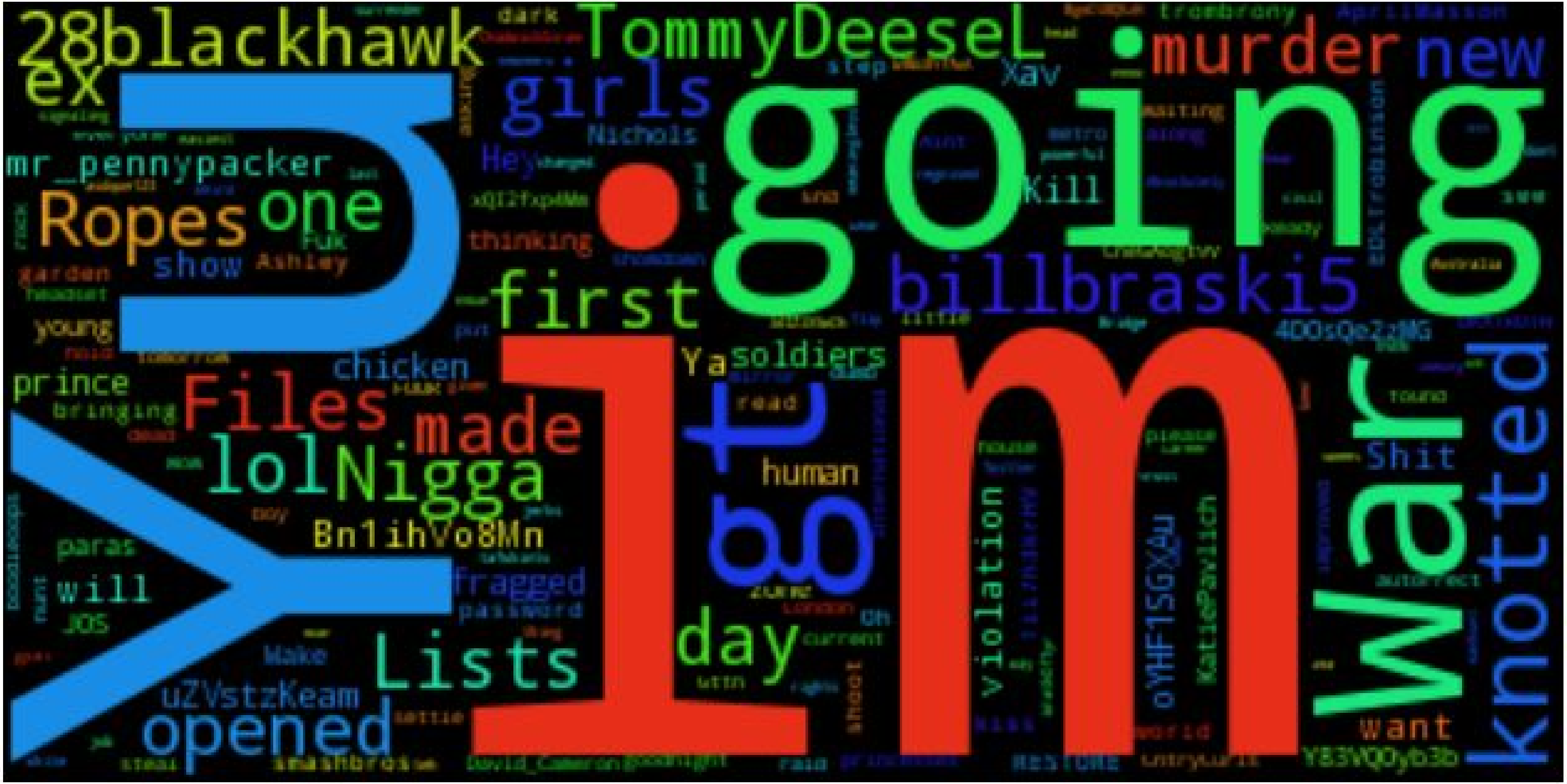
During the summer, I also learned to work with tools like Gephi. It did not provide any meaningful visualization with the data I have.

hateful

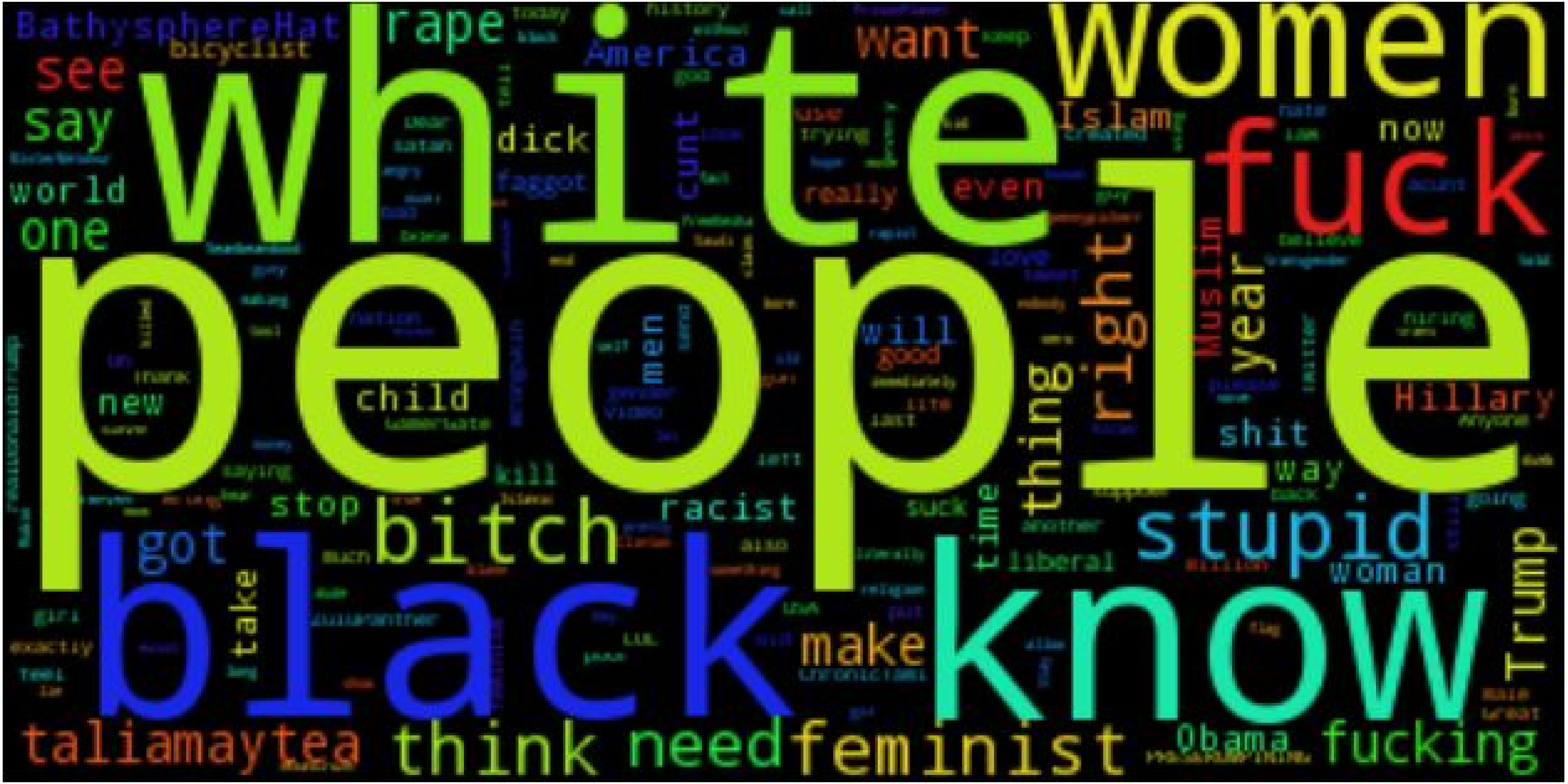


black
white
American
Islam
people
Muslim
faggot
gay
men
us
will
say
ppl
Jew
make
right
time
fuck
feminist
one
religion
America
man
anti
support
kill
want
live
need
hate
world
women
woman
shit
bitch
Jewish
gaelbill
liberal
terrorist
now
look
victim
nszat601
Yeah
know
sure
untry
Christian
peaceful
Racist
stupid
fucking
love
every
Orlando
Ricky_Vaughn99
immigrant
transgender
culture
matter
good
violent
wrong
imagine
trump

threat



you
im
going
war
get
28blackhawk
TommyDeeseL
murder
new
ex
girls
Nichols
Xav
Hey
mr_pennypacker
Ropes
one
Kill
billbraski5
first
show
Ashley
thinking
garden
young
soldiers
Ya
chicken
prince
bringing
Files
made
read
lolNigga
human
knoted
Shit
violation
Bn1ihVo8Mn
paras
will
fragged
password
day
Lists
uZVstzKeam
opened
want
Y83VQOyb3b
4DOsQe2zMG
oYHF15GXAu
KatiePavlich
smashbros
goodnight
dark
trombrony

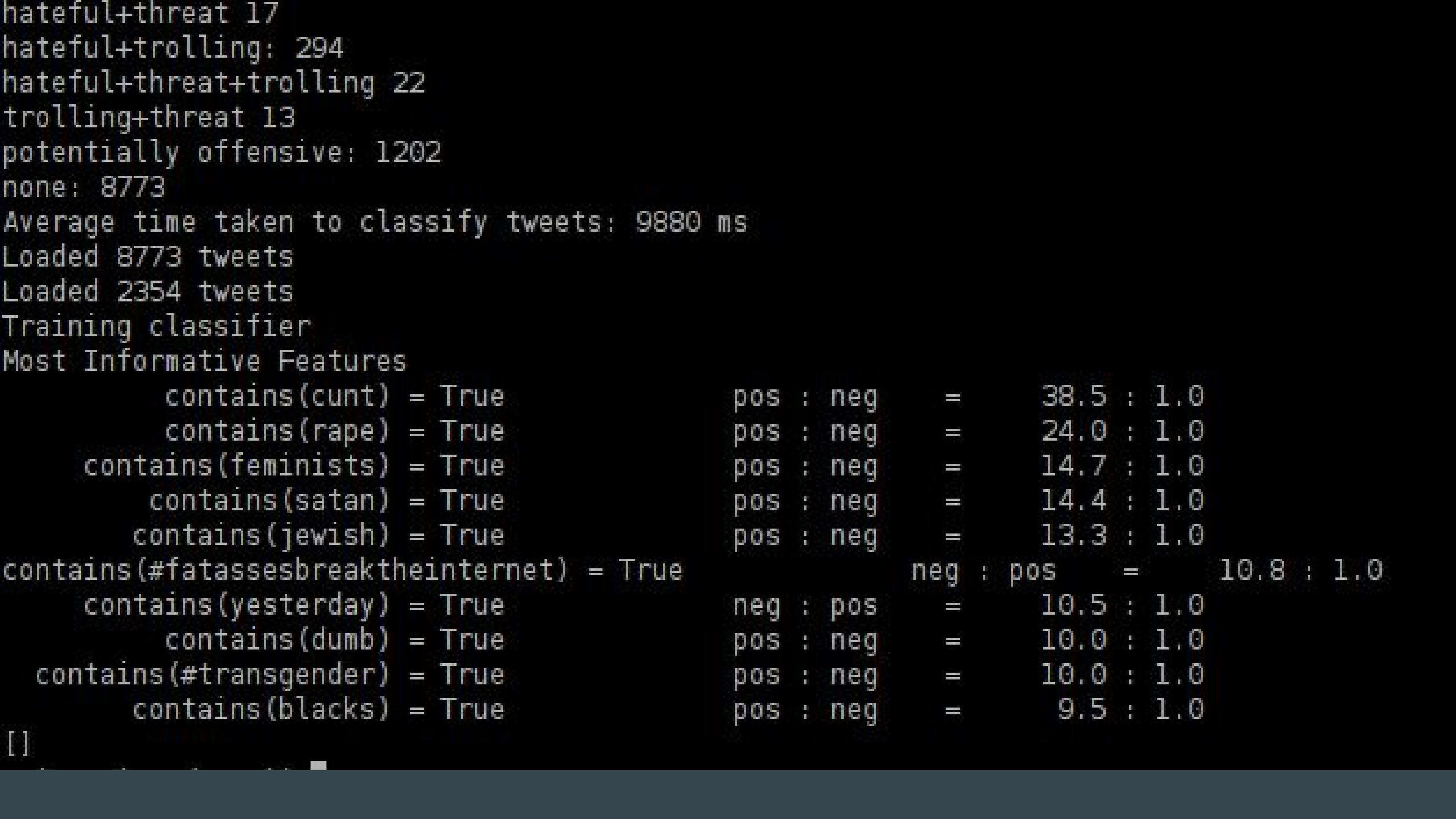

trolling

people
white
women
black
know
fuck
stupid
bitch
feminist
think
need
taliamaytea
fucking
Obama
rape
want
America
BathysphereHat
bicyclist
see
say
world
one
Islam
now
dick
faggot
cunt
really
even
right
thing
year
Muslim
Hillary
shit
way
woman
Trump
liberal
time
suck
racist
stop
got
take
make
child
new
men
will
good
kill
love

Natural Language Processing

Having no prior experience with NLP, I wrote code using the Python NLTK to extract useful information from the coded datasets, by considering the all the tweets classified into one of the categories (Trolling, Threat, Hateful, Potentially Offensive, …) as a positive training set, while considering the tweets classified as “None” to be the negative training set.

While doing this, I found that the dataset used here (tweets) is too terse to get useful information only using Natural Language Processing. I did get some meaningful data out of it (an offensive word list), but it may have many mislabeled words as it doesn’t take context or semantics into account.



```
hateful+threat 17
hateful+trolling: 294
hateful+threat+trolling 22
trolling+threat 13
potentially offensive: 1202
none: 8773
Average time taken to classify tweets: 9880 ms
Loaded 8773 tweets
Loaded 2354 tweets
Training classifier
Most Informative Features
          contains(cunt) = True              pos : neg    =     38.5 : 1.0
          contains(rape) = True              pos : neg    =     24.0 : 1.0
     contains(feminists) = True              pos : neg    =     14.7 : 1.0
         contains(satan) = True              pos : neg    =     14.4 : 1.0
        contains(jewish) = True              pos : neg    =     13.3 : 1.0
contains(#fatassesbreaktheinternet) = True              neg : pos    =     10.8 : 1.0
     contains(yesterday) = True              neg : pos    =     10.5 : 1.0
          contains(dumb) = True              pos : neg    =     10.0 : 1.0
  contains(#transgender) = True              pos : neg    =     10.0 : 1.0
        contains(blacks) = True              pos : neg    =      9.5 : 1.0
[]
```

Agent

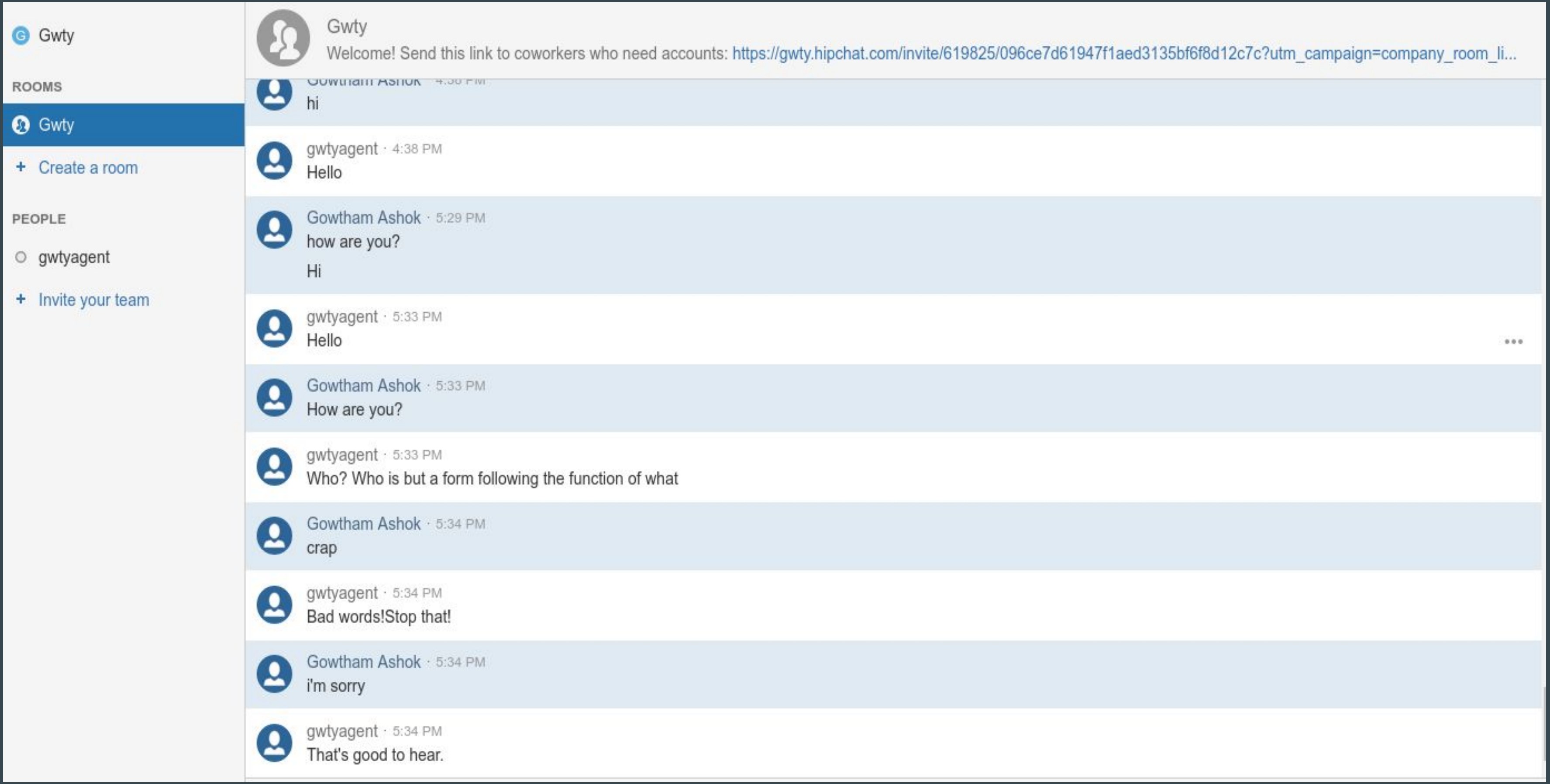
I designed an agent that could detect trolling. It is in a very primitive stage and is being improved upon continuously. I integrated it with HipChat to be easily accessible.

It uses a “bad word” dataset extracted from the most common words used in offensive tweets (Trolling, Threat, Hateful) to alert the chatroom if any such word is detected.

I did mockups on how this agent should behave and how it should handle trolling.

In designing this agent, I read a lot on the current state of Knowledge-based AI, and familiarized with myself in the techniques used in the field.

Chatting with the agent using HipChat



Gwty

Welcome! Send this link to coworkers who need accounts: https://gwty.hipchat.com/invite/619825/096ce7d61947f1aed3135bf6f8d12c7c?utm_campaign=company_room_li...

Gowtham Ashok · 4:38 PM
hi

gwtyagent · 4:38 PM
Hello

Gowtham Ashok · 5:29 PM
how are you?

Hi

gwtyagent · 5:33 PM
Hello

Gowtham Ashok · 5:33 PM
How are you?

gwtyagent · 5:33 PM
Who? Who is but a form following the function of what

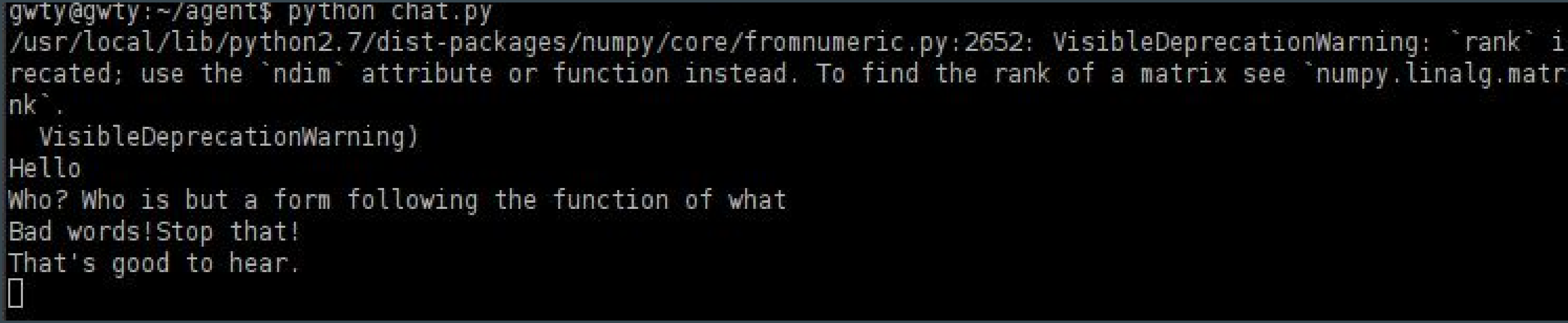
Gowtham Ashok · 5:34 PM
crap

gwtyagent · 5:34 PM
Bad words!Stop that!

Gowtham Ashok · 5:34 PM
i'm sorry

gwtyagent · 5:34 PM
That's good to hear.

Running the agent program on my local computer



```
gwty@gwty:~/agent$ python chat.py
/usr/local/lib/python2.7/dist-packages/numpy/core/fromnumeric.py:2652: VisibleDeprecationWarning: `rank` i
recated; use the `ndim` attribute or function instead. To find the rank of a matrix see `numpy.linalg.matr
nk`.
  VisibleDeprecationWarning)
Hello
Who? Who is but a form following the function of what
Bad words!Stop that!
That's good to hear.
```

Team Dynamics

Professor Jennifer Golbeck helped me in the design of the interface, and introduced me to the world of Social Media Research

The rest of the team was mostly involved in coding the tweets. I participated in discussions about what constitutes a “Troll”

They helped me in user studies about the interface. I made changes to the interface based on their feedback

Coursework

Introduction to HCI - For designing the Interface

Design Methods - For designing the Interface and Agent

Advanced Usability Testing - For performing user testing on the Interface

Research Methods - For performing Literature Review, analyzing data for Research

Social Media Analysis - For understanding social networks and analyzing them

What I learned from this internship

Good Design and Good Research very hard to do correctly

How to be self-guided

Designing and implementing interfaces that satisfies everyone’s needs is hard

Social Media Research is an interesting field

Natural Language Processing is hard

Conclusion

This internship was very useful to me, in making me know what Social Media Research was about. I learned about coding tweets, some Natural Language Processing, and some visualization techniques. It made me excited about this field. I plan on doing a capstone project in the same area.