Multilingual Toxic Comment Classification

Author: Drenizë Rama

TABLE OF CONTENTS

"*If you want to feel shit about the state of the world and everyone in it, Facebook comments are a great place to start.*" a LinkedIn article states.

How is it possible, though, that the very democratic element of the internet - the **comment section** - has become such a tool of cruelty?

Even the number one internet culture curator *WIRED* magazine wipes the slates clean, by stating:

> INTERNET RULE #1: Never read the comments.

The goal of this project is to build a natural language algorithm that classifies text input (social media comments) into toxic and non-toxic categories in multiple languages.

In order to achieve this, I decided to go with pretrained RNN models like BERT or LASER, which include a variety of languages.

EDA - The Data

4 datasets:

```
train_data.head(2)
```

	id	comment_text	toxic	severe_toxic	obscene	threat	insult	identity_hate
0	0000997932d777bf	Explanation\nWhy the edits made under my usern...	0	0	0	0	0	0
1	000103f0d9cfb60f	D'aww! He matches this background colour I'm s...	0	0	0	0	0	0

```
valid_data.head(2)
```

	id	comment_text	lang	toxic
0	0	Este usuario ni siquiera llega al rango de ...	es	0
1	1	Il testo di questa voce pare esser scopiazzato...	it	0

```
test_data.head(2)
```

	id	content	lang
0	0	Doctor Who adlı viki başlığına 12. doctor olar...	tr
1	1	Вполне возможно, но я пока не вижу необходимо...	ru

EDA - The Data

And an extended dataset for further exploration:

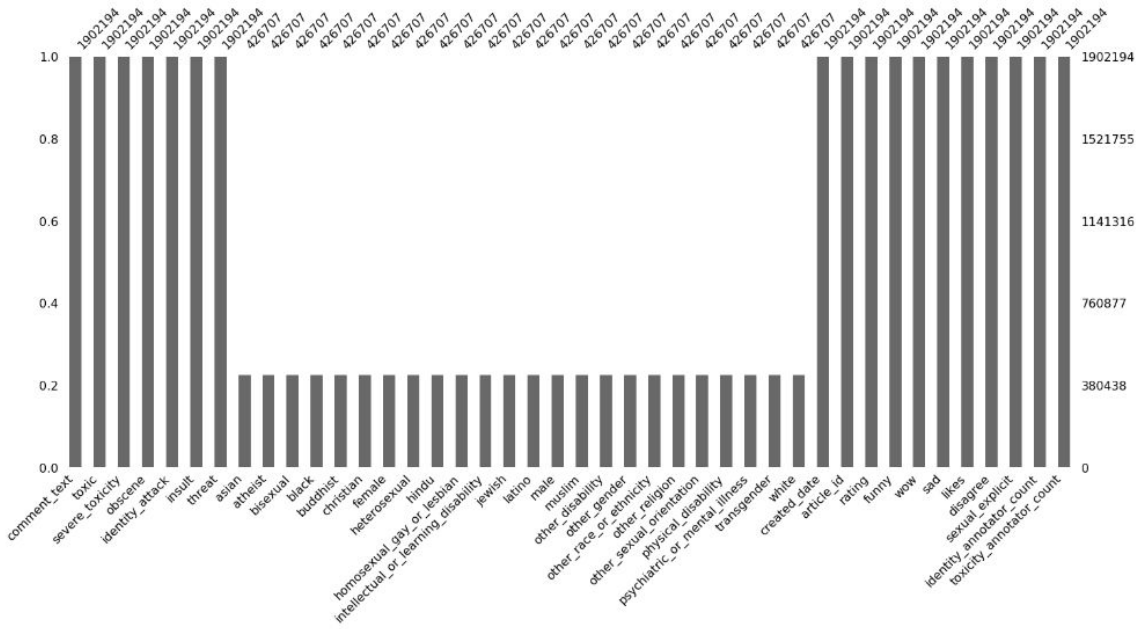
```
data.columns
```

```
Index(['comment_text', 'toxic', 'severe_toxicity', 'obscene',
       'identity_attack', 'insult', 'threat', 'asian', 'atheist', 'bisexual',
       'black', 'buddhist', 'christian', 'female', 'heterosexual', 'hindu',
       'homosexual_gay_or_lesbian', 'intellectual_or_learning_disability',
       'jewish', 'latino', 'male', 'muslim', 'other_disability',
       'other_gender', 'other_race_or_ethnicity', 'other_religion',
       'other_sexual_orientation', 'physical_disability',
       'psychiatric_or_mental_illness', 'transgender', 'white', 'created_date',
       'article_id', 'rating', 'funny', 'wow', 'sad', 'likes', 'disagree',
       'sexual_explicit', 'identity_annotator_count',
       'toxicity_annotator_count'],
      dtype='object')
```

```
data.shape
```

```
(1902194, 42)
```

The Data

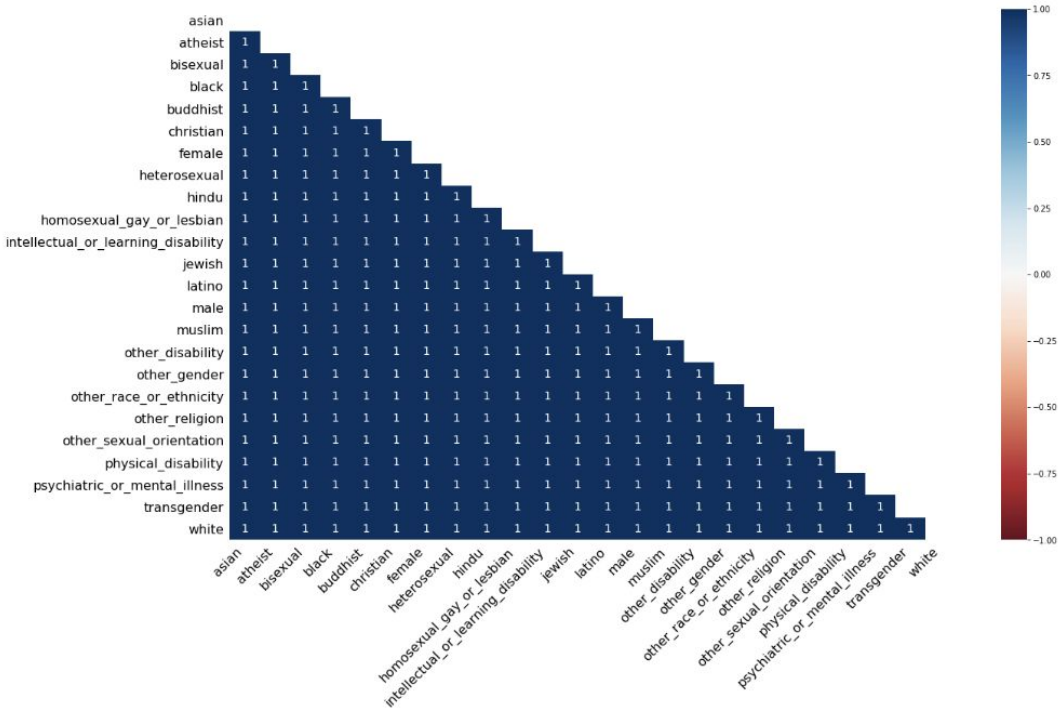


What do we see in the bar chart above?

- There is an equal amount of missing values in all of the identitiy features.

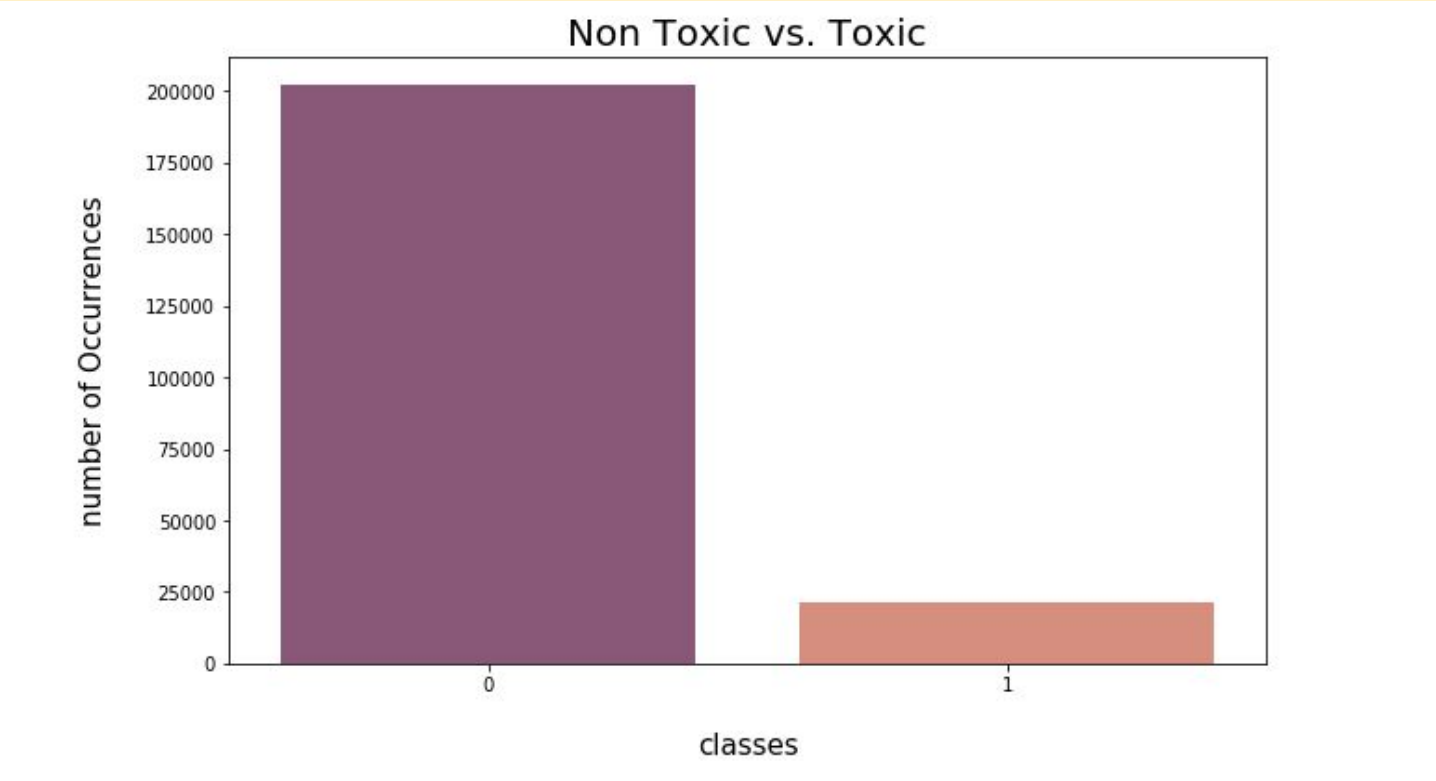
This may indicate that the type of identity hate has been evaluated only later on. However, it is more likely that two datasets have been joined.

The Data

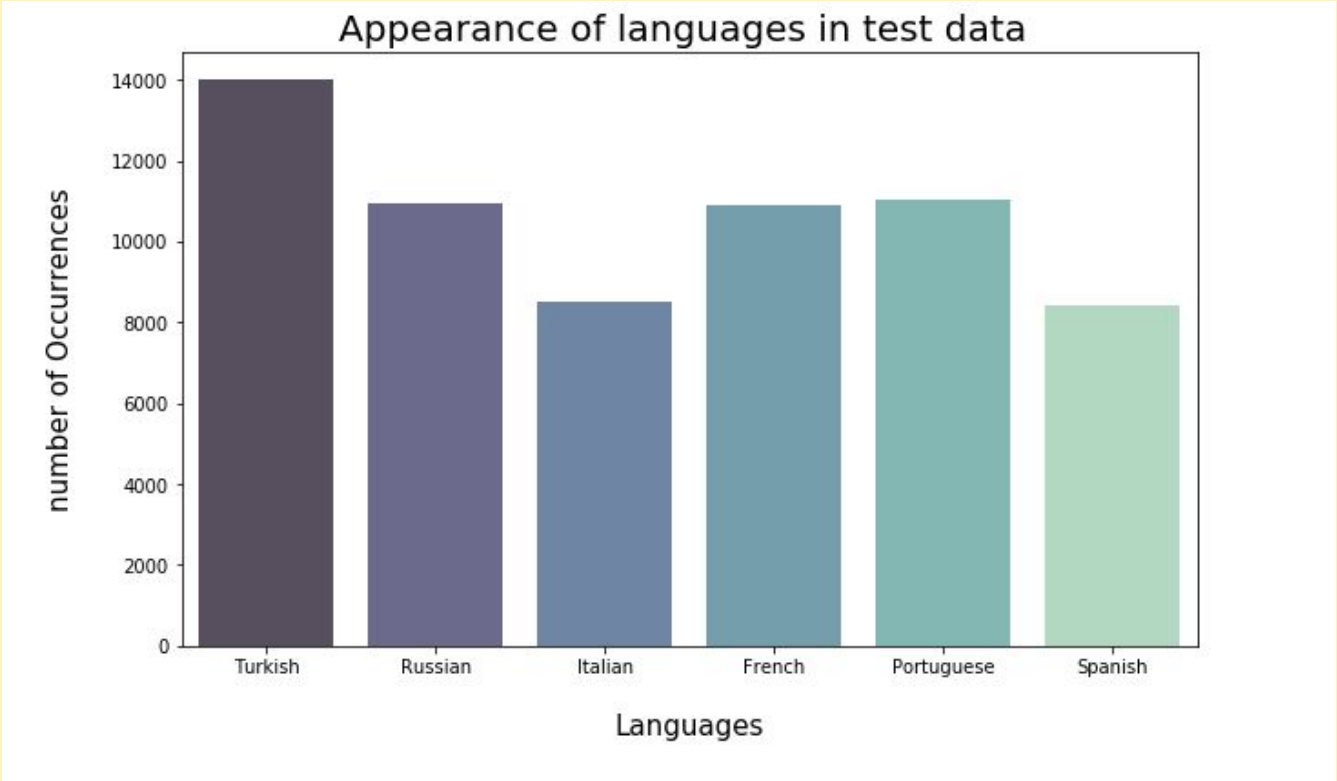


The perfect value of 1 in the *missing values heatmap* confirms that if one identity feature is present the other identity features are also present, suggesting that they have been added at the same time. This is a systematic error, so just going with the 'complete cases analysis' and reducing the sample to complete cases won't do. This ignorance might reduce analysis precision significantly. And that's surely not what we want.

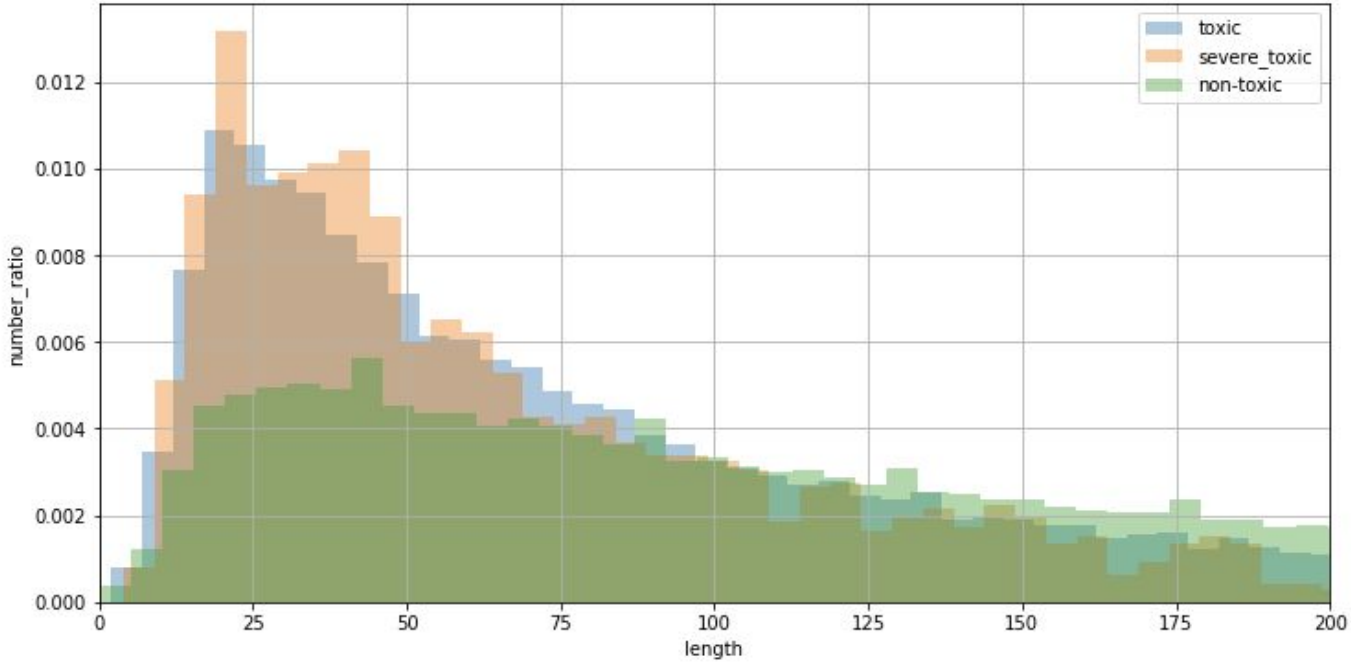
The Data


Non Toxic vs. Toxic
number of Occurrences
200000
175000
150000
125000
100000
75000
50000
25000
0
0
1
classes

The Data

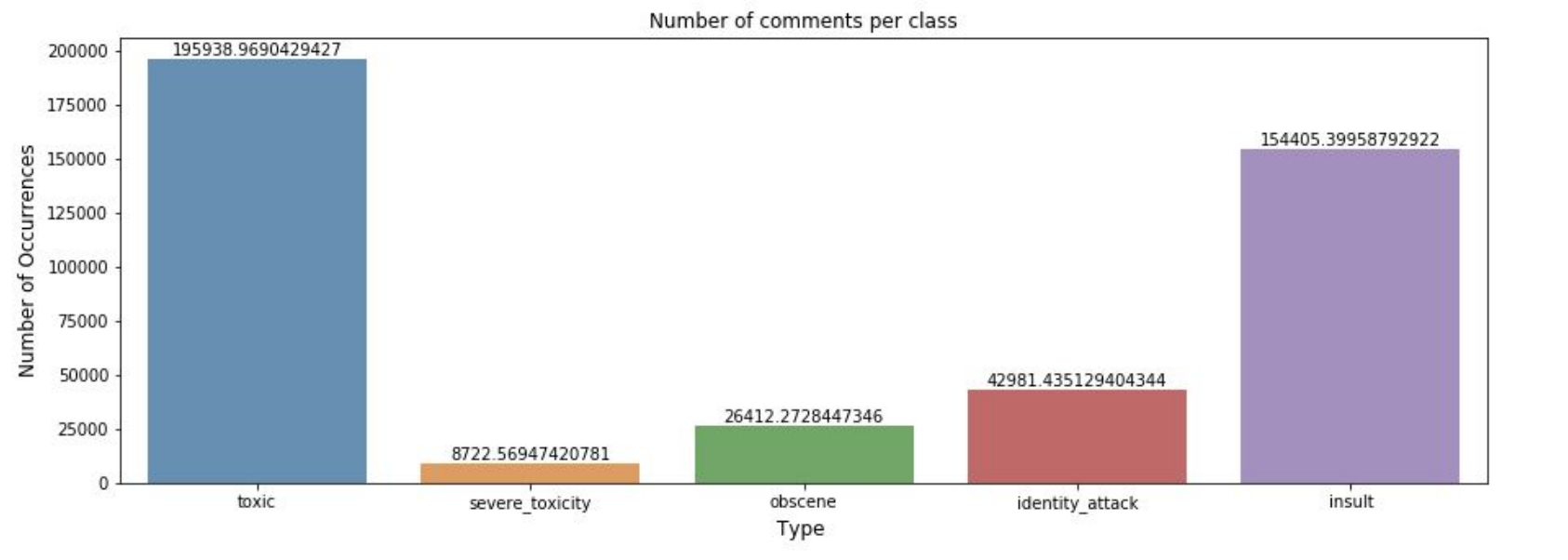

Appearance of languages in test data
number of Occurrences
14000
12000
10000
8000
6000
4000
2000
0
Turkish
Russian
Italian
French
Portuguese
Spanish
Languages

The Data

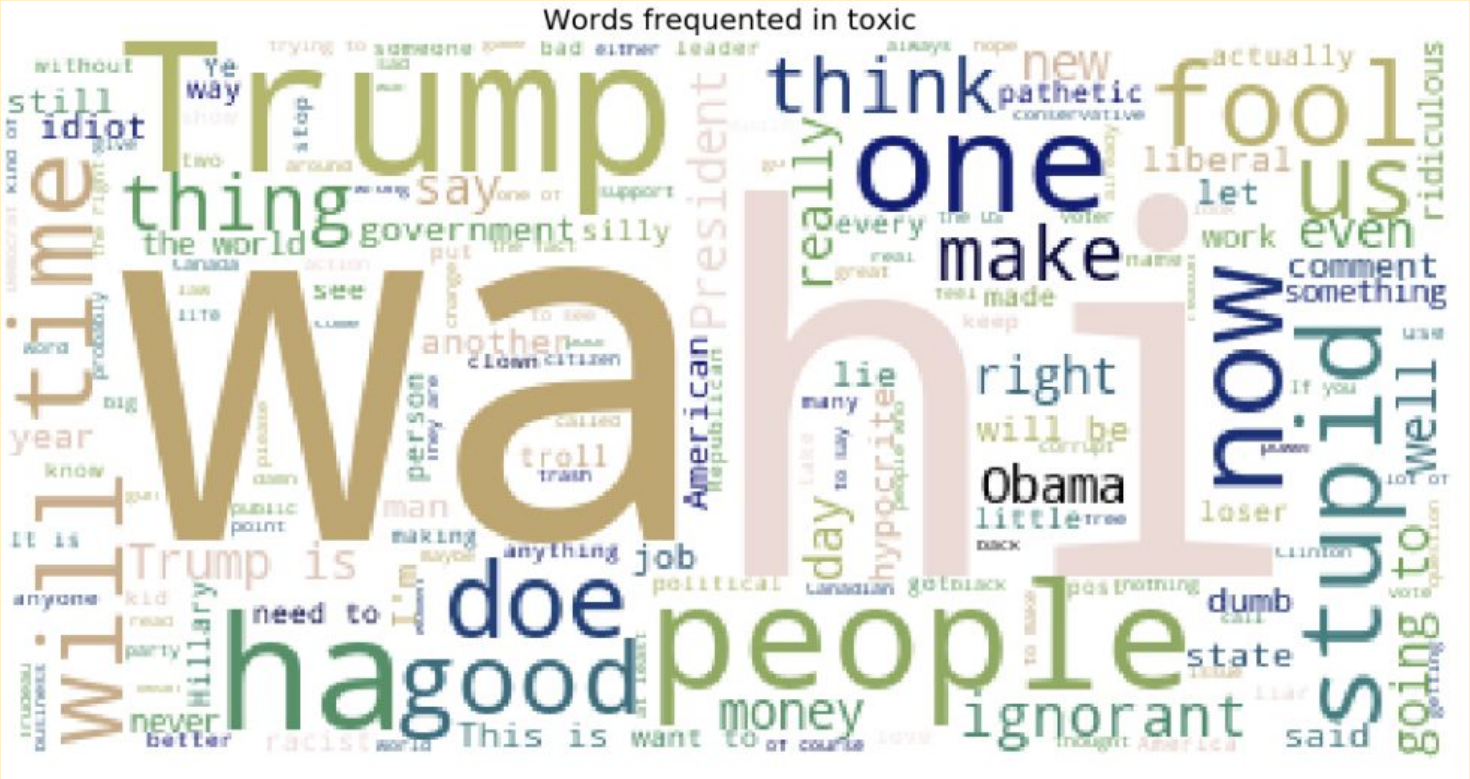


Length of comments by toxicity

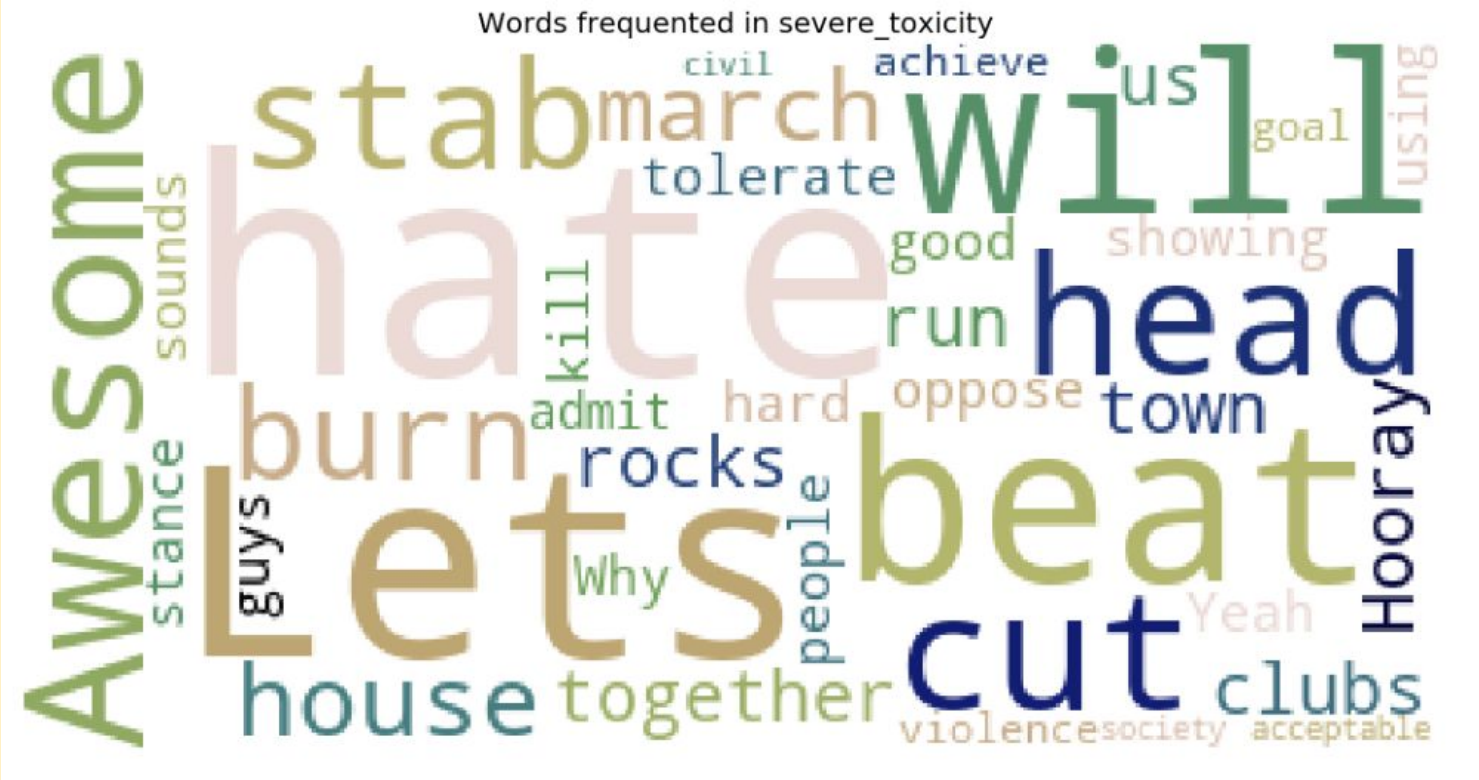
The Data


Number of comments per class
Number of Occurrences
200000
175000
150000
125000
100000
75000
50000
25000
0
195938.9690429427
8722.56947420781
26412.2728447346
42981.435129404344
154405.39958792922
toxic
severe_toxicity
obscene
identity_attack
insult
Type

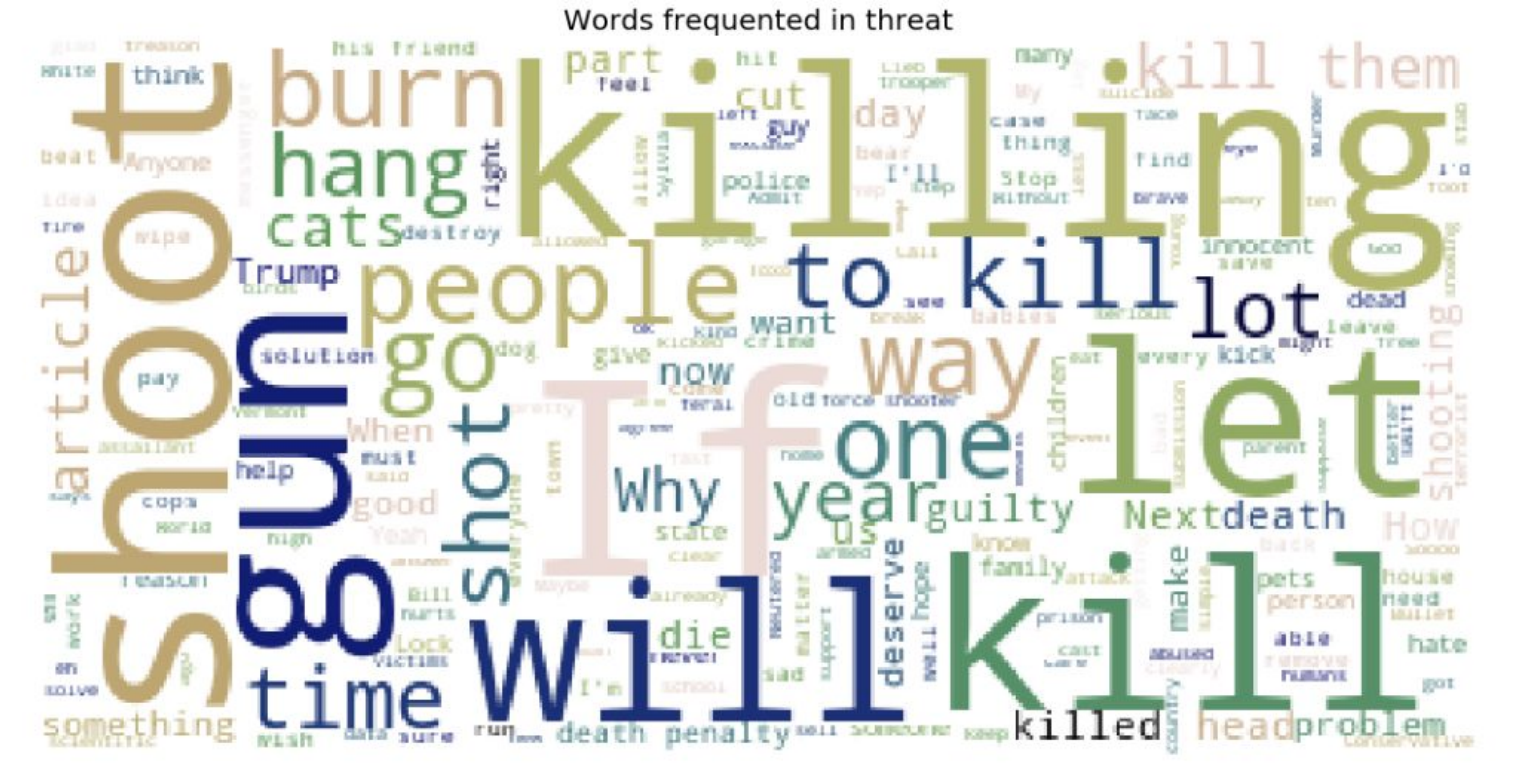
The Data


Words frequented in toxic
Trump
wa
hi
one
fool
think
people
doe
good
ha
stupid
now
us
time
will
thing
make
right
Obama
money
ignorant
President
American
Republican
Hillary
Trump is
need to
This is
want to
pathetic
liberal
government
the world
silly
say
really
lie
hypocrite
day
job
anything
troll
clown
citizen
person
man
year
idiot
still
without
way
actually
ridiculous
let
work
even
comment
something
well
going
dumb
state
said
loser
never
better
racist
anyone
making
political
canadian
conservative
new
every
made
will be
little
trash
back
gotback
nothing
trying to
someone
bad
either
leader
always
nope
stop
support
put
the fact
canada
action
see
another
public
point
kid
read
party
kind of
the right
democrat
republican
probably
big
word
know
call
vote
clinton
trump
america
thought
world
of course
love
truth
liar
tweet
tell
to make
post

The Data

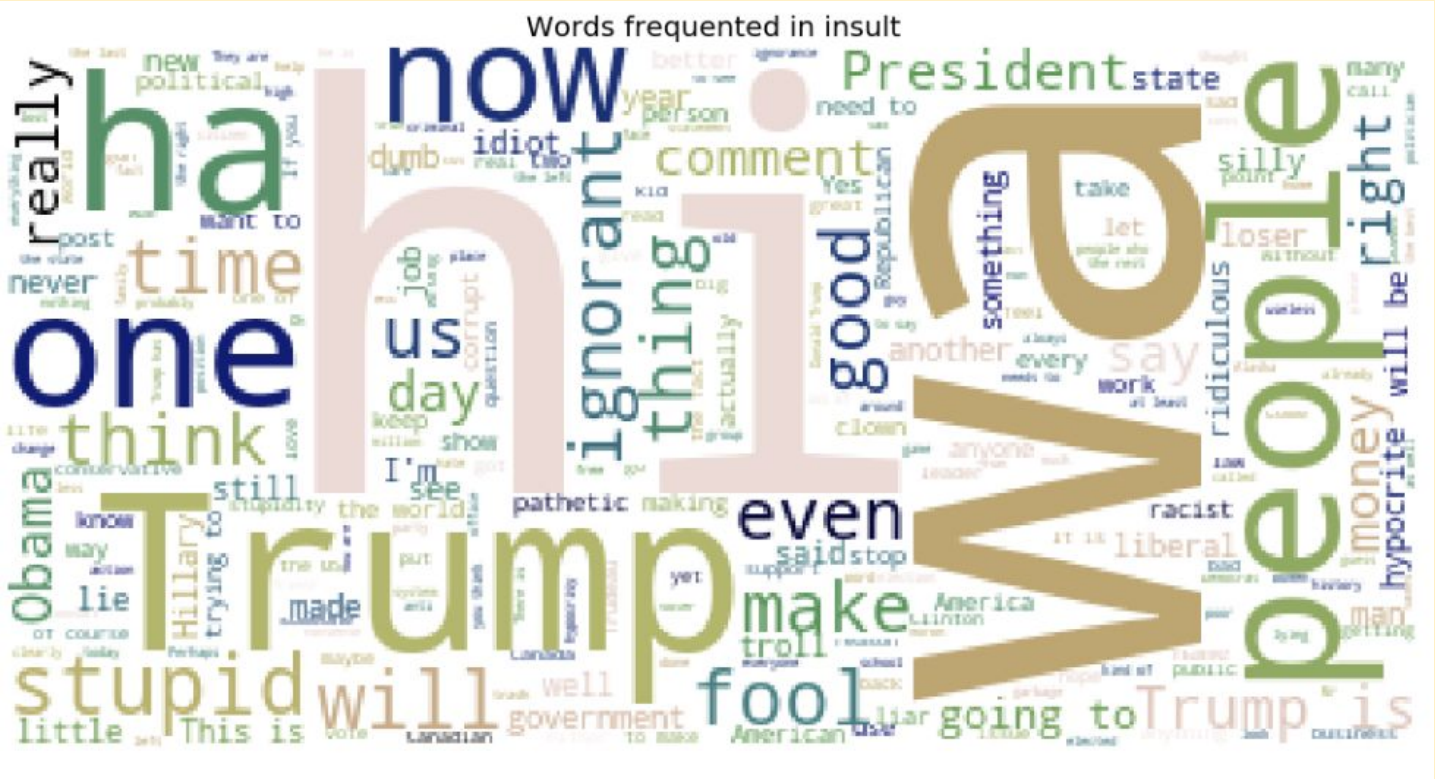

Words frequented in severe_toxicity
civil
achieve
us
stab
march
will
goal
using
tolerate
hate
good
showing
Awesome
sounds
kill
run
head
admit
hard
oppose
town
burn
rocks
beat
stance
guys
people
Hooray
Lets
Why
Yeah
cut
house
together
clubs
violence
society
acceptable

The Data

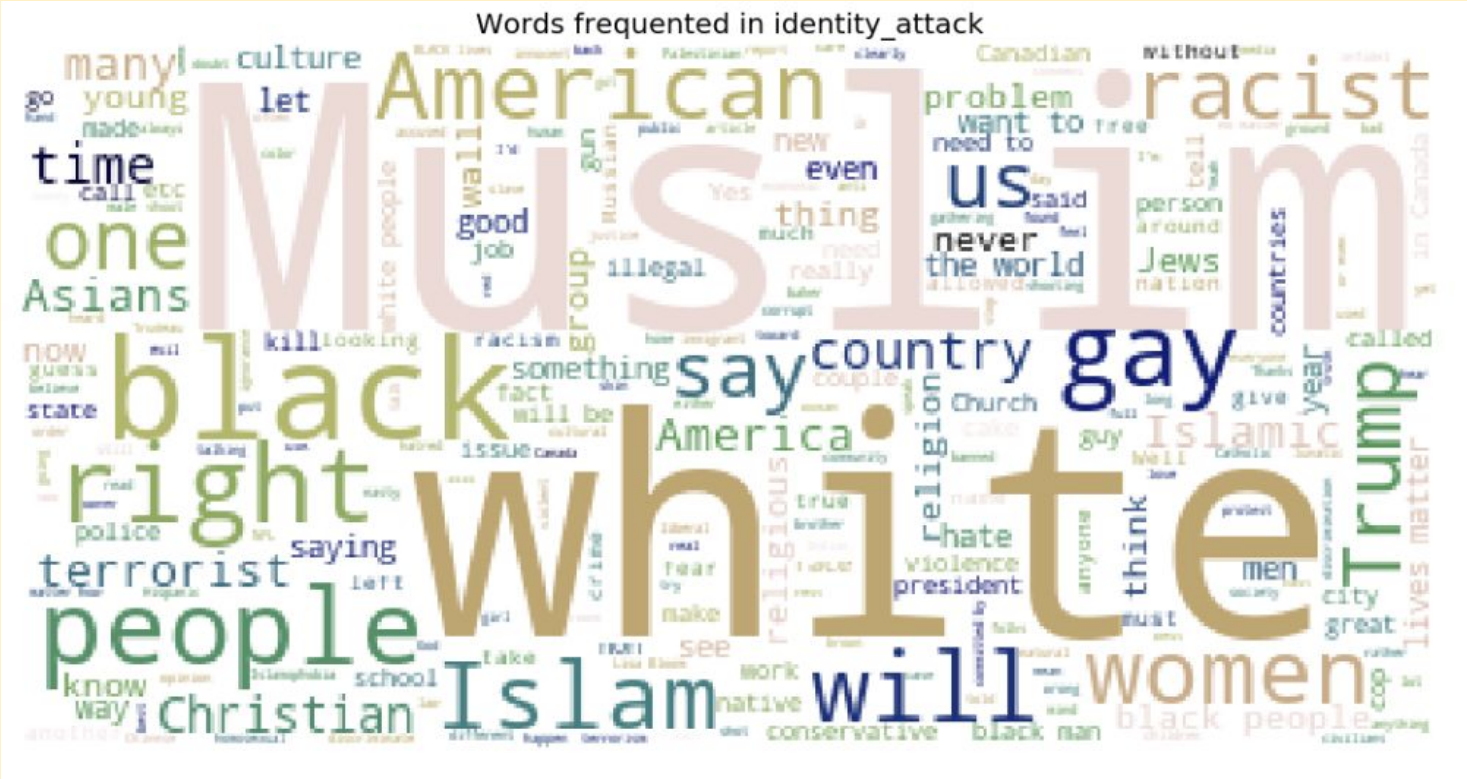


Words frequented in threat

The Data


Words frequented in insult
him
was
Trump
people
now
one
ha
President
stupid
will
fool
even
make
time
think
really
ignorant
thing
good
comment
us
day
Obama
Hillary
liar
going to
Trump is
government
little
This is
American
troll
America
said
stop
racist
liberal
money
hypocrite
right
will be
silly
loser
ridiculous
something
another
every
work
idiot
dumb
pathetic
still
lie
never
post
new
political
year
person
need to
state
Republican
clown
anyone
say
take
let
want to
job
corrupt
see
I'm
know
may
conservative
made
trying to
keep
show
the world
making
support
Canada
Canadian
well
man
public

The Data


Words frequented in identity_attack

THANKS!

CREDITS: This presentation template was created by Slidesgo, including icons by Flaticon, and infographics & images by Freepik.