Cyber bullying

Final project in Hebrew natural language processing course


18+

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

מהנשמע עם איך שהיא התלבשה היוסססססם??? 12:26

אומיגאד. זונהההההההההה 12:28

חחחחחחחחח מחפשת שיהלי יסתכל עליה 12:28

תמר

אופיר 12:30

ראיתי אותה אתמול בוכה בכניסה לבצפר 12:31

תמר

שהאדידס שלה זה חיקוי? איריריריריריריריריריר

תמר

הכי עולבבבבבבבבבב 12:32

אופיר

החלפת את הנושא אל "להחרים את כפיר"

את/ה הצטרפת

הצטרף/ה

הצטרף/ה

יאללה חבר'ה לא לשכוח מחר אף אחד לא מדבר עם כפיר 13:25

לא לדבר איתו? לא להשתין אפילו לכיוון שלו. כלוסססססם! 13:30

כל מי שחבר שלו בפייס יאללה להעלים אותו 13:32

שלח

Wednesday

היי הסיפור שלך עם הרופא והגלולות אמיתי? 08:24

Today

כן 11:36

אז תפסיקי לעשן סיגריות ולהחדיין עם כל אחד מה הבעיה? 07:40

תודה על העצה. 07:59

ותתגלחי בבית השחי זה מגעיל 08:05

אף אחד לא אוהב אותך

מה?

כולם שונאים אותך

מה?

תמות כבר, אתה סתם מציק לכולם

מה?

אין לך חברים תתאבד כבר

לא מדברים עם איתי השמן!!!

שרון, טל, עידו, אנסטסיה, גבריאלה

שם הקבוצה שונה ל- "לא מדברים עם איתי השמן!!!"

איתי עזב/ה את הקבוצה

איתי התאבד/ה

17:45

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

21:40

The Domain and The Problem

- The domain- cyber bulling
- 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?

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

How we got the data?

- Keepers

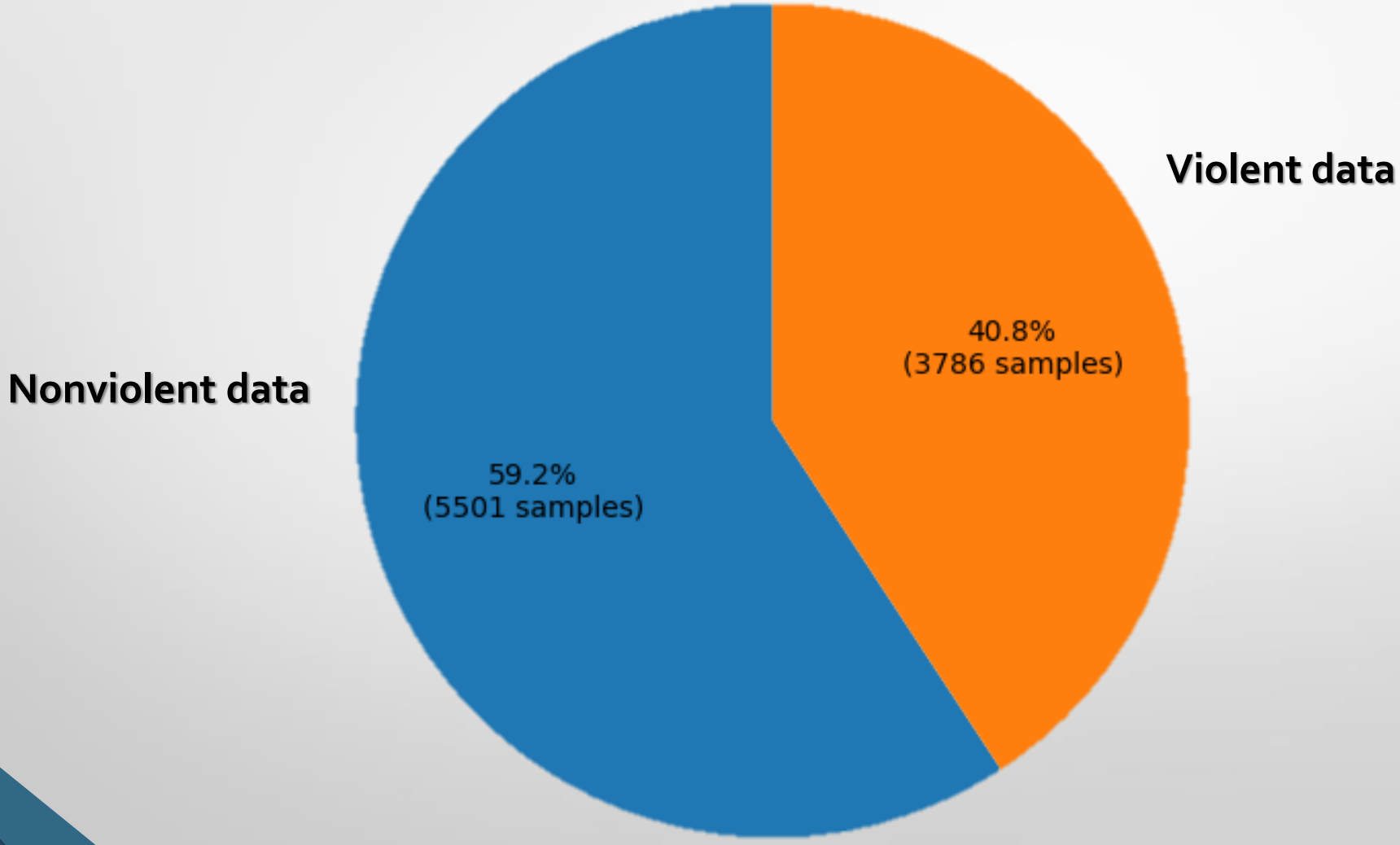
How many?

- 5501 nonviolent massage
- 3787 violent sentences

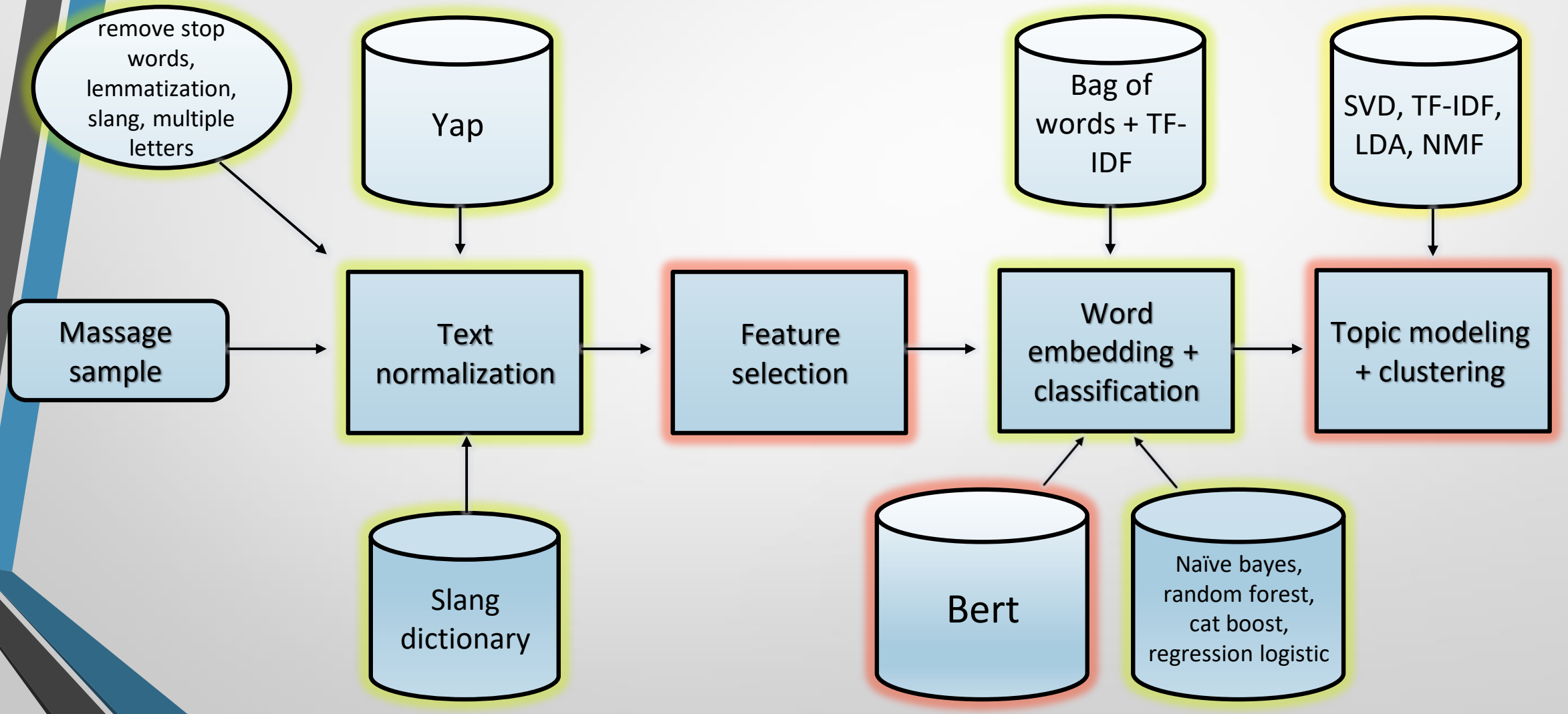
Special characterizes

- Slang, spelling mistake, duplicate characters
- Short massages

Data - Pie charts



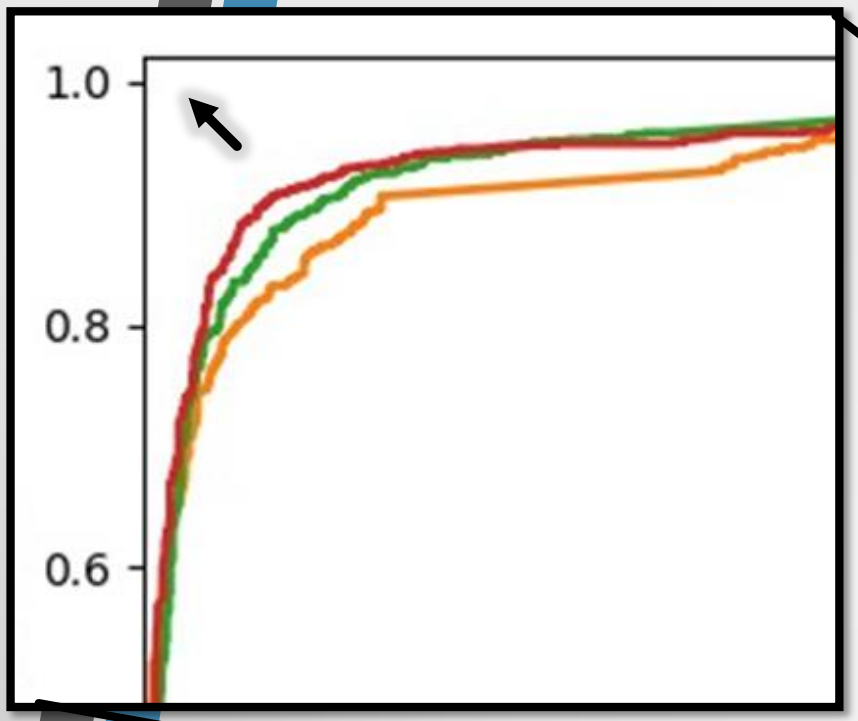
The Pipeline

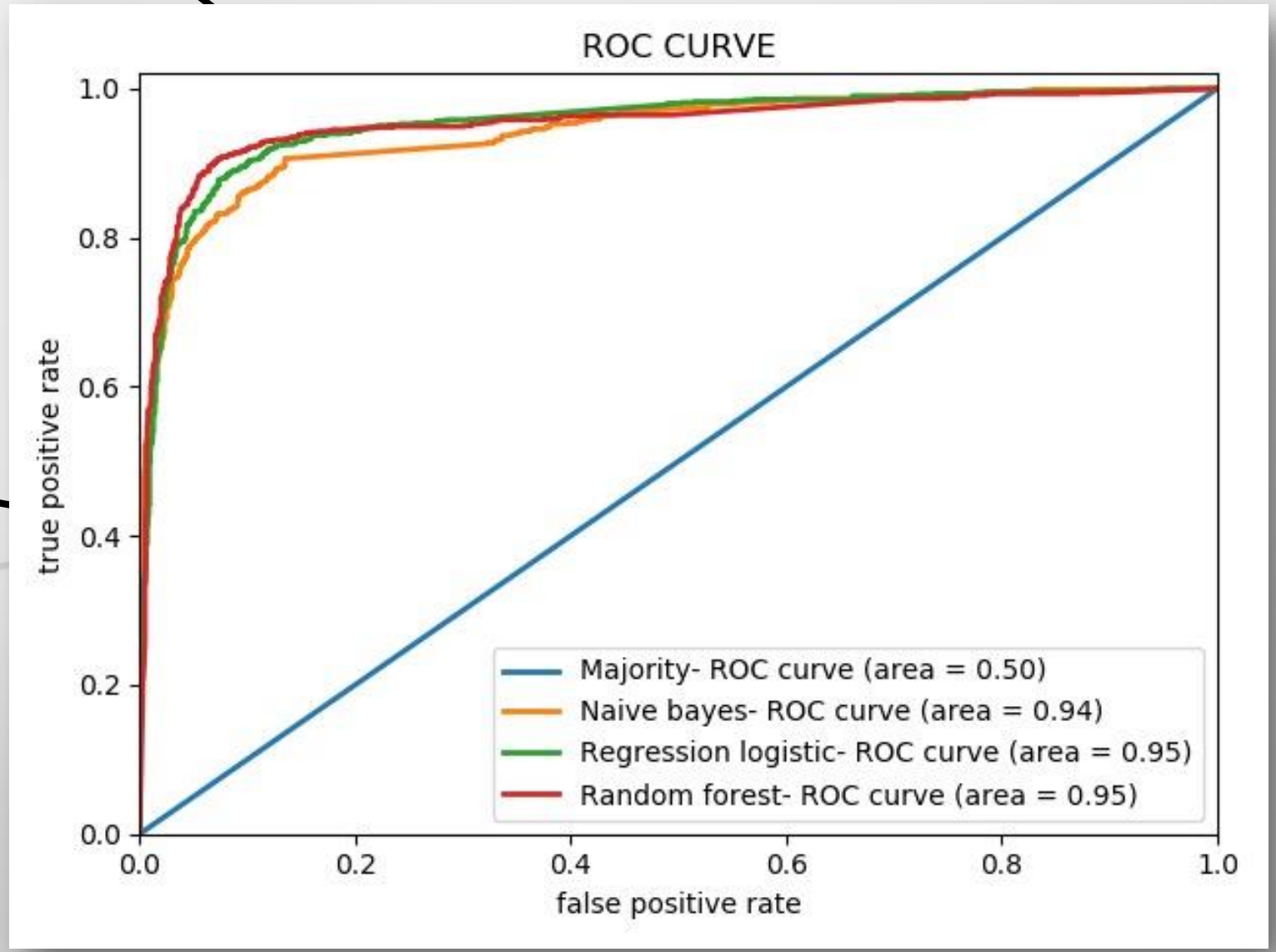


First Results

	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


1.0
0.8
0.6


ROC CURVE
true positive rate
false positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Majority- ROC curve (area = 0.50)
Naive bayes- ROC curve (area = 0.94)
Regression logistic- ROC curve (area = 0.95)
Random forest- ROC curve (area = 0.95)

The Types of Cyberbullying

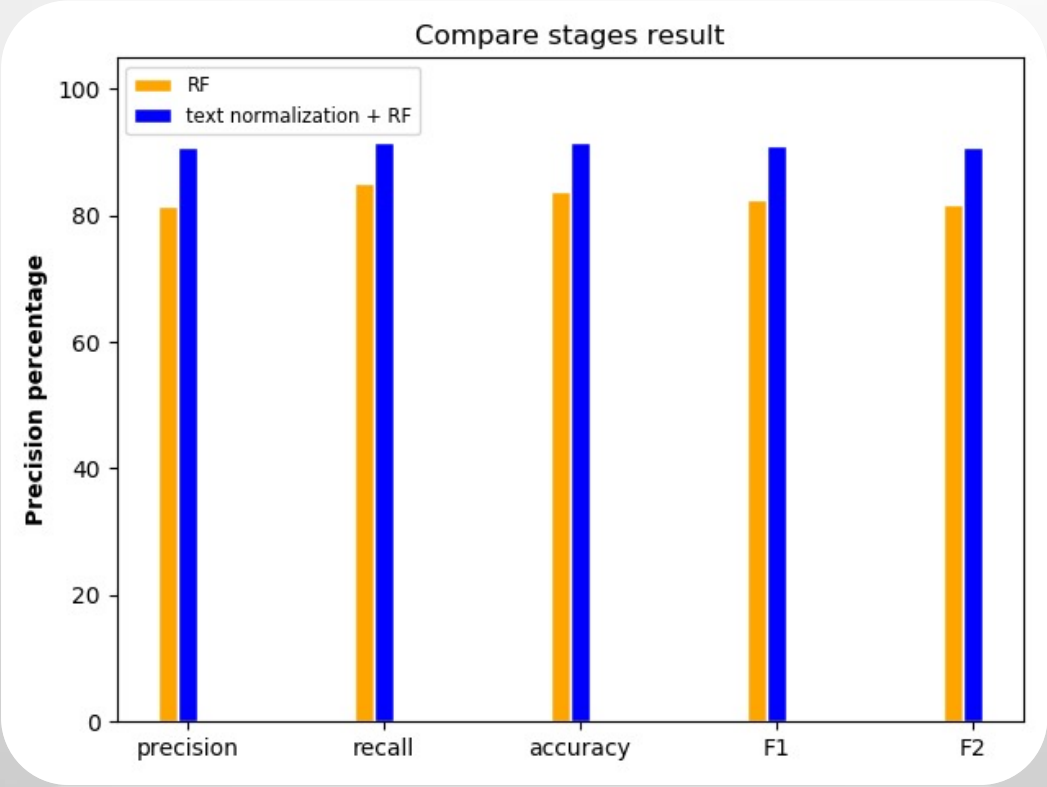
Sexual Comments

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 – "מצתערת", "אמשך", "טוווווב", "חיימשלי"

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

Examples – "מכות -> מך"

- **Clustering**

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