



CYBERBULLYING DETECTOR

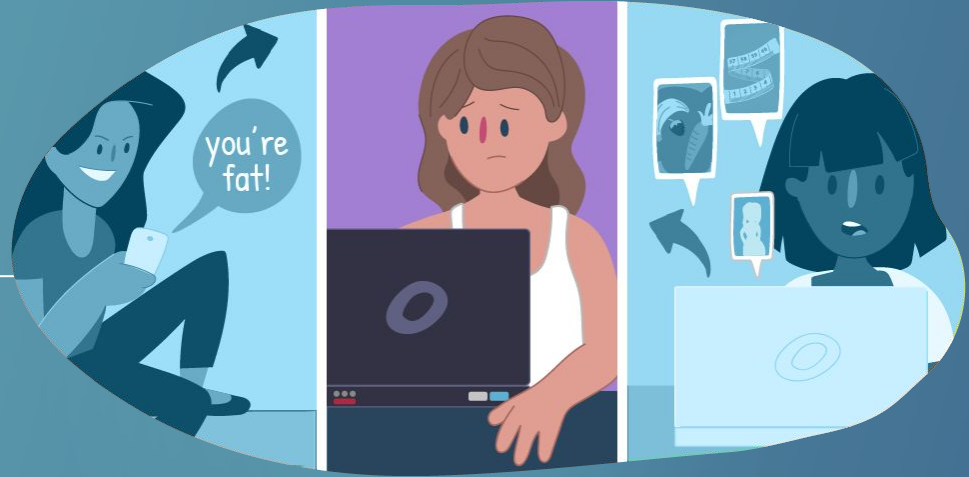
<ITC>

Areej Aweida
May Steinfeld-Kalisher
Sheryl Sitruk

What is Cyberbullying?

Cyberbullying


Sending, posting or sharing negative, harmful, false or mean content about someone else.



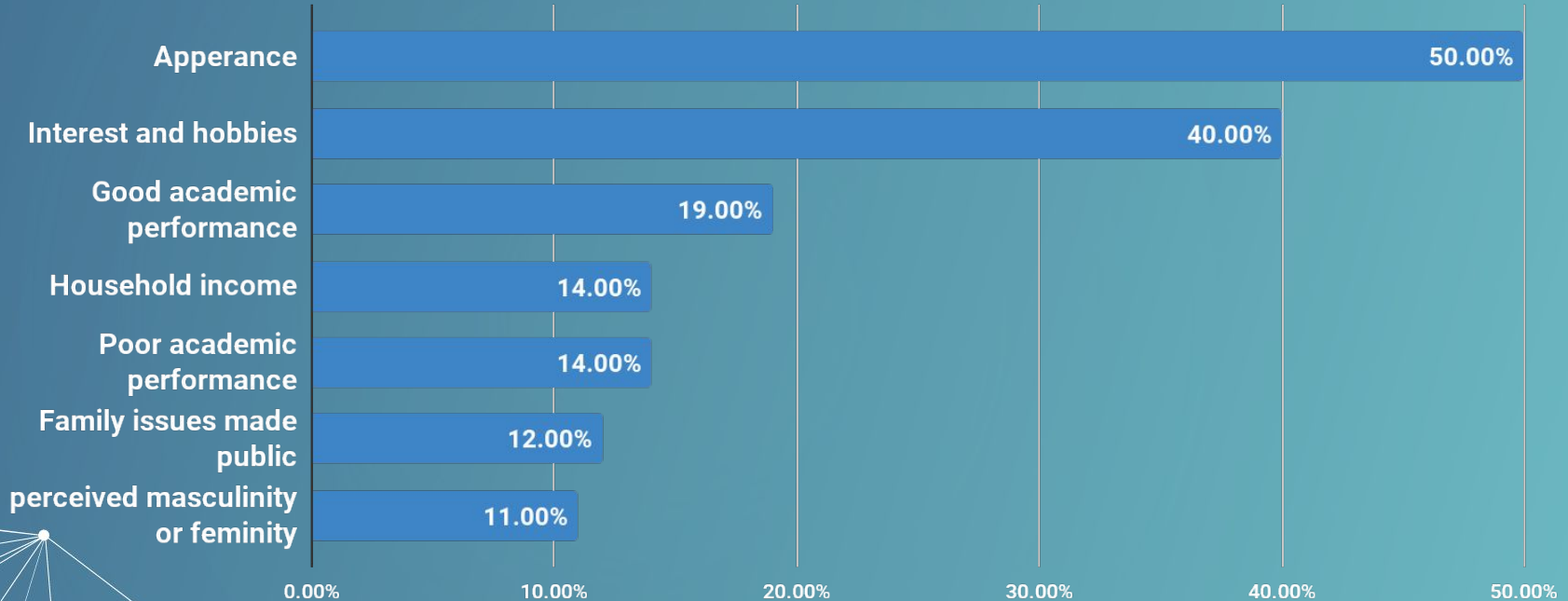


Social issue

73% of students feel they have been
bullied in their lifetime and 44% say it has
happened in the last 30 days



Top reasons for Cyberbullying among young people




Negative effects of Cyberbullying on victims





Business Cyberbullying

Can cause lost revenue, a decrease in employee morale and a downtick in a company's persona and prestige



Starbucks Case



black power alt bro

@vidalwuu



y'all realize there are no coloured hands in the press photos right
[@Starbucks](#) [#RaceTogether](#)

7:21 AM - 17 Mar 2015

1,889 RETWEETS 1,422 FAVORITES



The goal

Predict whether a given post
is cyberbullying

Pipeline



Business Problem

Define our business problem



EDA & Preprocessing

Cleaning and Preprocessing the data



Modeling

Try different ML models



Analyze results

Evaluate models looking on accuracy and recall metrics.



NN models

Go deeper and run deep learning models.



Deployment

Deploy our app using Flask

Our Challenges

**Access to high
quality
datasets**

**Unbalanced
dataset**

**Complexity of
tweets**



Datasets



Hate-Speech

9924 posts classified as hate speech or not.



Labeled-Tweets

11091 tweets classified as offensive and non-offensive



Public Data-Davinson

24,783 tweets classified as hate_speech, offensive or neither

Some Numbers

45600

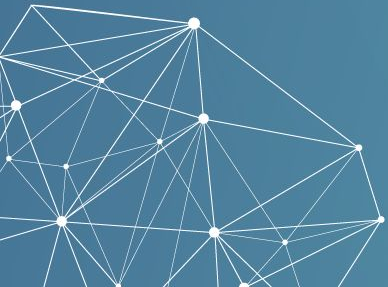
Samples

20944

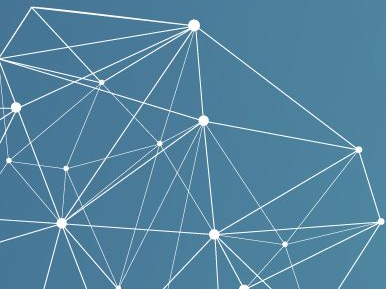
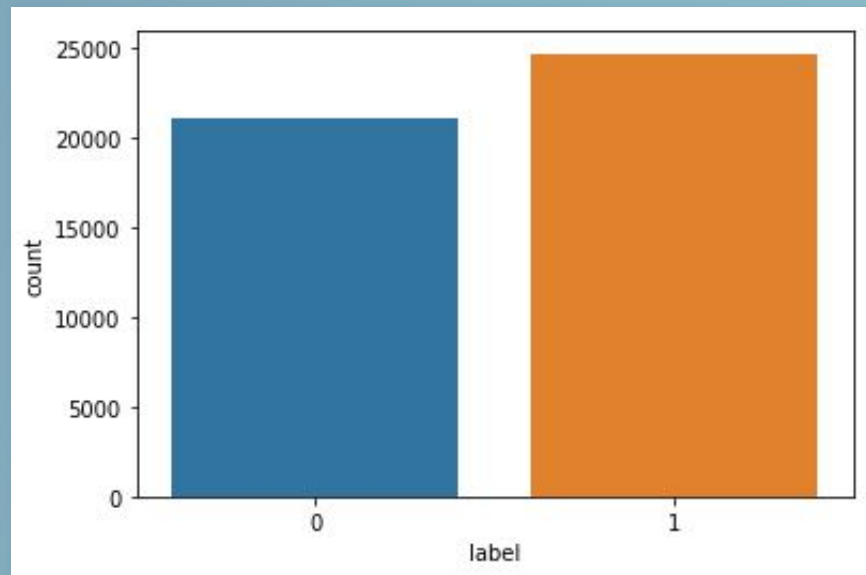
Non-cyberbullying

24656

Cyberbullying



Our Dataset

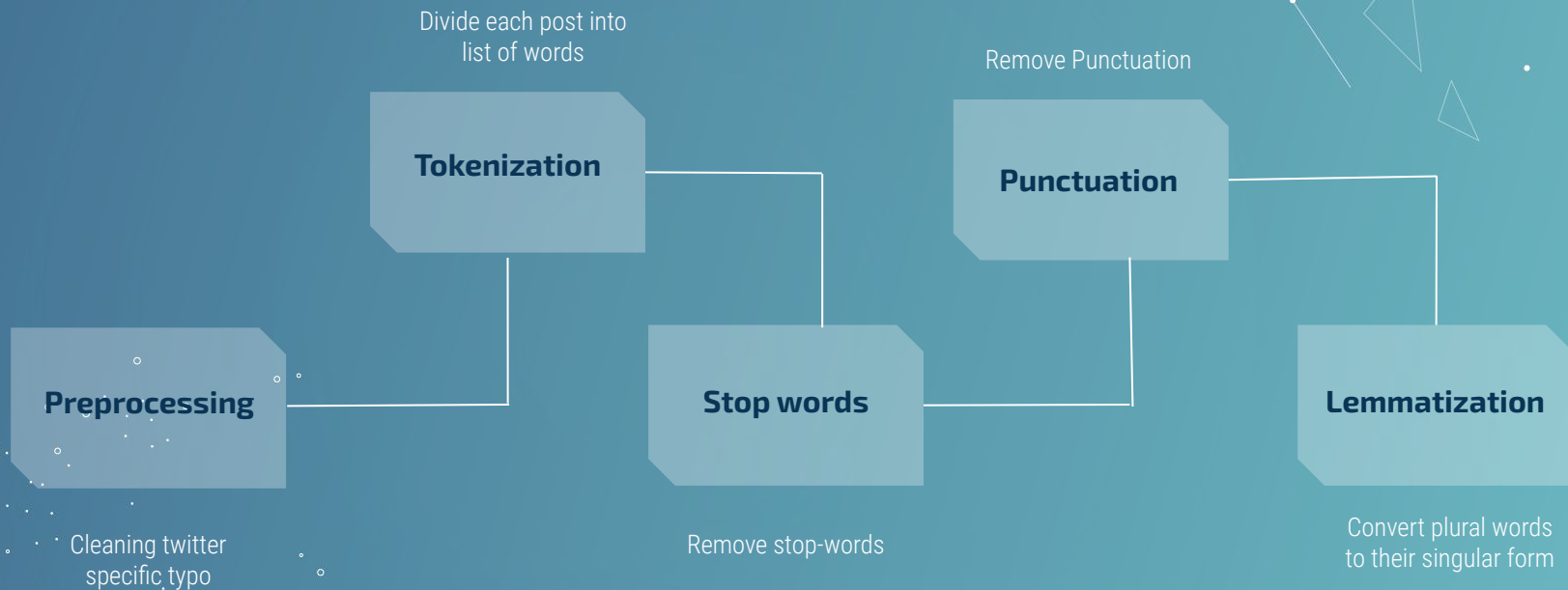


Tweets Example

"@Marlon_N_Gaines I am brown, while you are **black and ignorant**, your **stupidity** makes me sick, typical thug looking **nigger**."

"Bumped into a **bitch** that filed for sexual harassment at work against a co-worker...I aint been so shook in my life"

Preprocessing

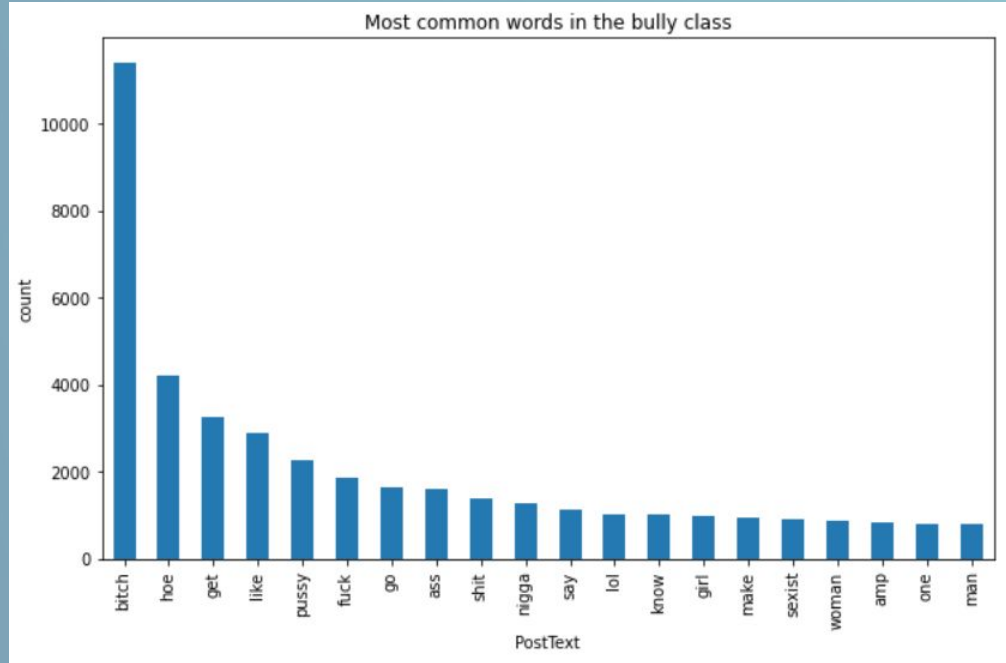


Data Visualization

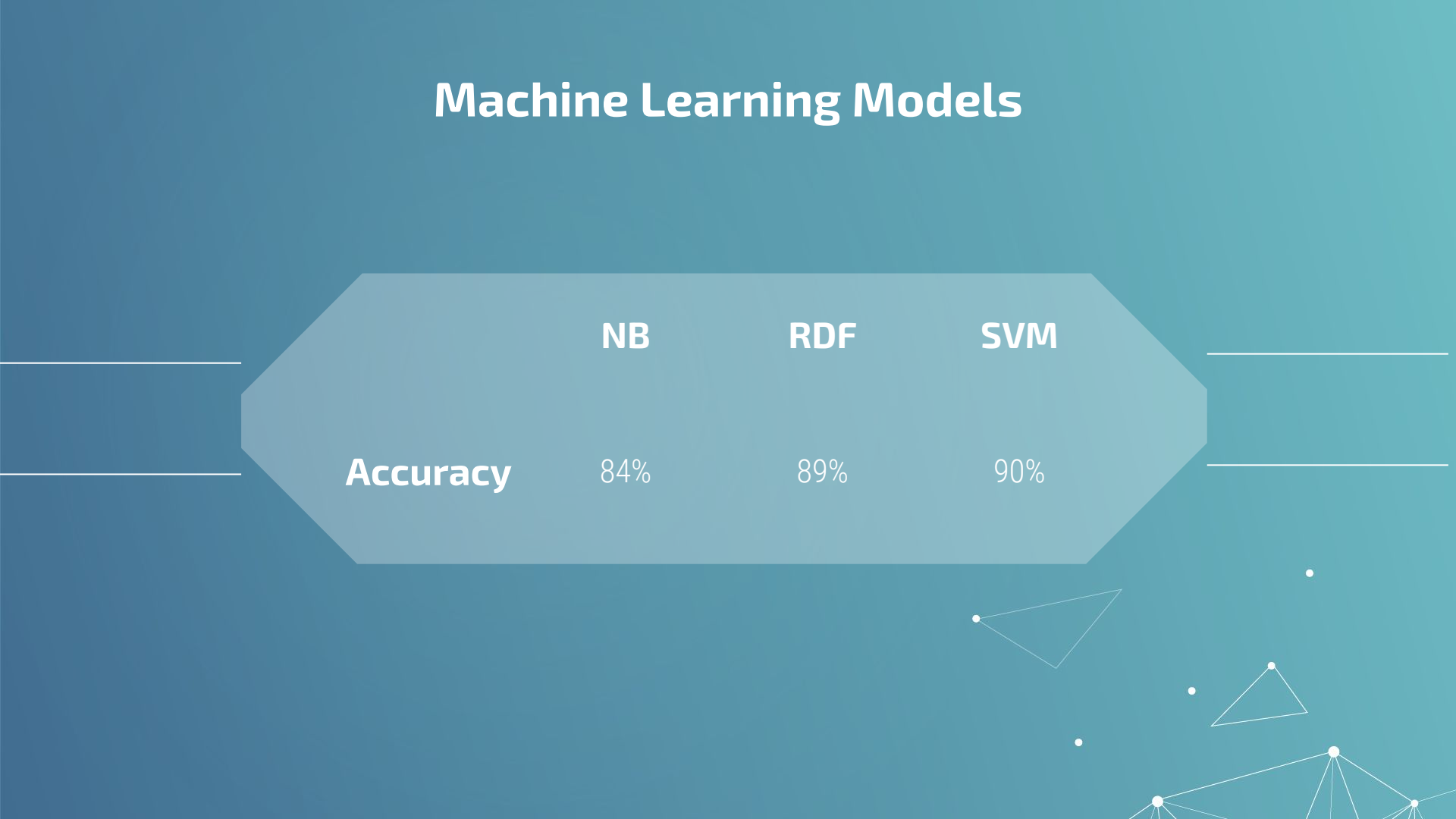


Words in Cyberbullying Tweets

Most common words in the bully class



Machine Learning Models



	NB	RDF	SVM
Accuracy	84%	89%	90%

Deep Learning Model

BERT

State of the art language model
for NLP

ACCURACY

93%



Comparing results

ML vs. NN

NN

Good accuracy



Heavy



Time



Basic ML models

Good accuracy



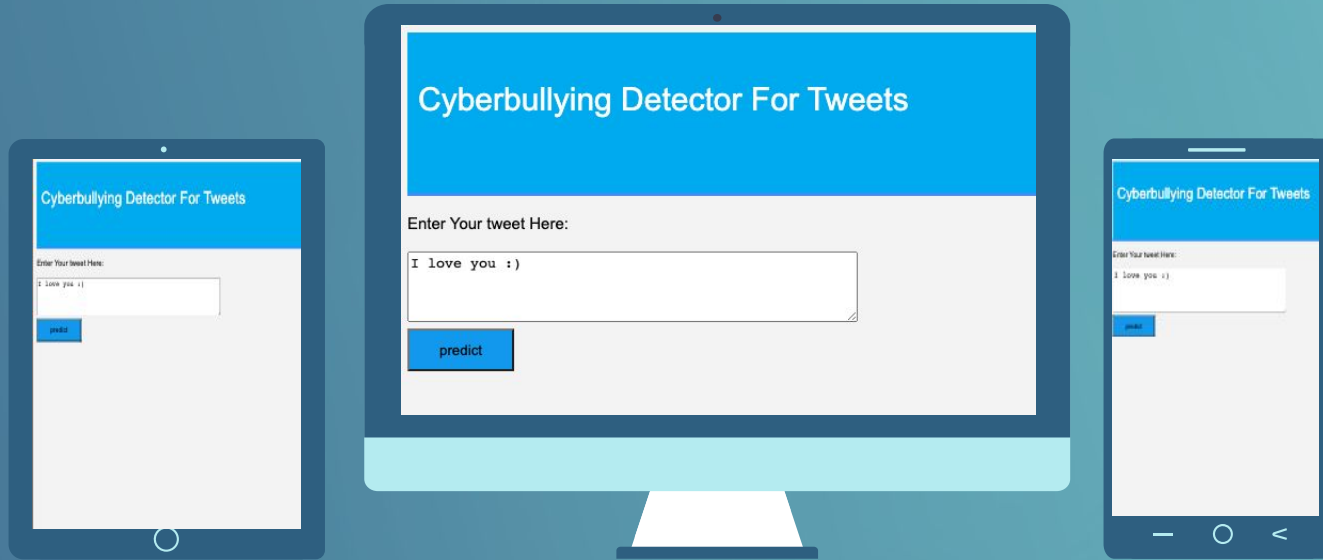
Light



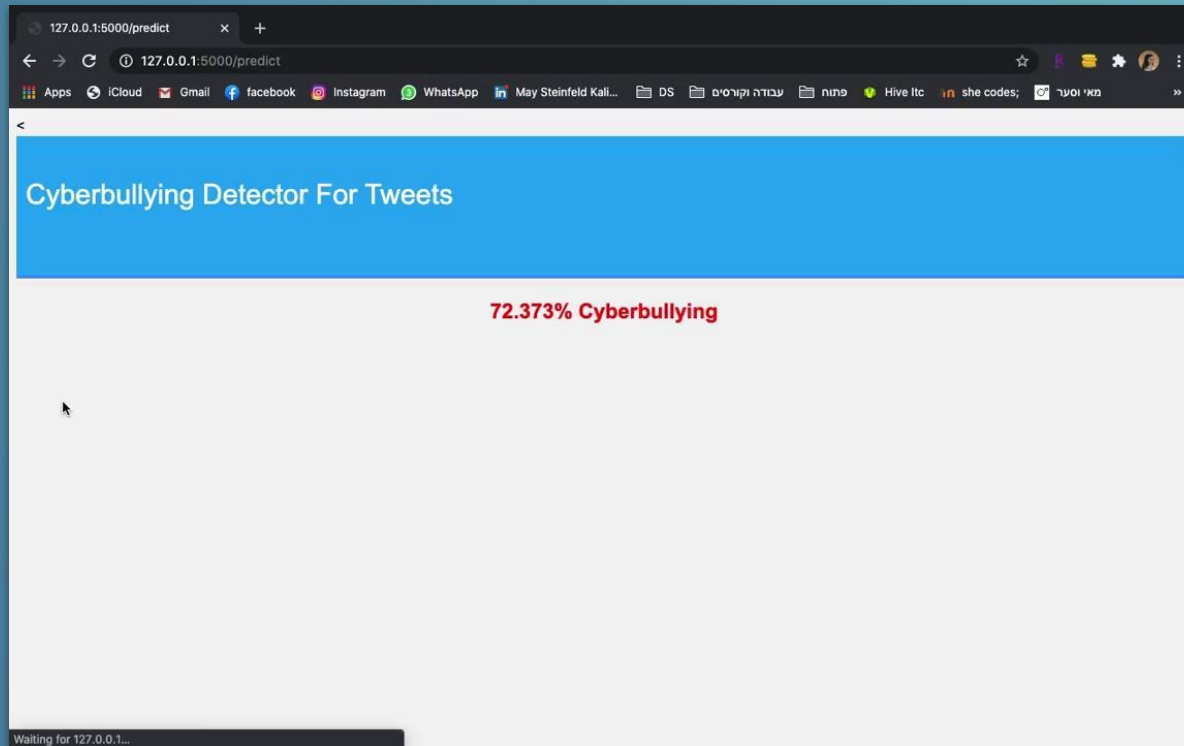
Preprocessing



Deployment



Demo



What's Next ?



More Scalability

Works only for long tweets
(around 6-9 words)...



Type of Cyberbullying

Detect specifics types of
cyberbullying : Sexism, Racism ...



Deploy for other platforms

42% of cyberbullying are detected
on Instagram and 37% on
Facebook

The leadership team



Sheryl Sitruk

Founder



Areej Aweida

Founder



May Steinfeld-Kalisher

Founder

RESOURCES

SOURCES

- <https://comparecamp.com/cyberbullying-statistics/>
- [Bert article](#)
- [GitHub](#)





THANKS

Does anyone have any questions?



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

Please keep this slide for attribution.