Identification of Toxic Comments in Online Platforms

Ehsan Montazeri
Mehvish Saleem
Ramanpreet Singh

INTRODUCTION

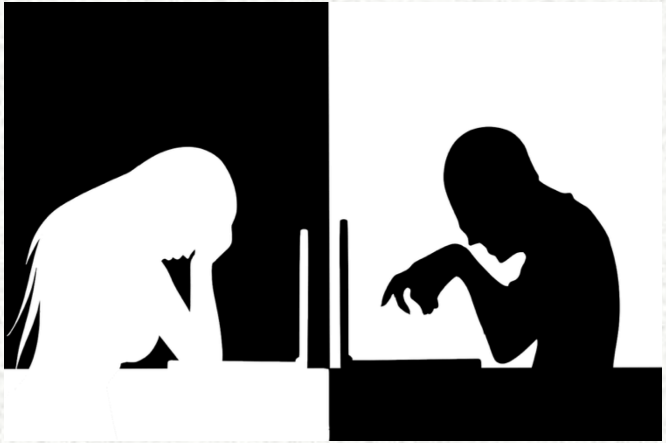
Introduction

- Identifying toxic comments in multiple online communities
    - Wikipedia Talk Pages, SFU Opinion and Comments Corpus (SOCC), Facebook
- Categorizing the different types of toxicities
    - Racism, sexism, bullying
- Comparing different communities

MOTIVATION

Motivation

- Toxicity in social interactions is very common
  - Over 43% of teens have been victims of cyberbullying. [1]
- Can have severe repercussions such as low self esteem, health problems, depression and isolation
  - Over 64% of victims reported that online toxicity negatively impacted their feelings of safety and ability to learn at school.
- Automatically detecting toxicity can help platform moderators to remove toxic comments and block users



FUCK FUCK
NIGGER NIGGER
FAT JEW
WANKER WANKER
hi moron
SHIT SHIT
PIG PIG
hate hate
Bark Bark
BALLS BALLS
Nipple Nipple
Go fuck
yourselfgo fuck
article
Wikipedia
faggot faggot
SUCK SUCK
fuck Go
think
people
edit
AIDS AIDS
SHUT FUCK
HUGE FAGGOT
piece shit
name
blocked
problem
fucksex fucksex
BOOBS BOOBS
ass ass
want
fuck yourselfgo
someone
much
look
SUPER GAY
way
SUCK COCK
need
fucking
little
NOOBS wiki
nothing
DONT CARE
going
make
DIE FAG
now
put
man
BUTTSECKS BUTTSECKS
penis penis
JEW FAT
FUCK SHUTJEW
FAG DIE
poop poop
heil hitler
bastard Pro
hitler heil
really
mean
got
also
COCKSUCKER MOTHJER
FREEDOM FREEDOM
know FGGT
something
even
Hanibal911You're bastard
Bastered Bastered
block
Please
time
thing
site
good
made
FUCK DOG
MEXICANS SUCK
FAGGOT HUGE
SUCK DICK
know
will
Pro Assad
wiki NOOBS
word
gu/
one
see
talk page
FGGT know
fact
delete
moron hi
dickhead dickhead
anything
faggots faggots
right
HATE WIKIPEDIA
MOTHJER FUCKER
editor
read
stop
idiot
life
never
stupid
say
GAY SUPER
source
page
DIE DIE
FUCKER COCKSUCKER
SUCK MEXICANS
CARE DONT
use
back
said

DATASETS

Datasets

- Train our model using Wikipedia and SOCC datasets [2-3]
  - Comments labeled by human raters for toxic content
  - Roughly about 160000 comments in the Wikipedia dataset and 1000 comments in the SOCC dataset
- Evaluate the performance on comments from multiple Facebook pages
  - Sampled data from three categories: Sports, Politics and Entertainment

CHALLENGES

Challenges

- Data Collection
    - Challenging to find good data sets for training
    - Toxic comments are deleted by moderators
- Data Cleaning
    - ASCII characters, GIFs, pictures, ...
    - Missing values
- Model Selection
    - Achieving good performance

PROGRESS

PROGRESS

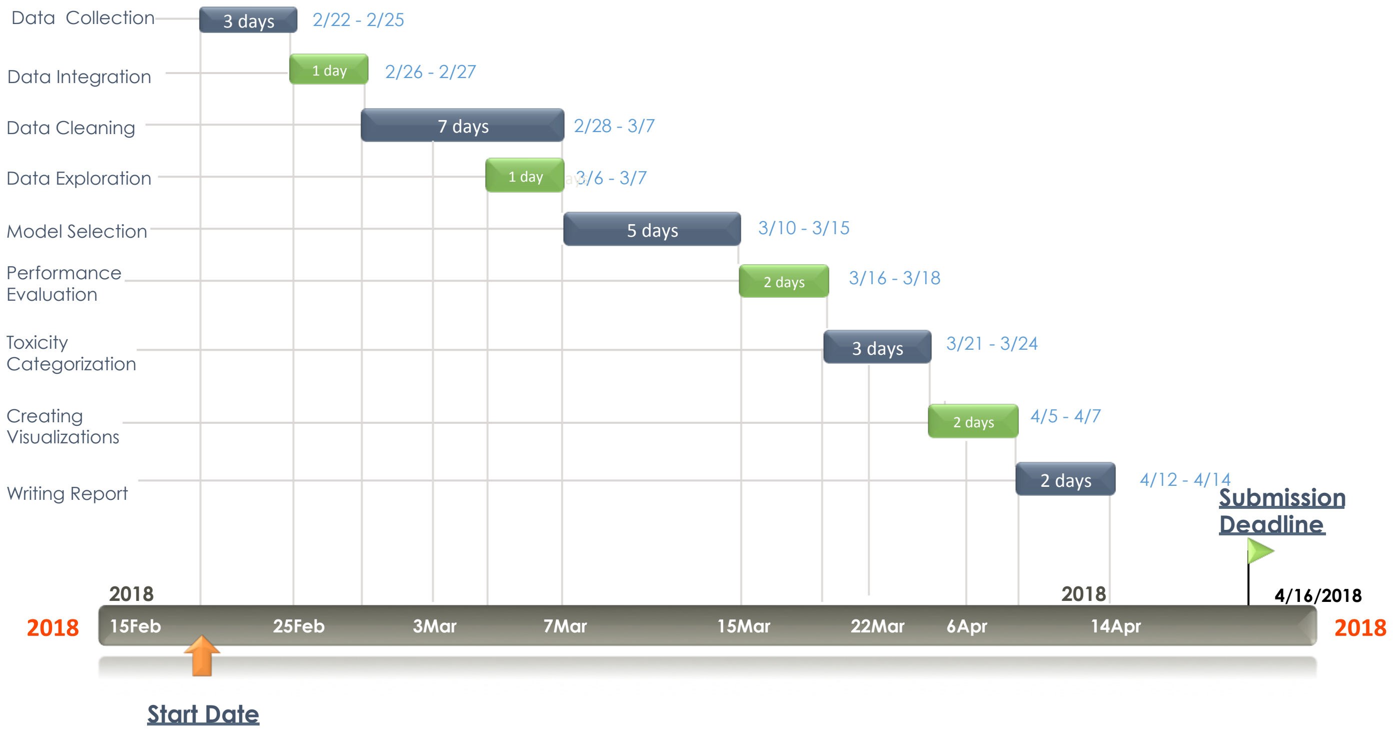
- Scraped data from facebook pages, sampled a subset of about 10000 comments from each category (Sports, Entertainment, Politics)
- Integrated data from multiple sources.
- Cleaning Data (In -Progress)

FUTURE WORK

Next Steps

- Data exploration (e.g. checking for class imbalances, missing values, ...)
- Data preparation (feature engineering)
- Building the model
- Toxicity categorization
- Visualization

Project Timeline



References

[1] DoSomething.org

[2]https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[3] https://github.com/sfu-discourse-lab/SOCC