



Cyber bullying

**Final project in Hebrew natural language
processing course**

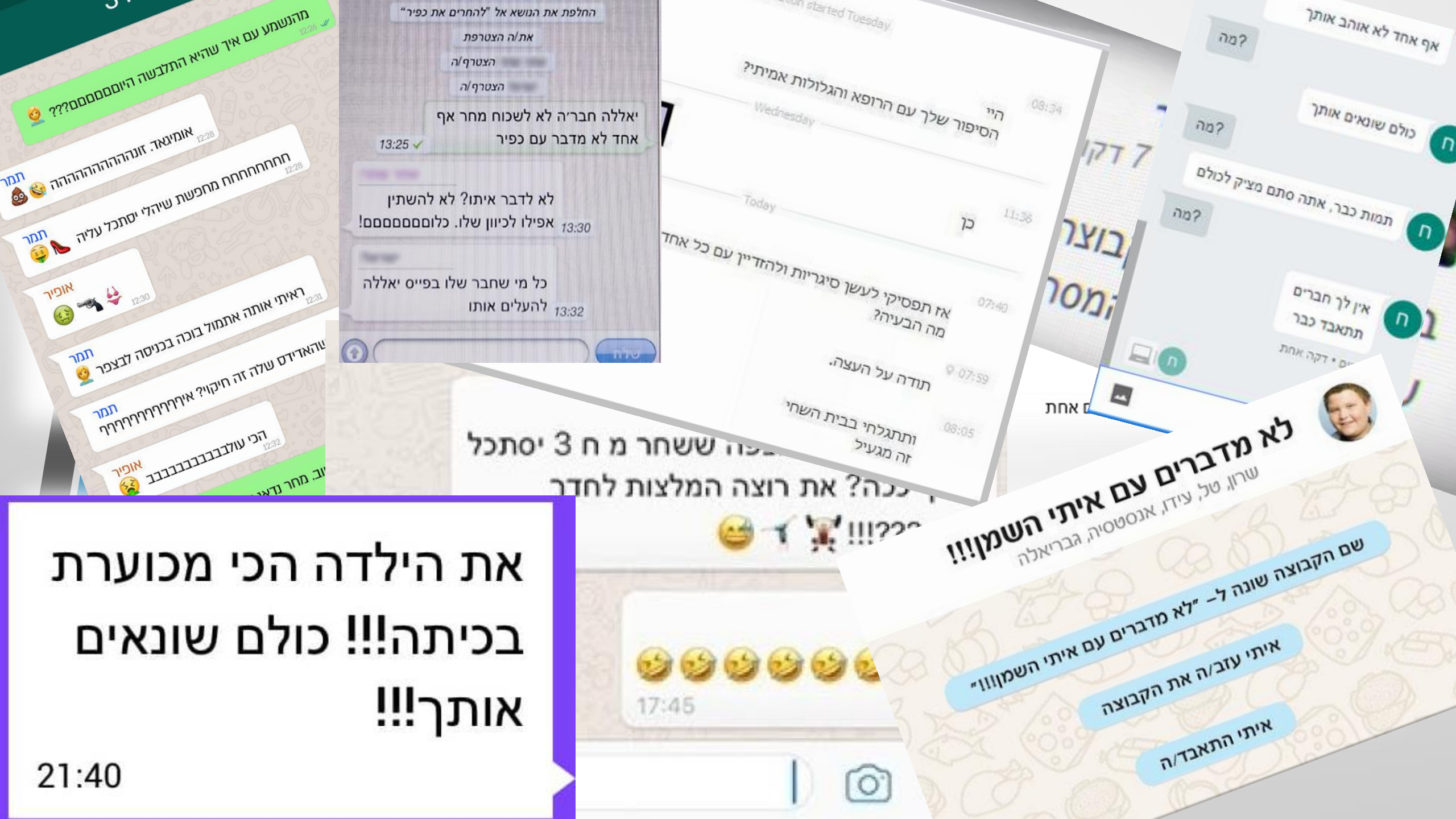


Background

- The internet can be a productive and empowering place for children
- There are many dangers on the Internet that can affect the development of children
- So what is Cyberbullying?
 - When someone bullies or harasses others on the internet
 - Posting rumors, threats, sexual remarks, a victims' personal information, or pejorative labels
 - Common among teenagers

את הילדה הכי מכווערת
בכיתה!!! כולם שונאים
אותך!!!

21:40





The Domain and The Problem

- The domain- Cyberbullying
- The problem- Detection of Cyberbullying in social networks

Research Questions

- Can social network messages be classified as violent or not?
- What are types of cyberbullying?



The Data



What is the data?

- Messages of children and teenagers from the social media
- Tagged data



How we got the data?

- Keepers



How many?

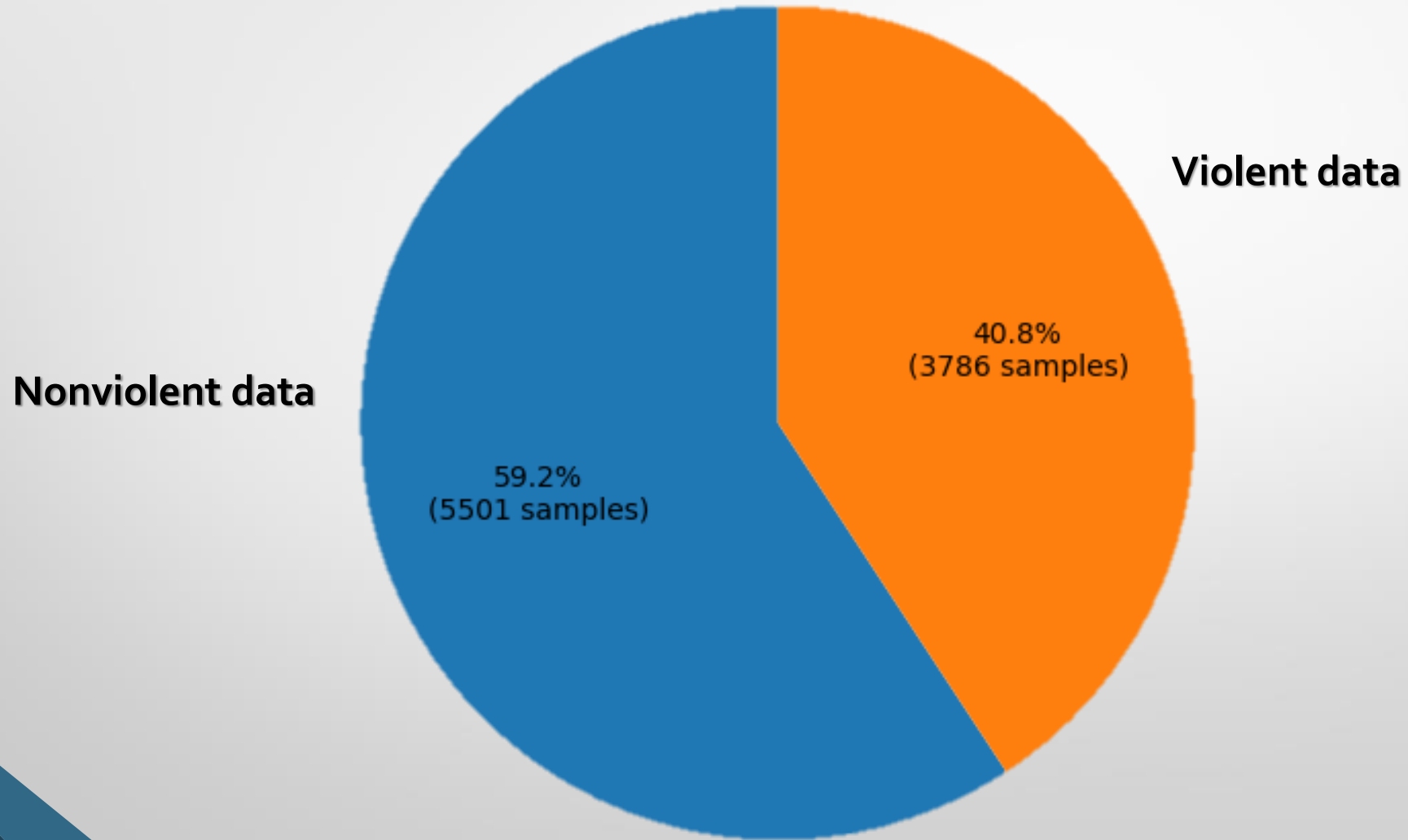
- 5501 nonviolent message
- 3787 violent sentences



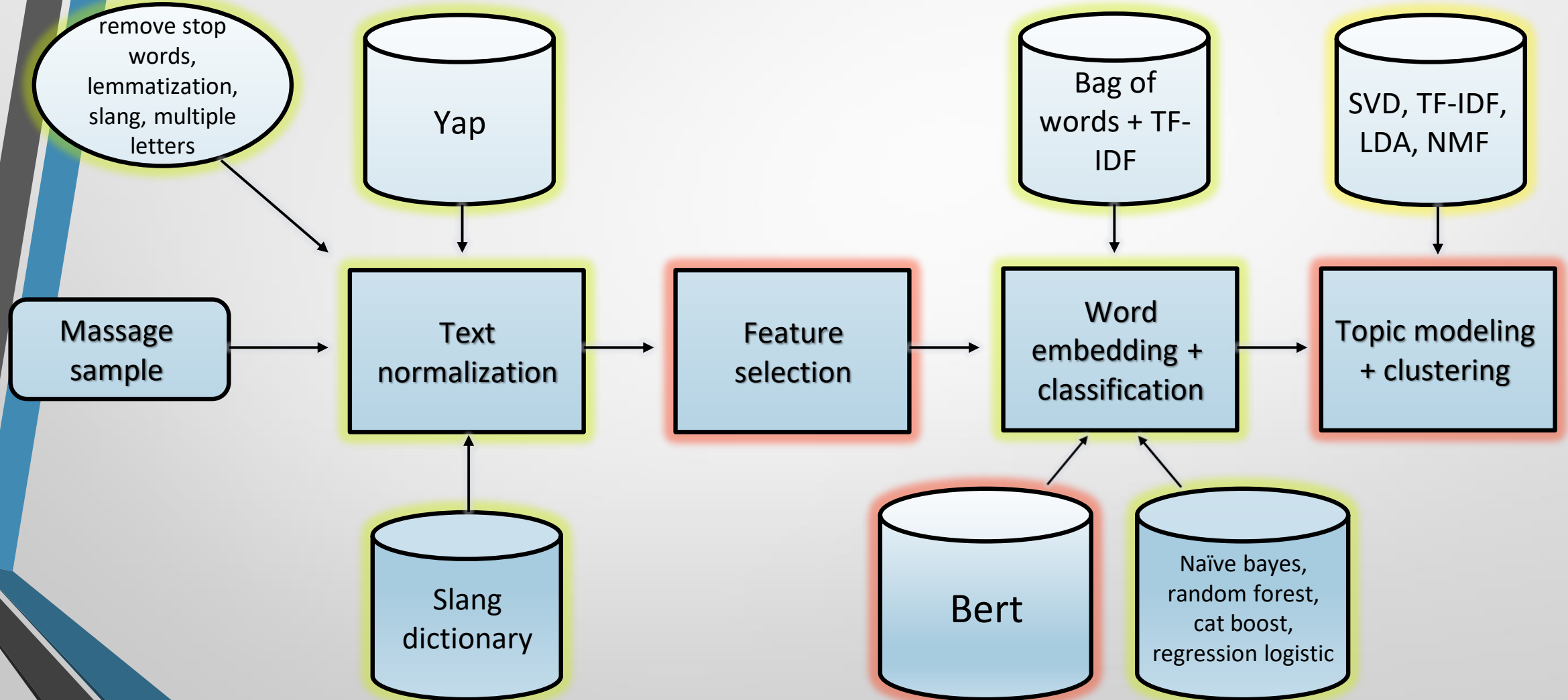
Special characterizes

- Slang, spelling mistake, duplicate characters
- Short messages

Data - Pie charts



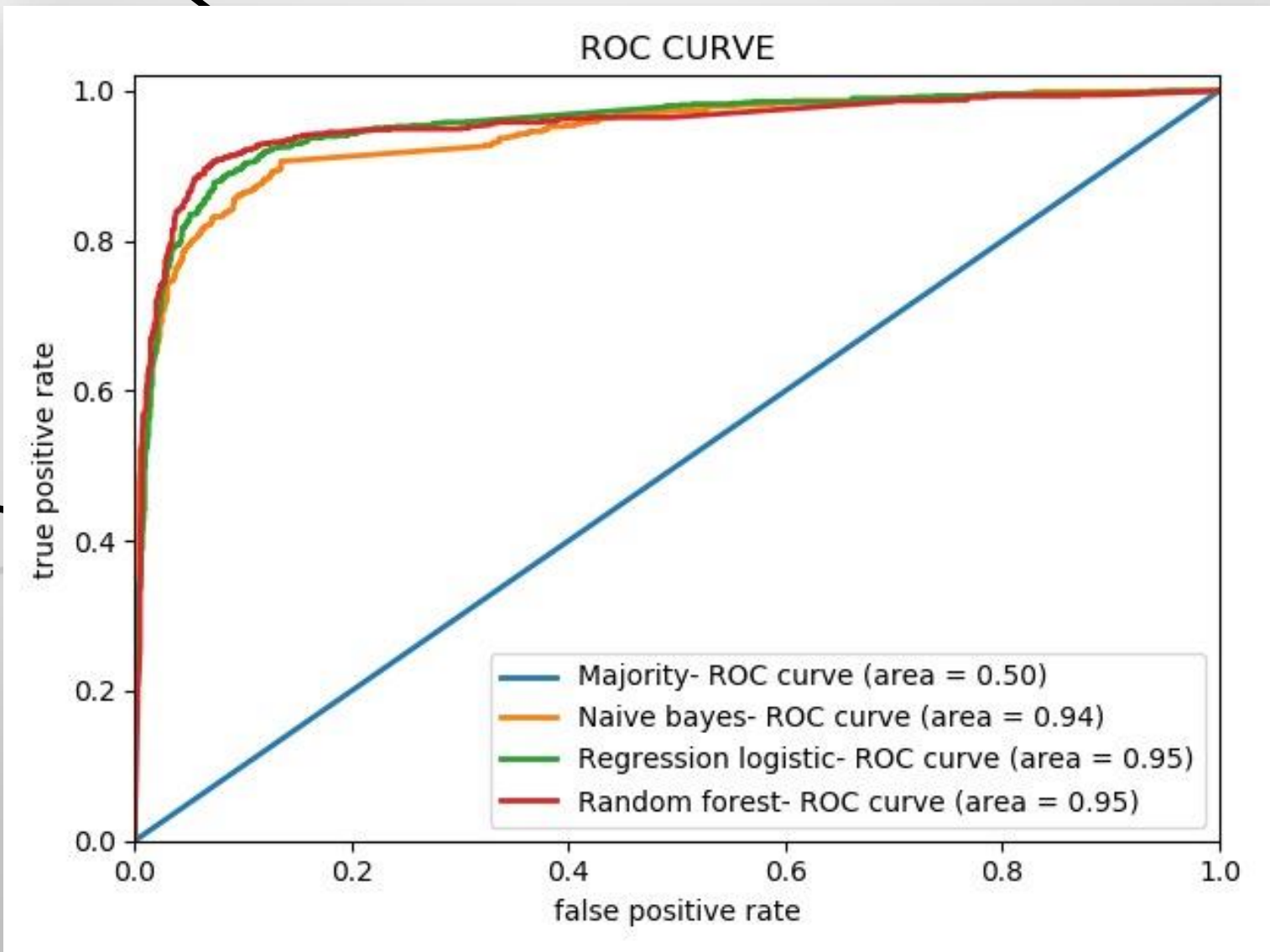
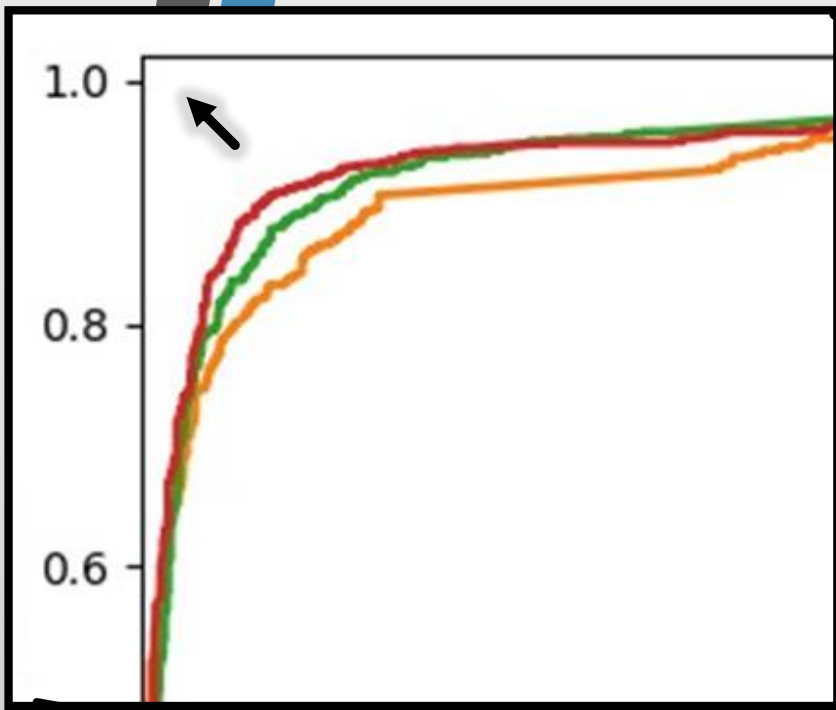
The Pipeline



First Results

	Recall score	Precision score	Accuracy score	F1 score	F2 score
majority	0.5	0.3049	0.6098	0.3788	0.4433
naive bayes	0.8765	0.88	0.8846	0.8782	0.8772
regression logistic	0.8632	0.9018	0.8867	0.8758	0.8668
random forest	0.9072	0.9163	0.9165	0.9113	0.9087

ROC Curve



The Types of Cyberbullying

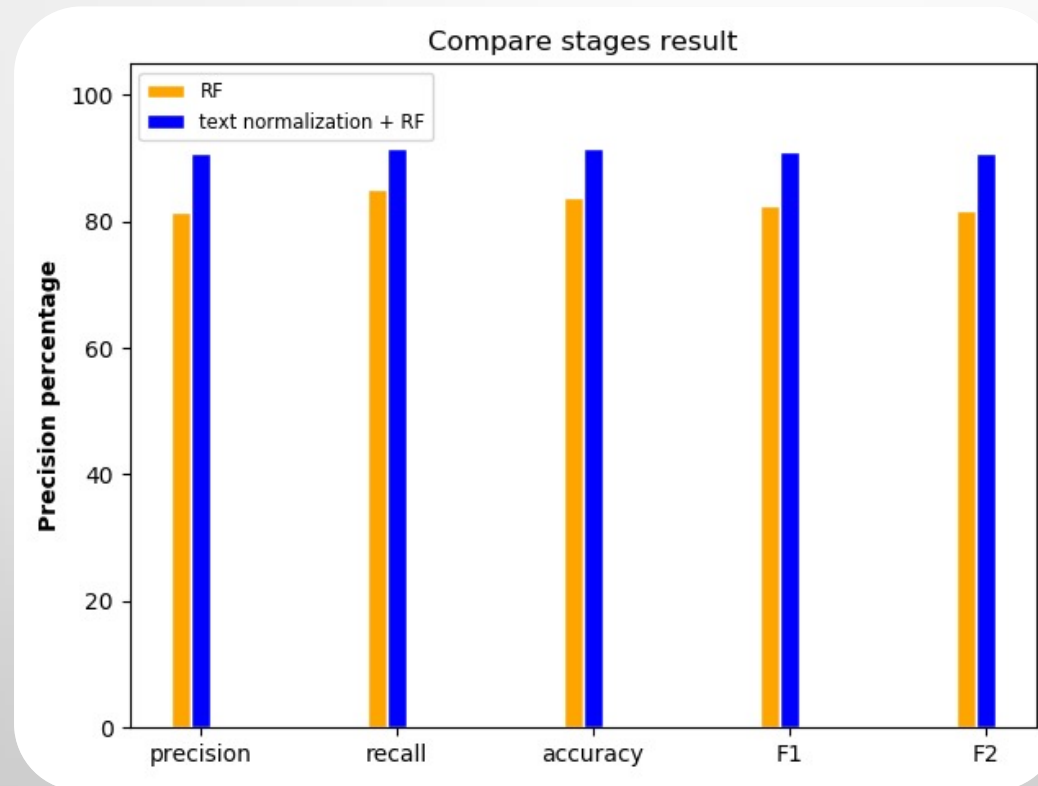
Sexual Comments

Threats

Pejorative

What went as expected

- Text normalization phase improved classifier results
- The better and more complex classifiers provided better results



Surprises and Insights so far

- **Data** – Hebrew slang, duplicate characters, spelling mistake

Examples – "חיימשלי", "טוּוּוּוּב", "אמשך", "מצתערת"

- **Yap** – incorrect lemmatization for strong word

Examples – "מכות < מך"

- **Clustering**
- **Topic Modeling**

Summery

- We succeeded in classifying messages into violence and non-violence.
- We partially found an answer to the types of cyberbullying.
- We need to continue to improve our classifiers and clusters. First, by adding a feature selection step.

Further work

- Violence score
- Trained the models with more various data
- Classified message by violence type
- Auto-fix violent messages to non-violent messages

