Analyzing People’s Reactions On Different Levels Of Anonymity

ECE 180 Project by Team 1

Index

- **Motivation**
  - Our goal
- **Dataset**
  - Reddit vs Facebook
  - Topics considered
- **Processing & Modeling**
  - Data Extraction
  - Model
- **Analysis**
  - General analysis
  - Detailed analysis with examples
- **Future Possibilities**

Motivation

In the virtual world, people tend to communicate without considering the courtesy and morality of the situation. For example, publishing **toxic comments** on social media brings hatred to the situation and may hurt other people.

Our goal is to **analyze the difference in aggression** people show on **different levels of anonymity** by comparing the toxicity of comments on Facebook and Reddit on a certain topic.

Dataset

To measure the level of aggression people show on different levels of anonymity

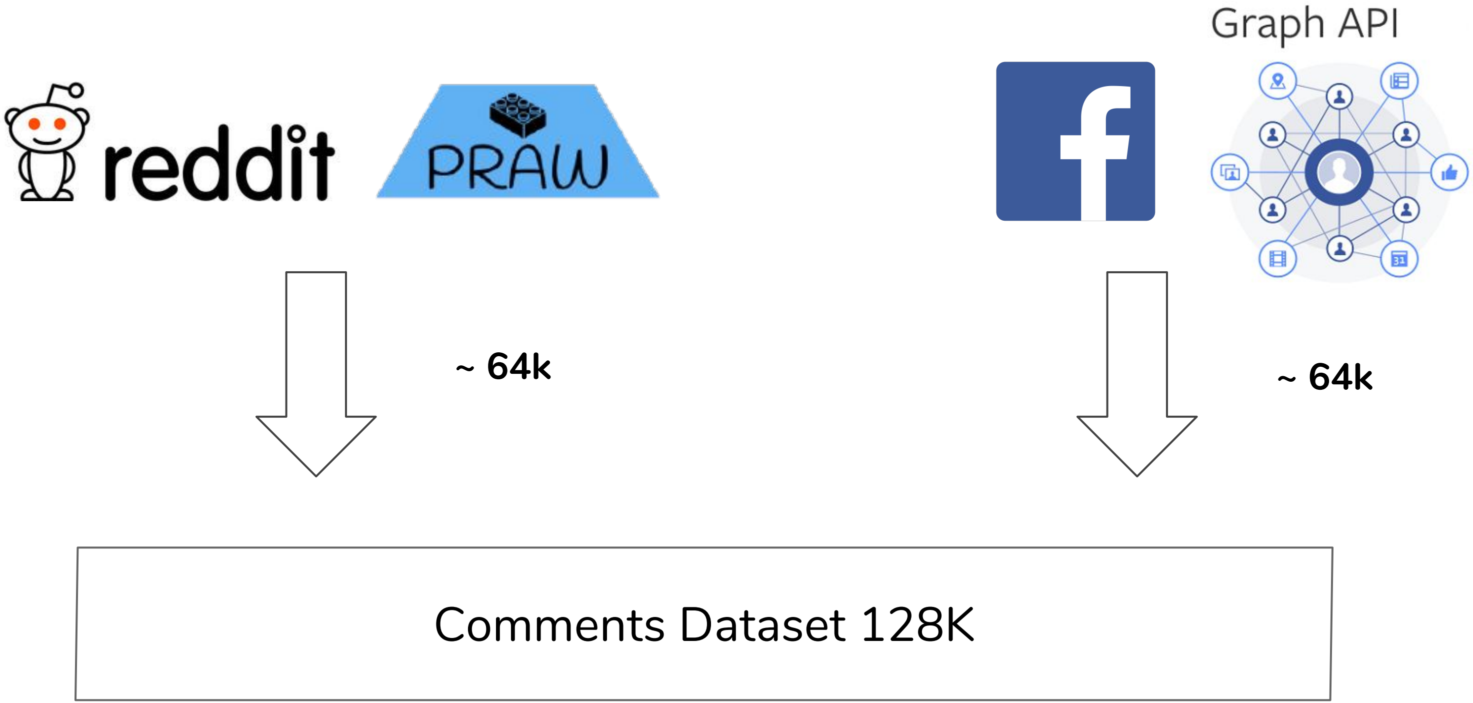
About 64,000 comments for each

Dataset

To measure the level of aggression people show on different levels of anonymity

reddit ~ 64k	f ~ 64k
/r/gaming	Steam & EA Games
/r/gunners (Arsenal FC subreddit)	Arsenal FC
/r/hockey	NHL
/r/news	NBC News & BBC News
/r/LeagueOfLegends	League of Legends

Dataset Extraction



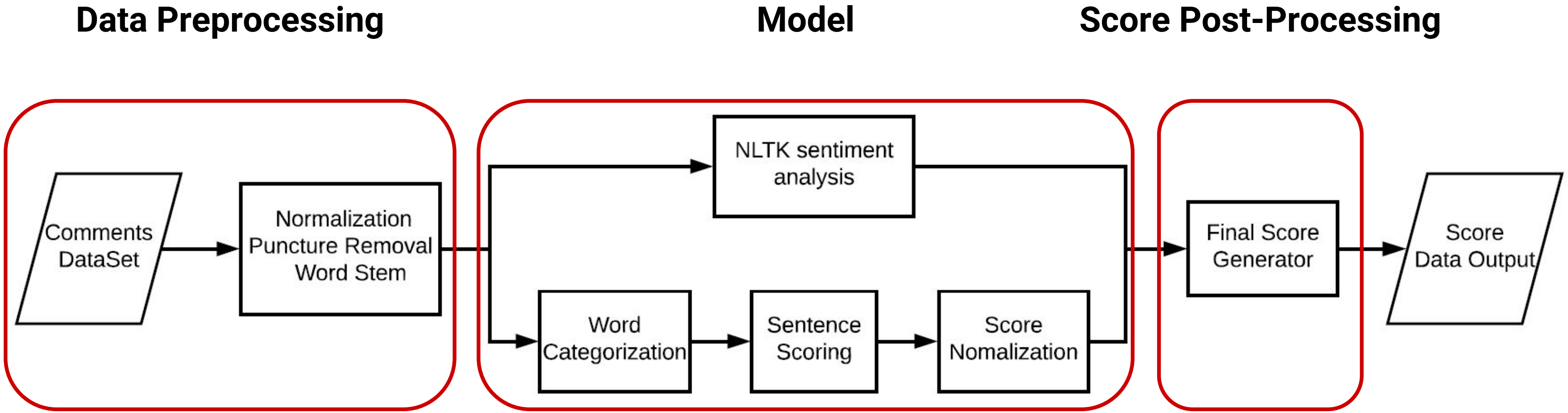
~ 64k

~ 64k

Comments Dataset 128K

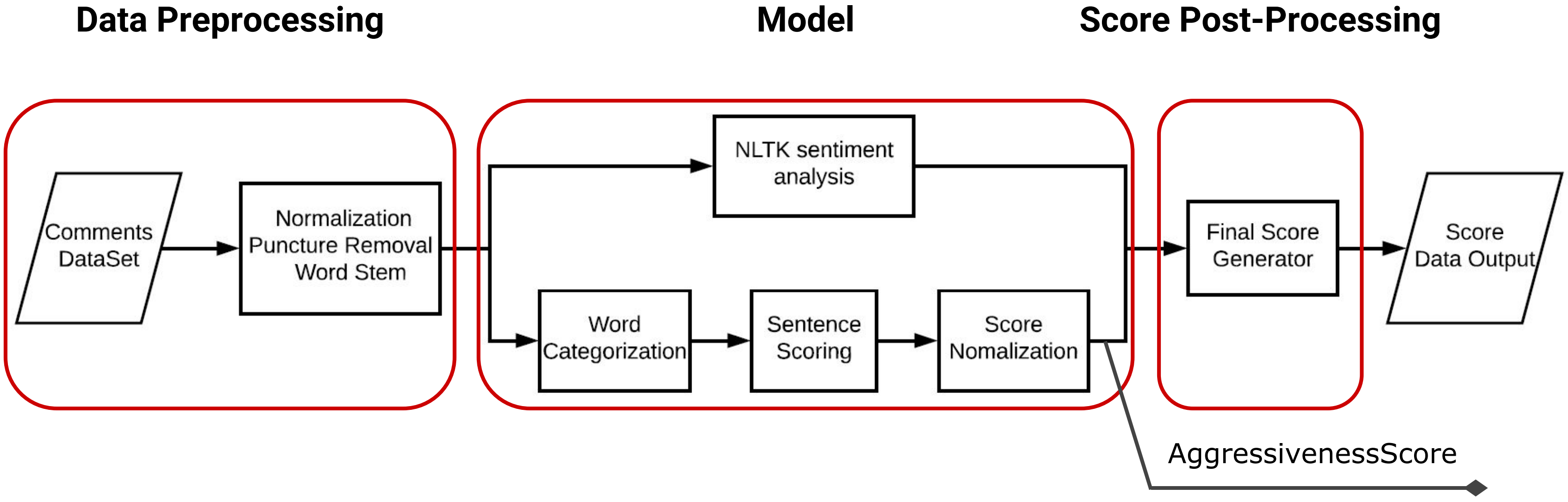
Processing & Modeling

Natural Language ToolKit (sentiment analysis), self-defined bad words library



Processing & Modeling

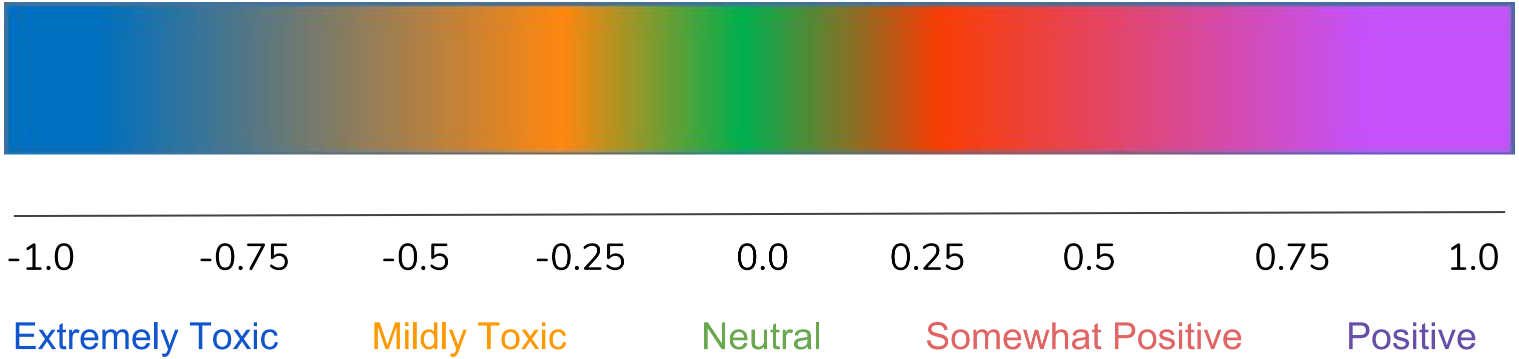
Natural Language Toolkit (sentiment analysis), self-defined bad words library



$$AggressivenessScore = normalization(\alpha + \sum_{\omega \in comments} count(\omega) * \theta_\omega)$$

$$CommentsScore = Max(Min(AggressivenessScore, -NLTK\ NegativeScore), -1)$$

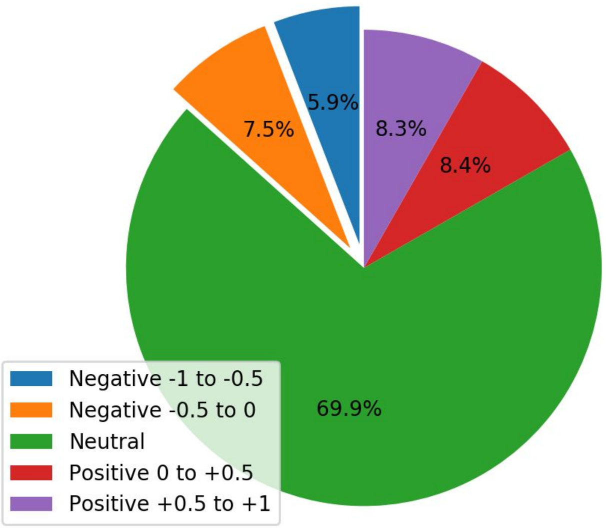
What does the score mean?



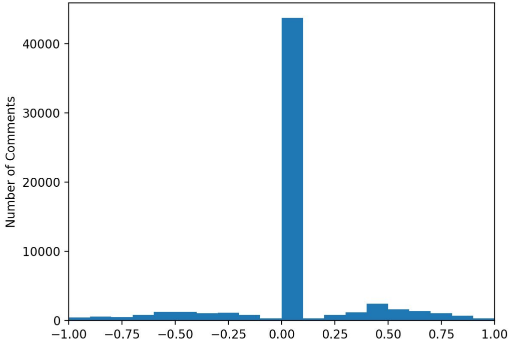
Analysis | General Analysis

To measure the level of aggression people show on different levels of anonymity

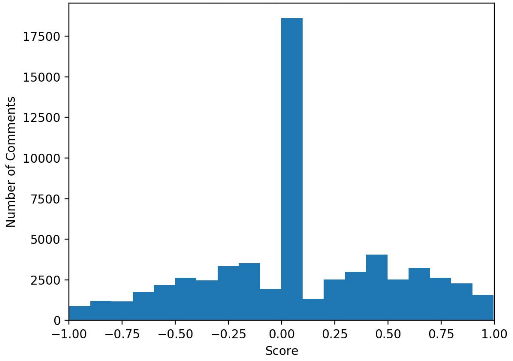
Overall_FB64K_Comments_Score



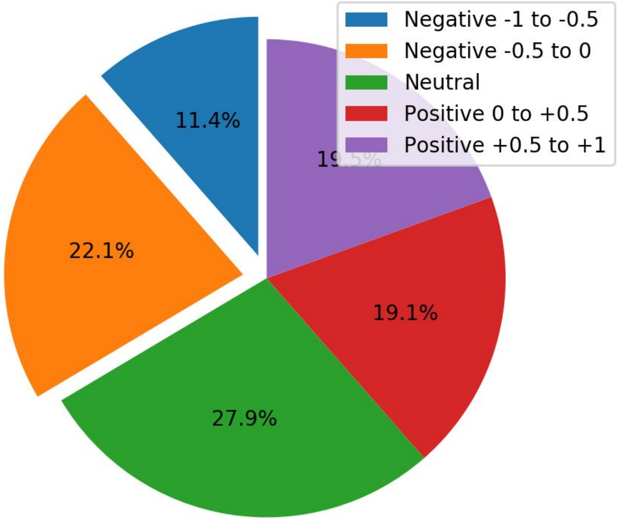
Overall_FB64K_Comments_Score



Overall_Reddit64K_Comments_Score

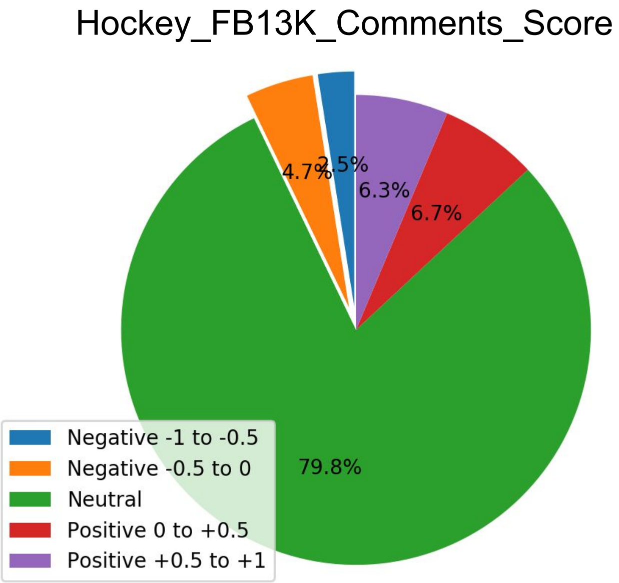


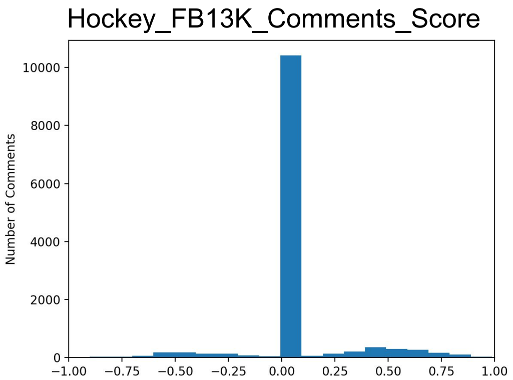
Overall_Reddit64K_Comments_Score

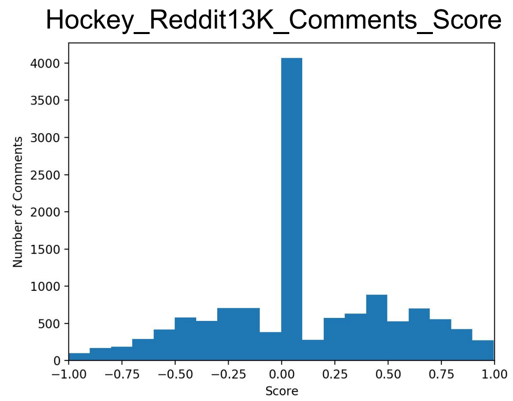


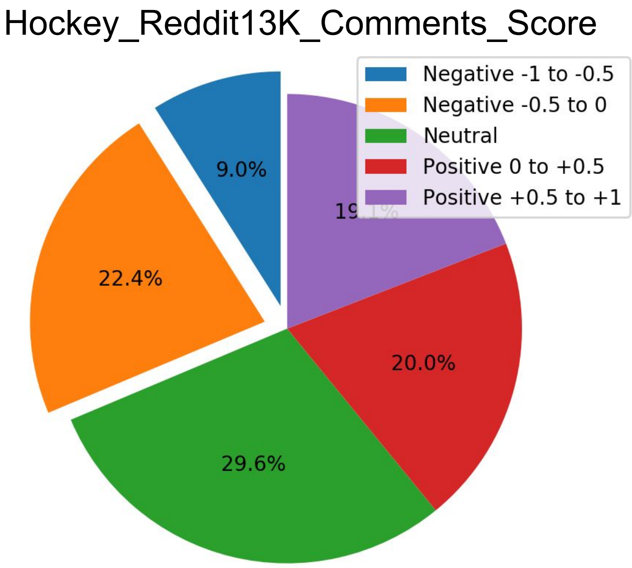
Analysis | General Analysis

To measure the level of aggression people show on different levels of anonymity



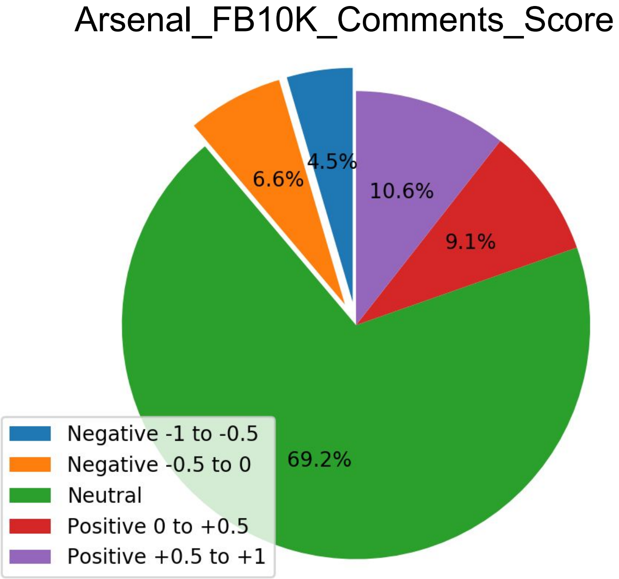


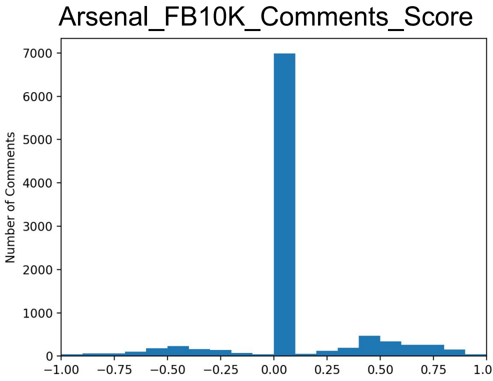


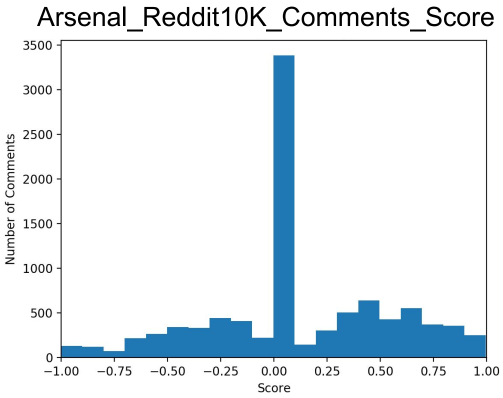


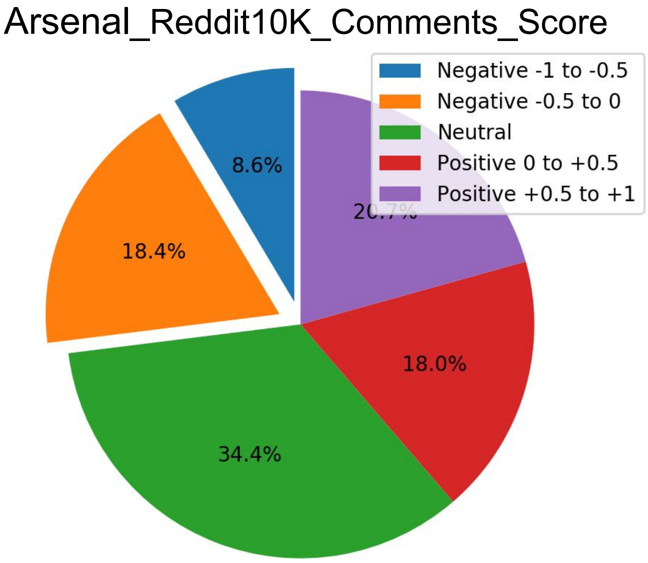
Analysis | General Analysis

To measure the level of aggression people show on different levels of anonymity



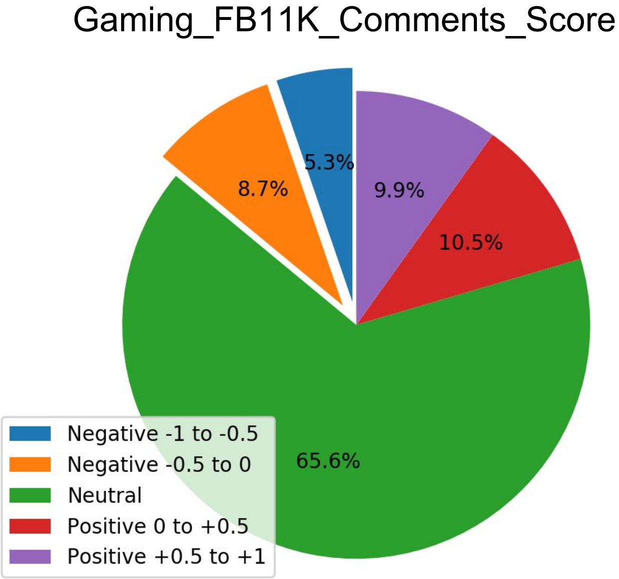


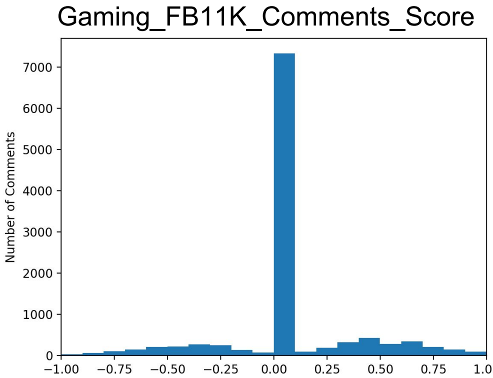


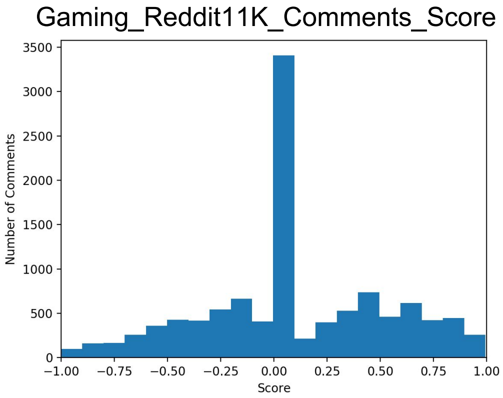


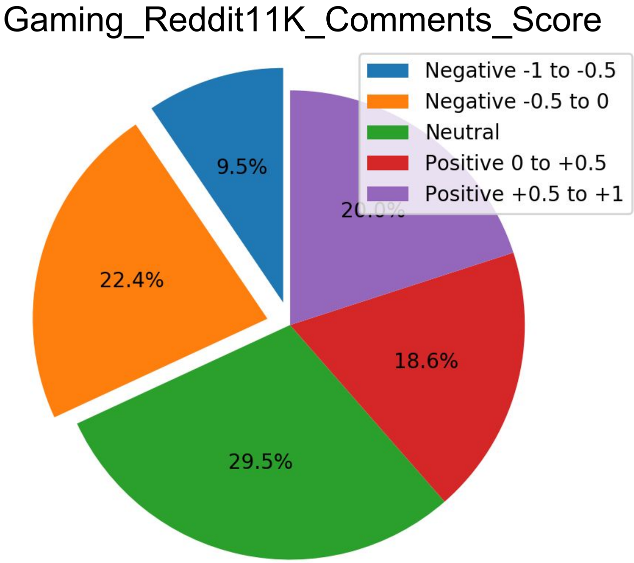
Analysis | General Analysis

To measure the level of aggression people show on different levels of anonymity



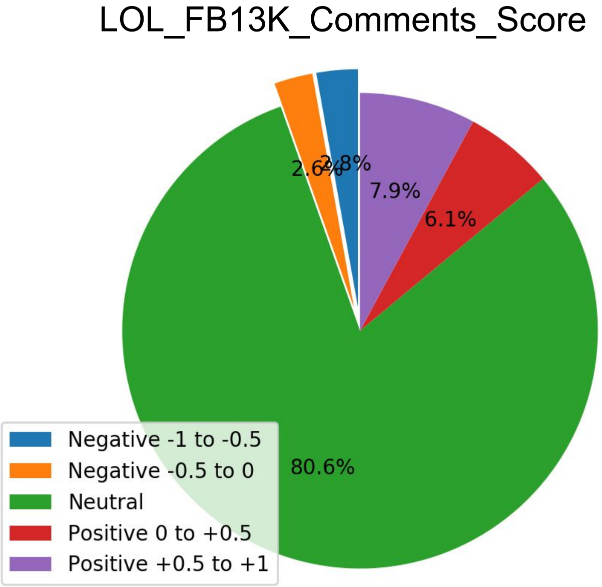




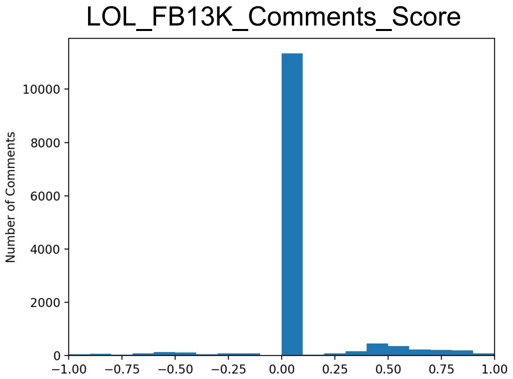


Analysis | General Analysis

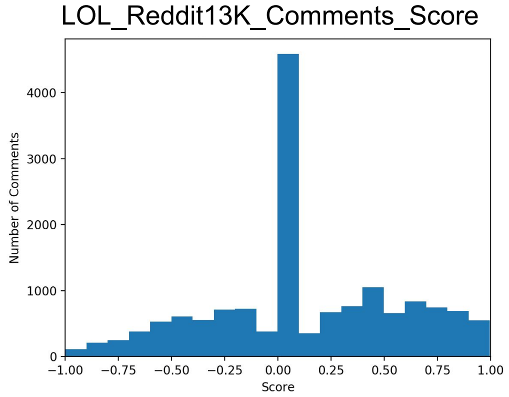
To measure the level of aggression people show on different levels of anonymity



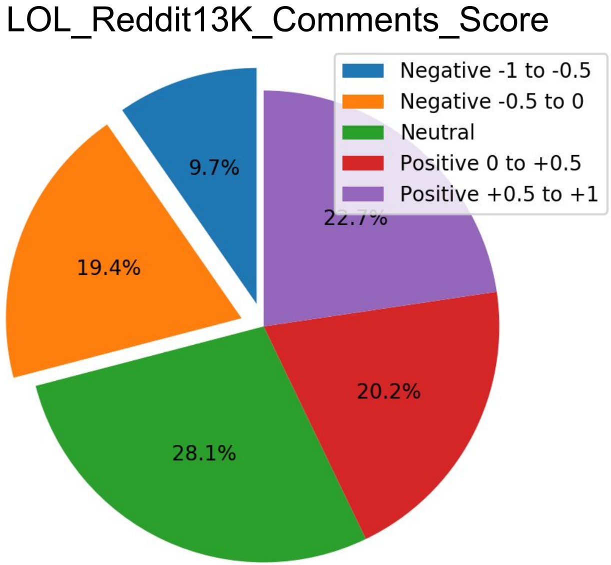
LOL_FB13K_Comments_Score



LOL_FB13K_Comments_Score



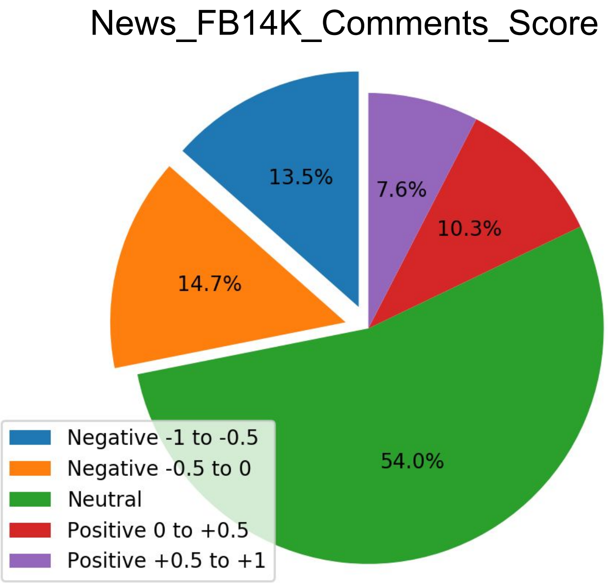
LOL_Reddit13K_Comments_Score

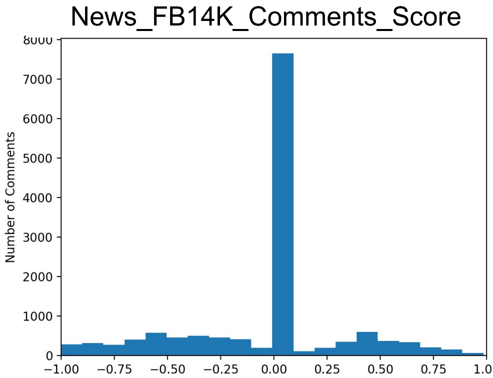


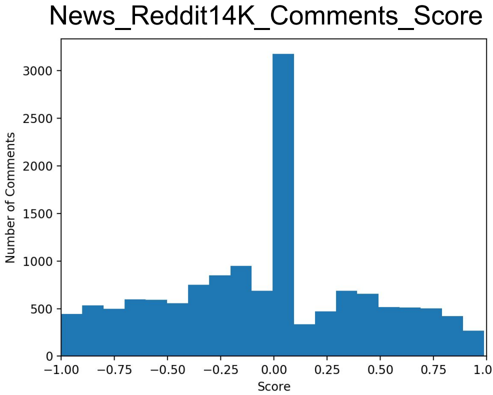
LOL_Reddit13K_Comments_Score

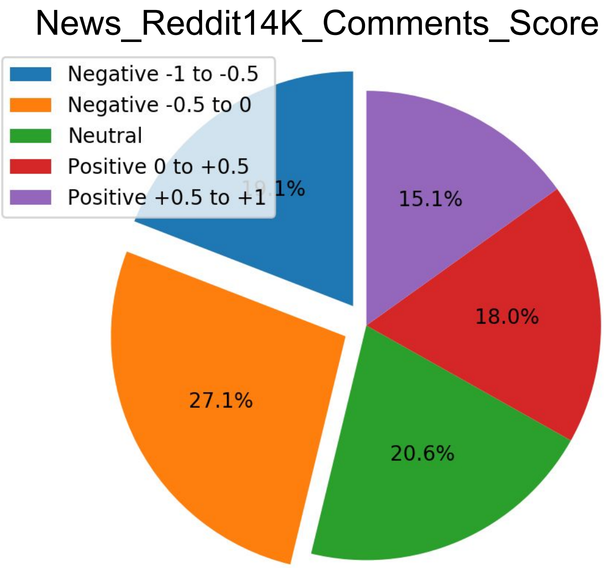
Analysis | General Analysis

To measure the level of aggression people show on different levels of anonymity





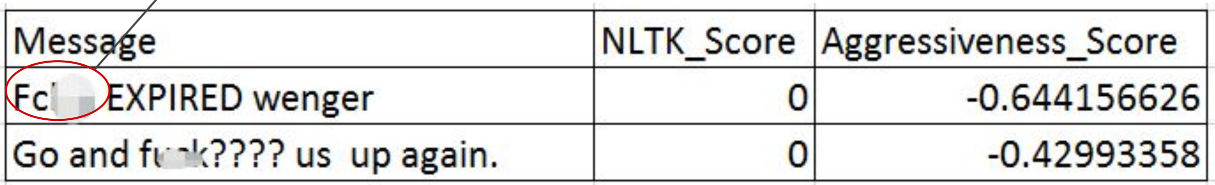




Analysis | Detailed Analysis

Our Model is better in detecting toxic comments
Some comments with toxic words are spelled incorrectly on purpose which the NLTK regards it as neutral. But our model can successfully detect those toxic comments.

Comment with ‘fc*’ instead of ‘fu**’



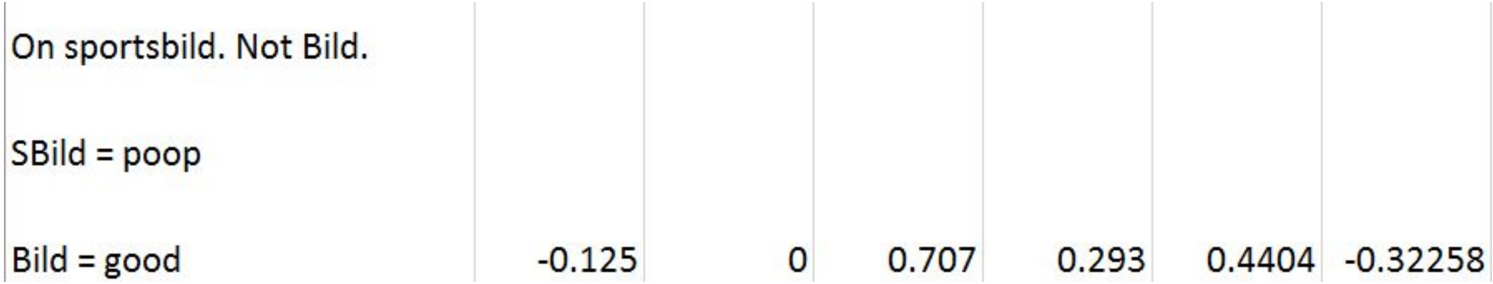
Message	NLTK_Score	Aggressiveness_Score
Fc[redacted] EXPIRED wenger	0	-0.644156626
Go and f[redacted]k???? us up again.	0	-0.42993358

Table: Cases our model perform better than NLTK model

Analysis | Detailed Analysis

To measure the level of aggression people show on different levels of anonymity

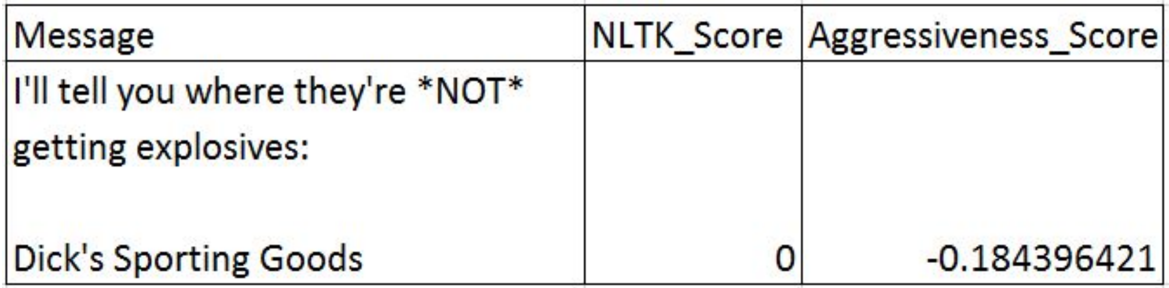
In some cases, NLTK does not understand some phrase and shows a positive result. But in our model, we don't have to understand the whole sentence so that we can detect those partial toxicity in the sentence.



On sportsbild. Not Bild. SBild = poop Bild = good	-0.125	0	0.707	0.293	0.4404	-0.32258

Analysis | Detailed Analysis

Case NLTK model performs better than our model
Our model sometimes is confused with words having multi-meaning. When names appears in our dictionary and the name also have toxic meaning, like ‘Dick’, our model cannot judge these names correctly. In this case, NLTK shows better result (score those comments with zero as ‘neutral’)

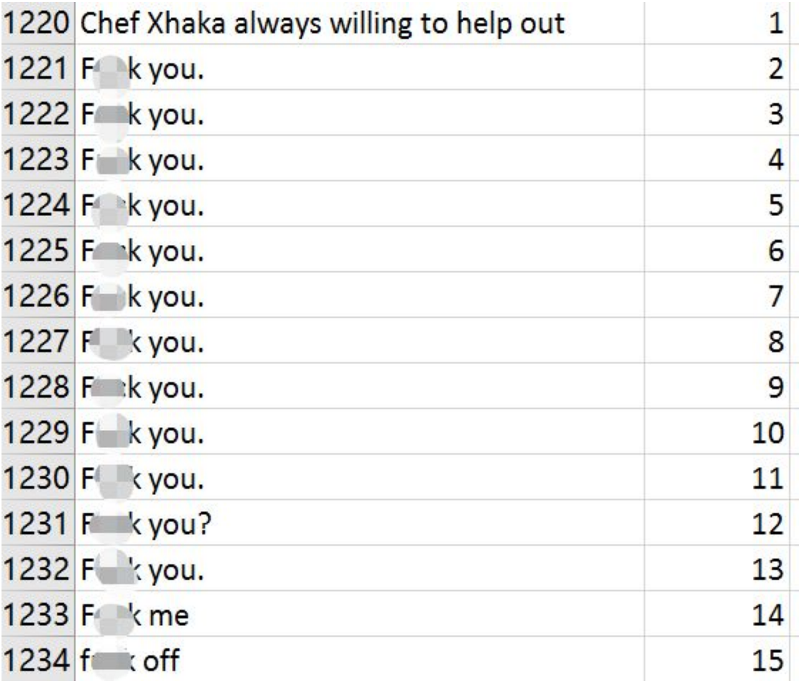


Message	NLTK_Score	Aggressiveness_Score
I'll tell you where they're *NOT* getting explosives: Dick's Sporting Goods	0	-0.184396421

Table: Special case of detecting name words

Analysis | Detailed Analysis

Chain Effect
When someone post an aggressive comment and a few people follow him with toxic replies, other people are likely to reply in the same way (with toxic expressions)



1220	Chef Xhaka always willing to help out	1
1221	F[redacted]k you.	2
1222	F[redacted]k you.	3
1223	F[redacted]k you.	4
1224	F[redacted]k you.	5
1225	F[redacted]k you.	6
1226	F[redacted]k you.	7
1227	F[redacted]k you.	8
1228	F[redacted]k you.	9
1229	F[redacted]k you.	10
1230	F[redacted]k you.	11
1231	F[redacted]k you?	12
1232	F[redacted]k you.	13
1233	F[redacted]k me	14
1234	f[redacted]k off	15

Table: Special case of detecting comment chain

Future Possibilities

- More comprehensive dictionary on our model
  - Detect toxicity more precisely
- Optimization
  - Very long running time
- Explores the reason behind that result
  - Learn more linguistic knowledge.

Github : compare-toxicity   (We also have **one-click demo** on the repo!!)

Thank You

ECE 180 Project by Team 1