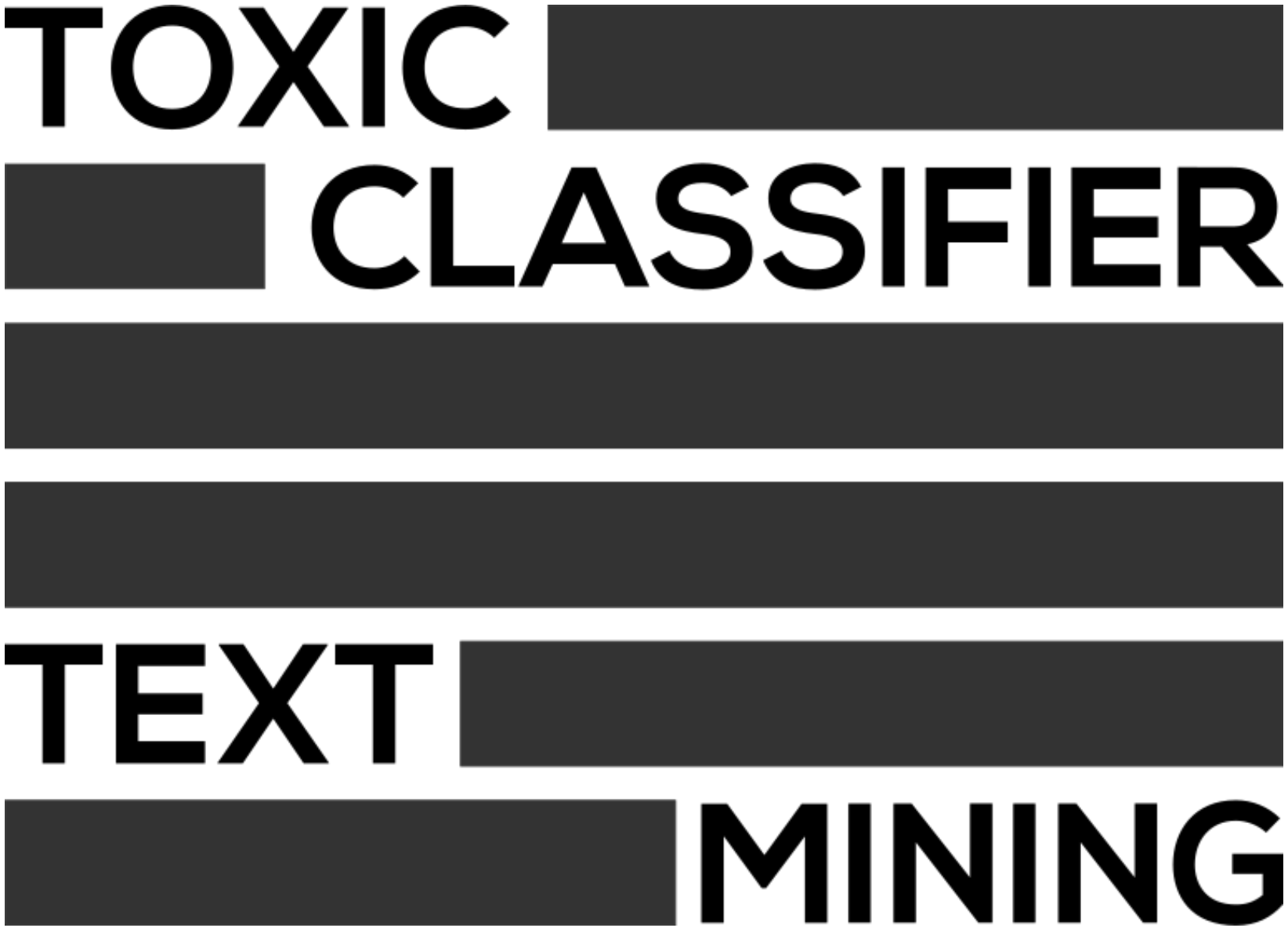
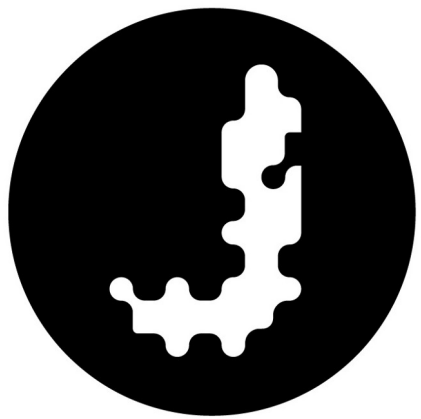

TOXIC
CLASSIFIER
TEXT
MINING

# *Data*



Jigsaw



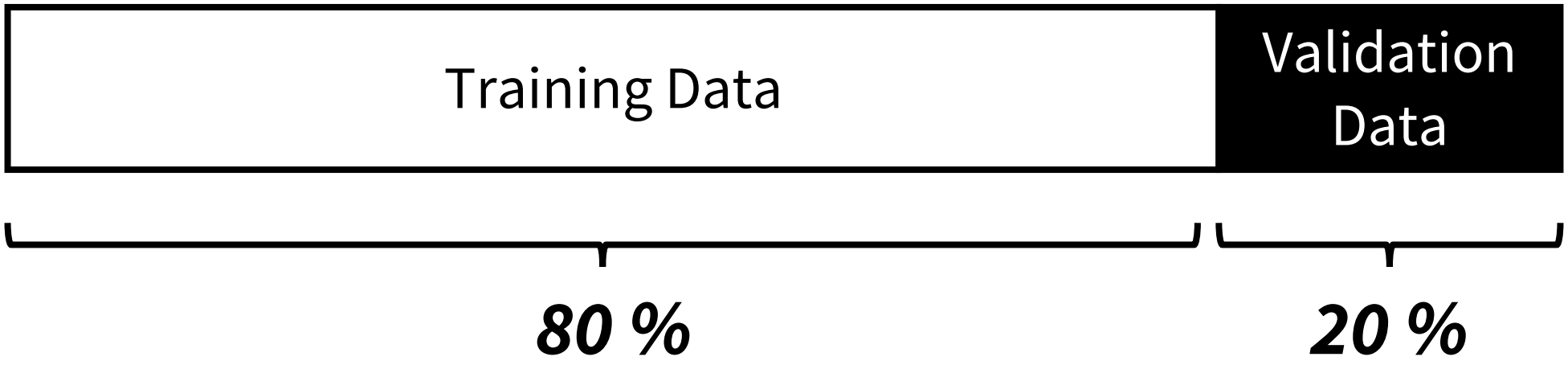
*Source:* https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge/data

# *Data*

***159 571 Observations***

Training Data

# *Data*



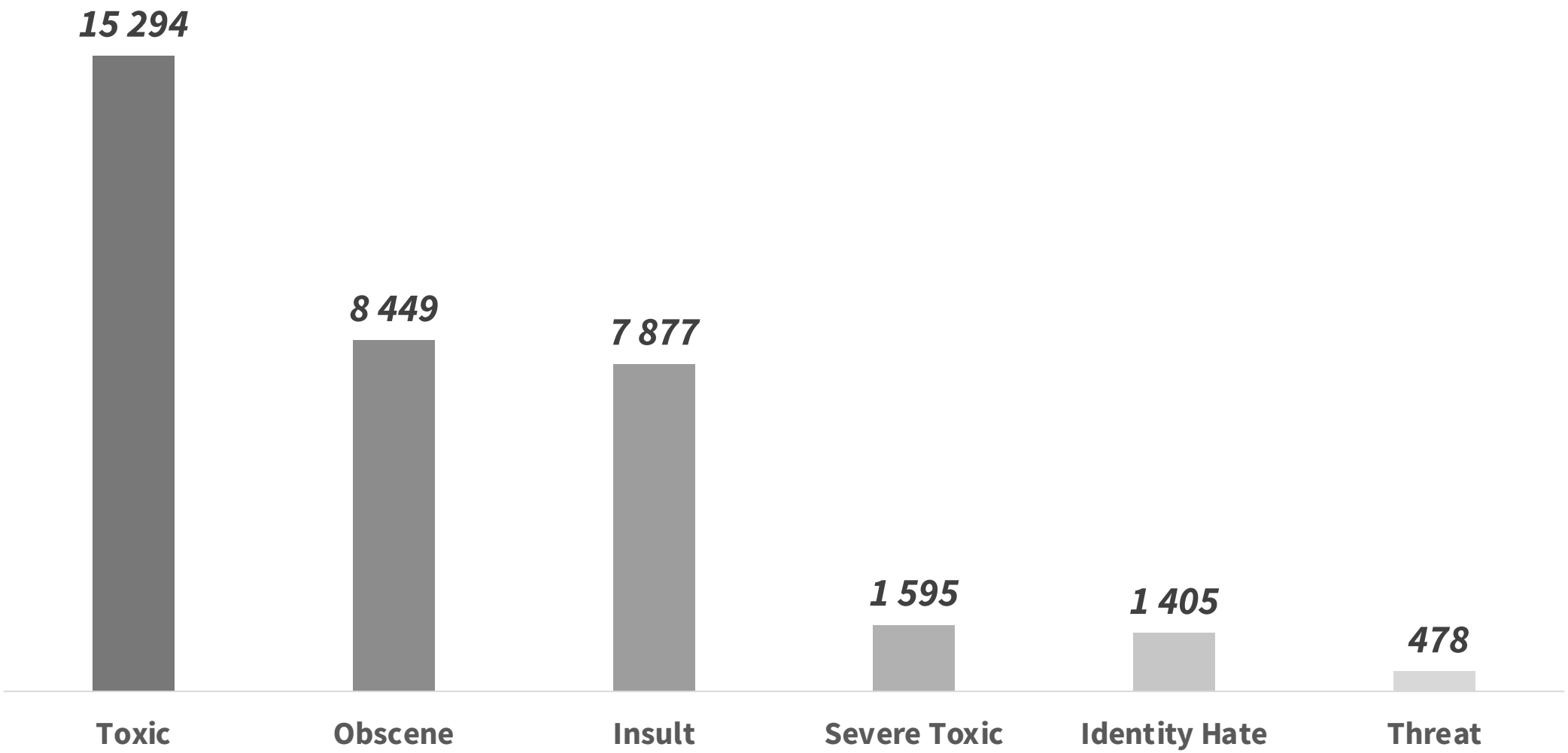
# *Preprocess*

| id | Comment Text | Toxic | Severe Toxic | Obscene | Threat | Insult | Identity Hate |
|---|---|---|---|---|---|---|---|
| 1 | … | 0 | 0 | 0 | 0 | 0 | 0 |
| 2 | … | 1 | 0 | 0 | 1 | 0 | 0 |
| 3 | … | 0 | 0 | 0 | 0 | 0 | 0 |

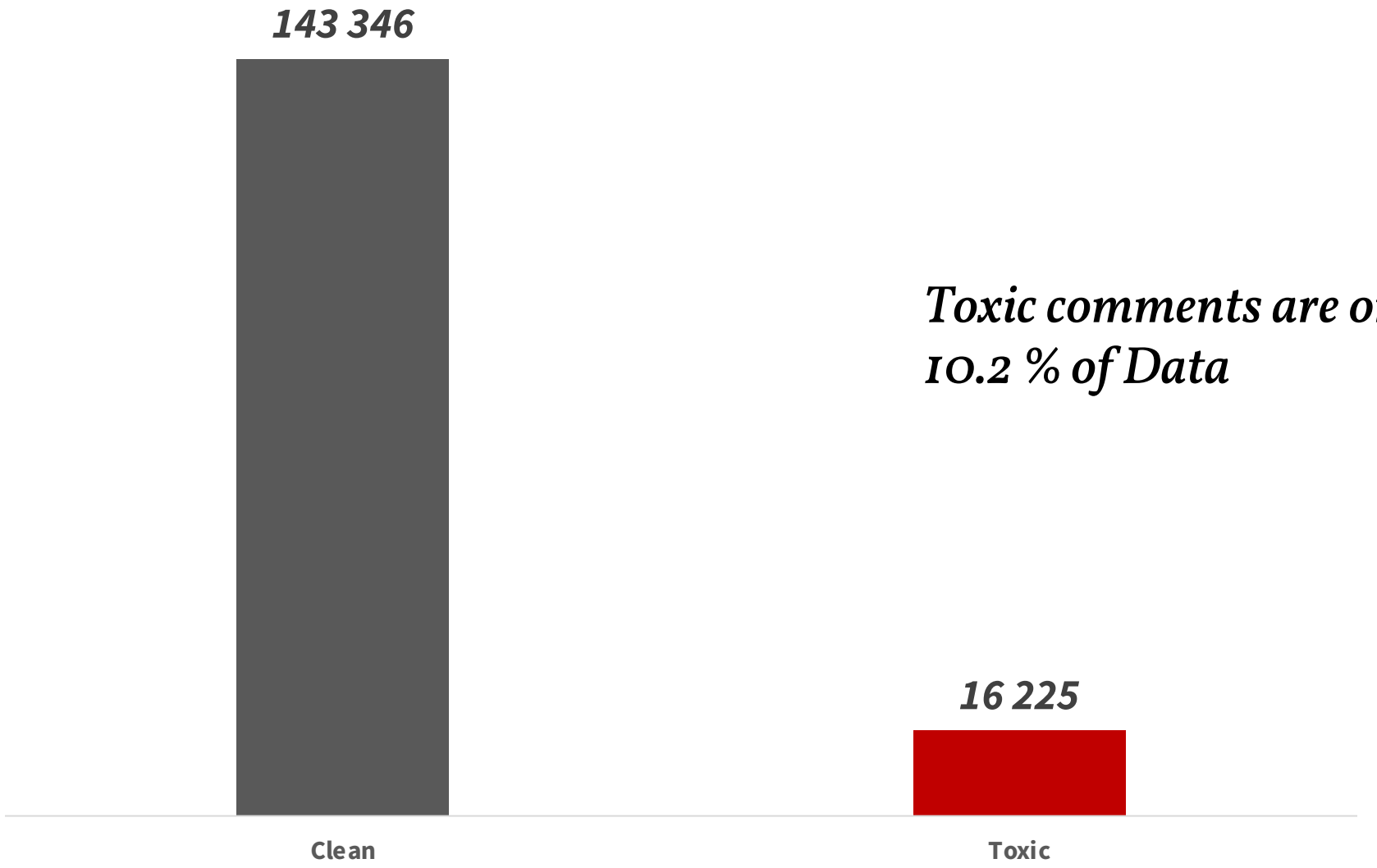
# *Preprocess*

| id | Comment Text | Toxic | Severe Toxic | Obscene | Threat | Insult | Identity Hate | **Target** |
|---|---|---|---|---|---|---|---|---|
| 1 | … | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| **2** | **…** | **1** | **0** | **0** | **1** | **0** | **0** | **1** |
| 3 | … | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

# *Preprocess*

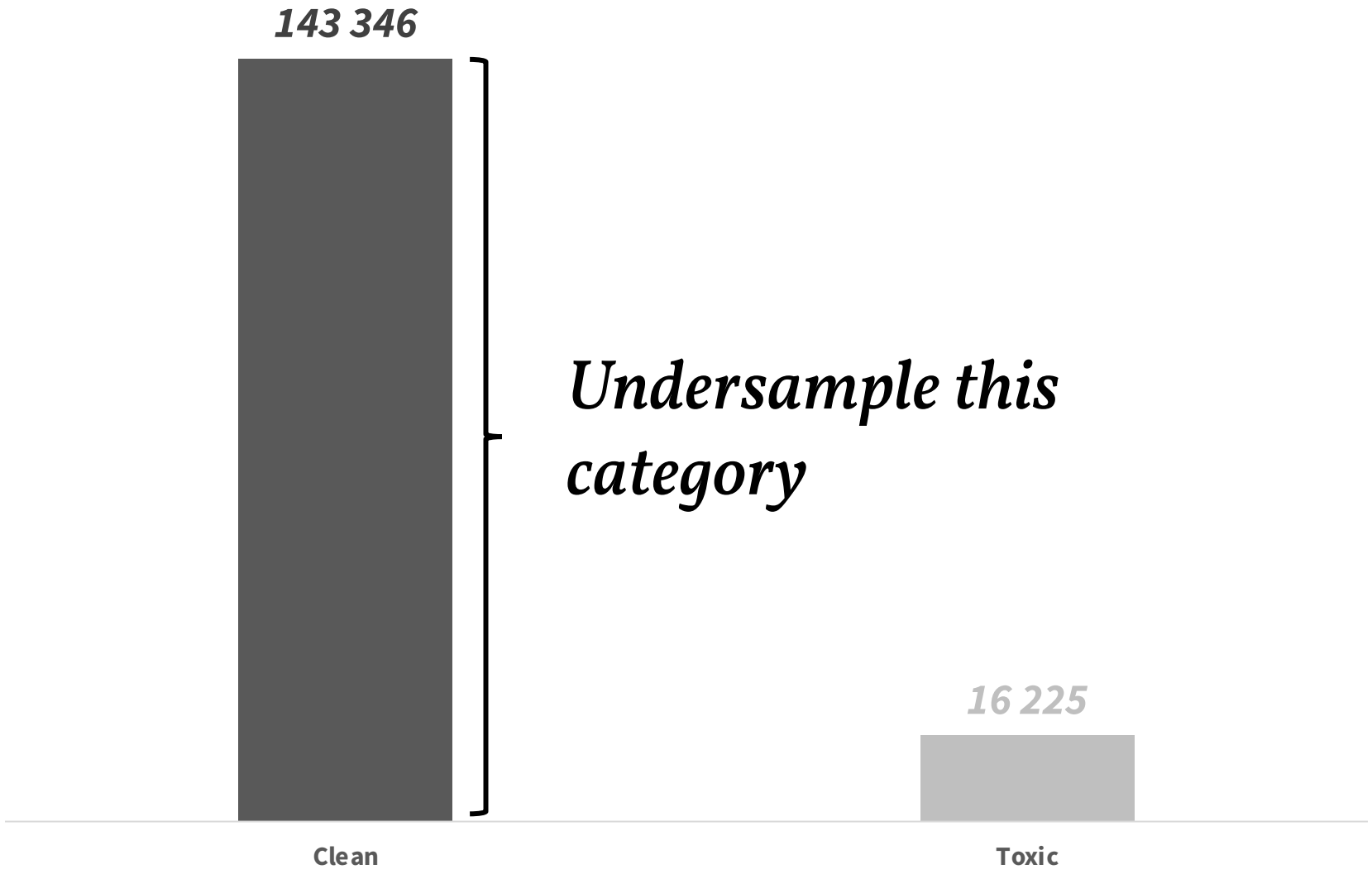

15 294
8 449
7 877
1 595
1 405
478
Toxic
Obscene
Insult
Severe Toxic
Identity Hate
Threat

# Preprocess

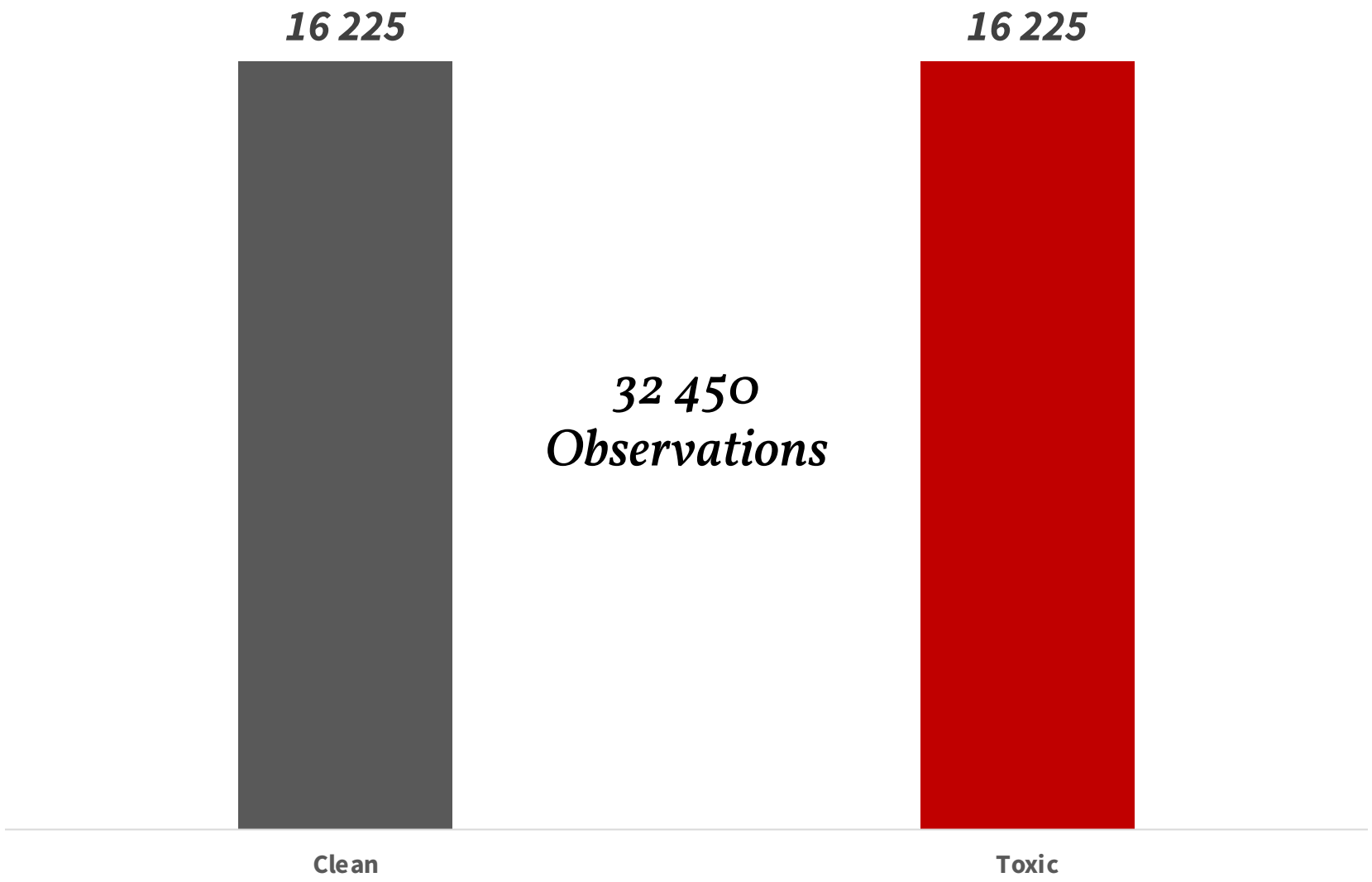



*Toxic comments are only 10.2 % of Data*

# Preprocess



# *Preprocess*



# *Preprocess*

## *1. Balance Dataset*

# *Preprocess*

## *1. Balance Dataset*

## *2. New Features*

- Character count
- Word count
- Word density
- Ponctuation count
- Title word count
- Upper case word count

# *Preprocess*

## *1. Balance Dataset*

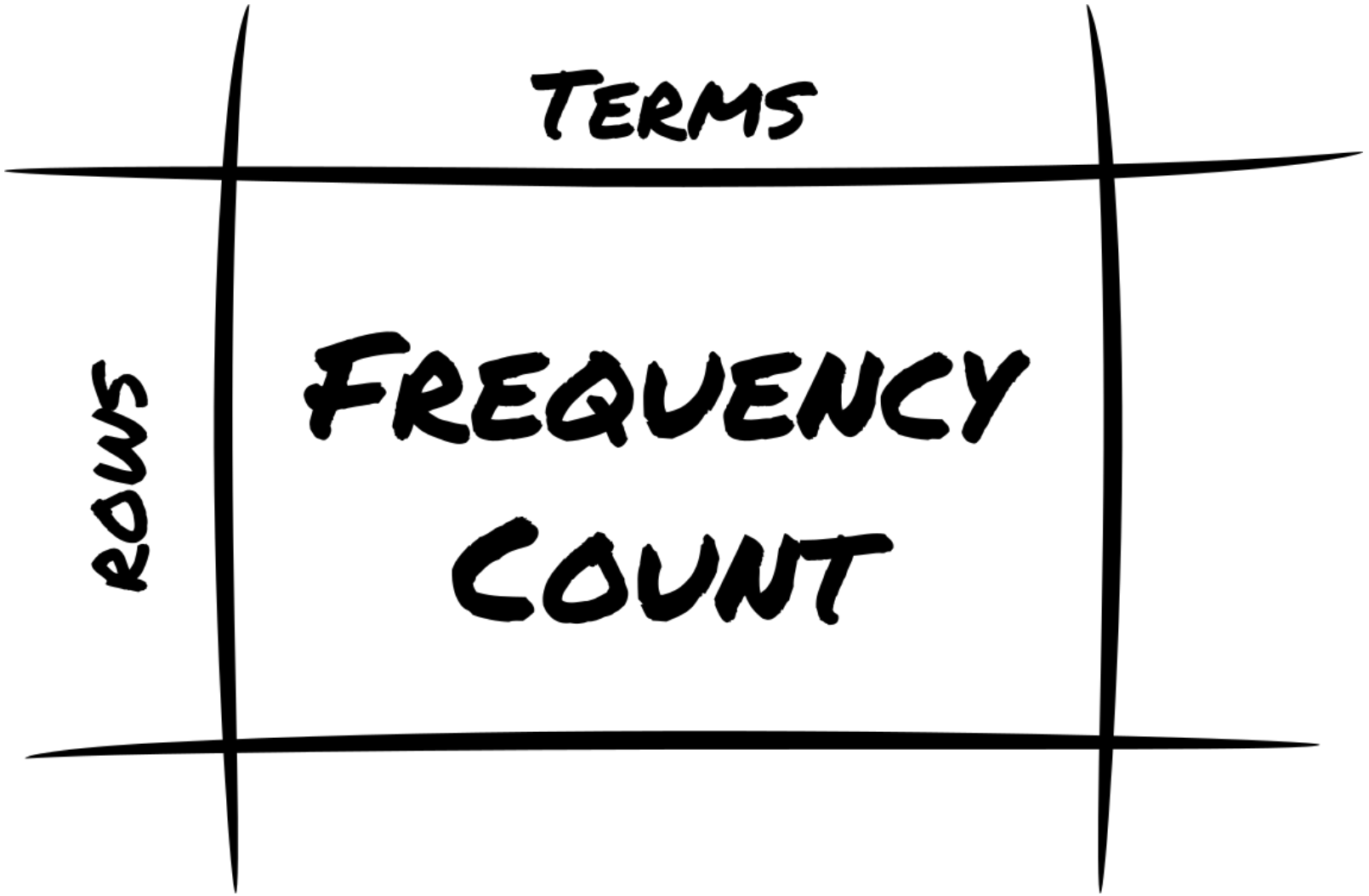
## *2. New Features*

- Character count
- Word count
- Word density
- Ponctuation count
- Title word count
- Upper case word count

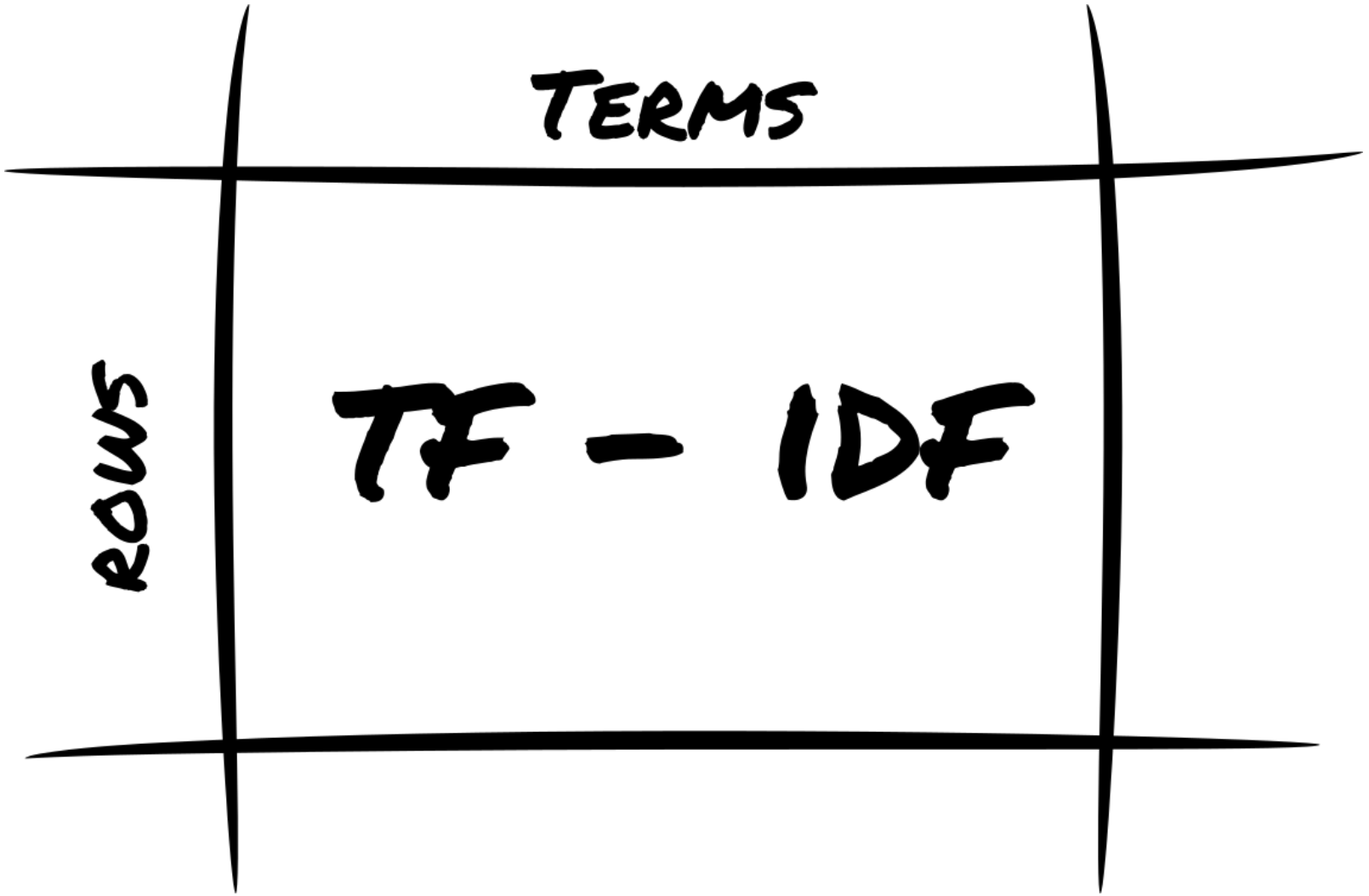
## *3. Clean Text*

- Lower Case
- Remove ponctuation
- Tokenize
- Remove stopwords
- Stem words

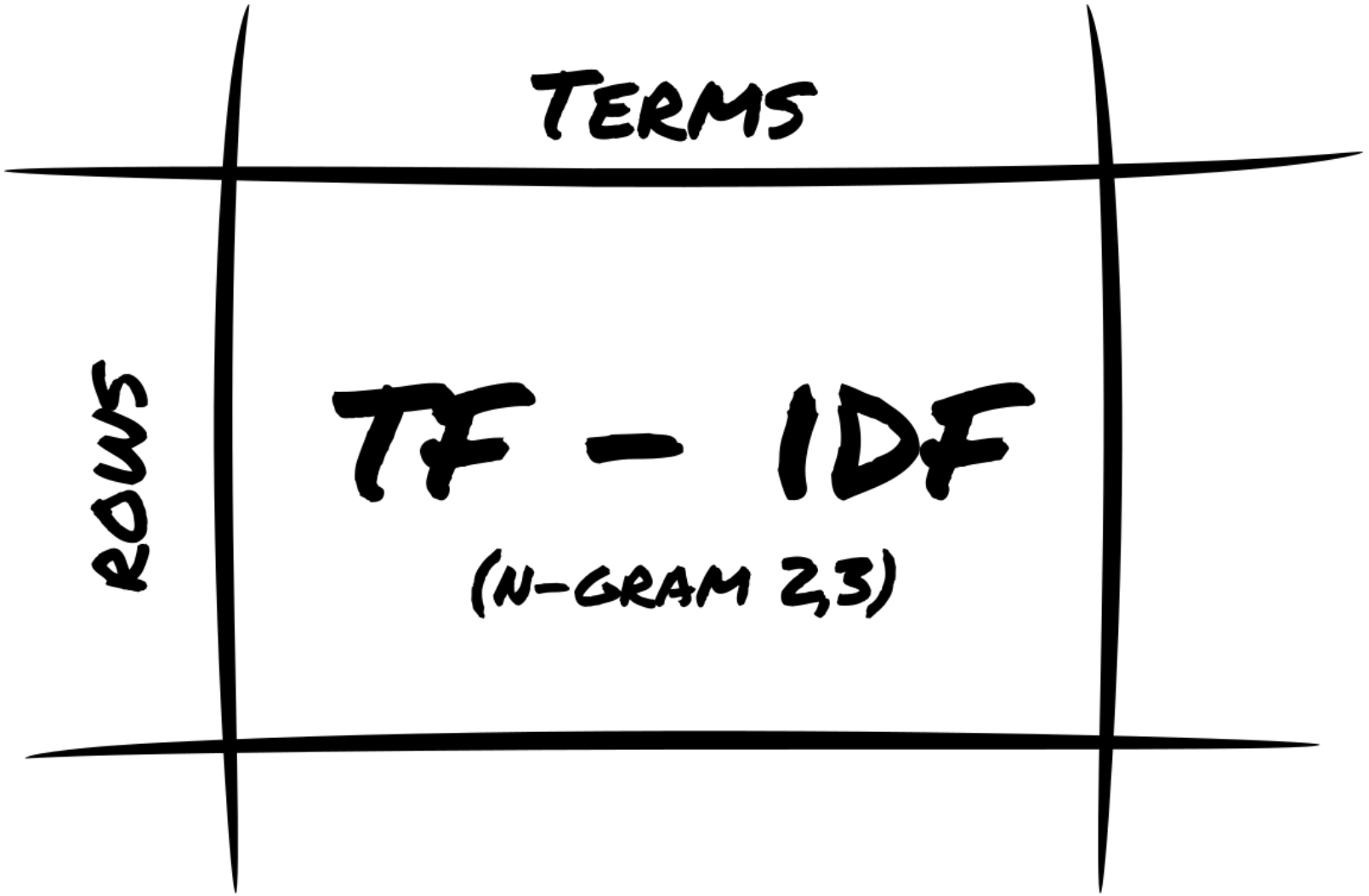
# Feature Engineering - Vectorizers



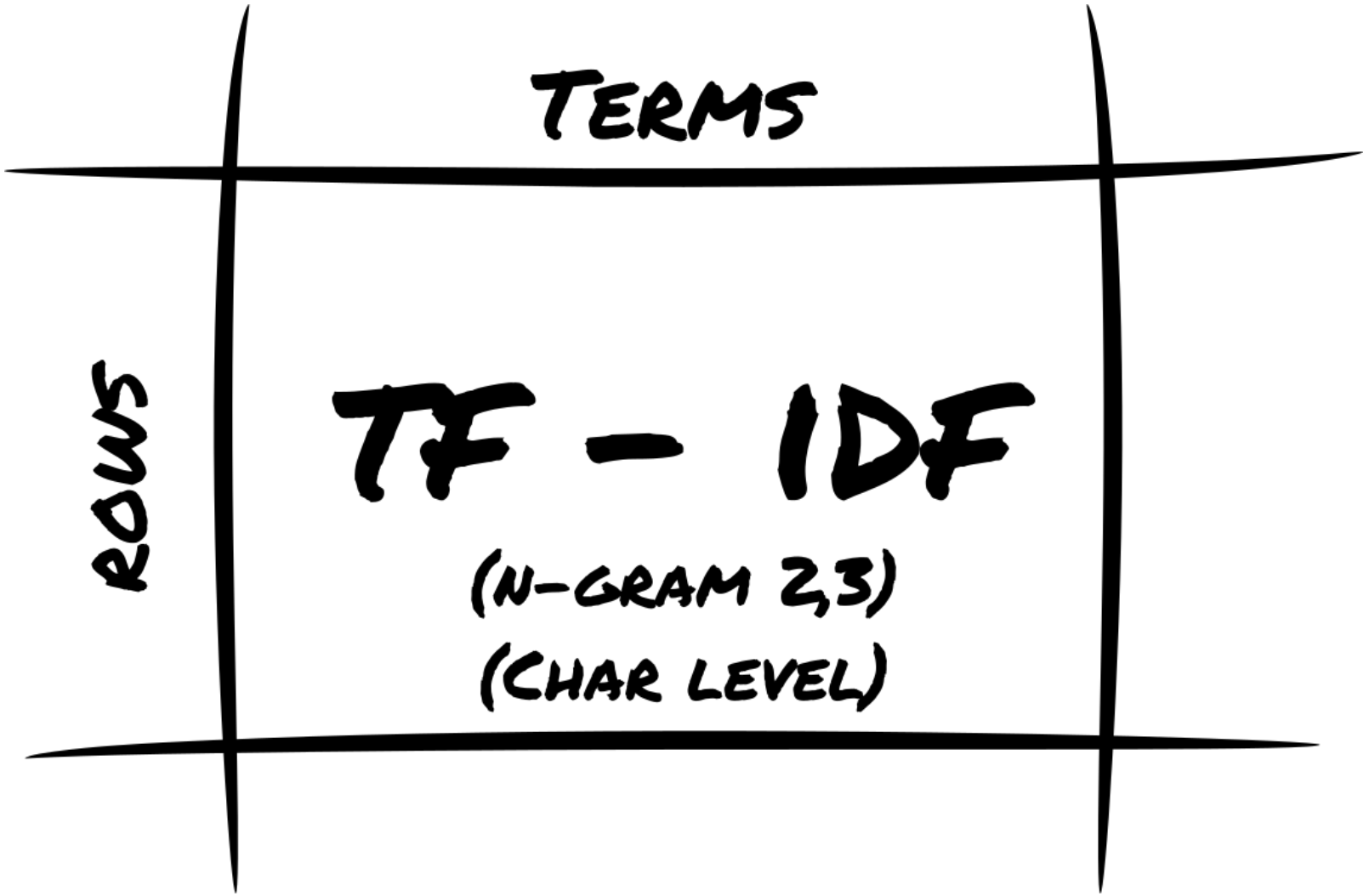
# *Feature Engineering - Vectorizers*



# *Feature Engineering - Vectorizers*



# Feature Engineering - Vectorizers



# *Models*

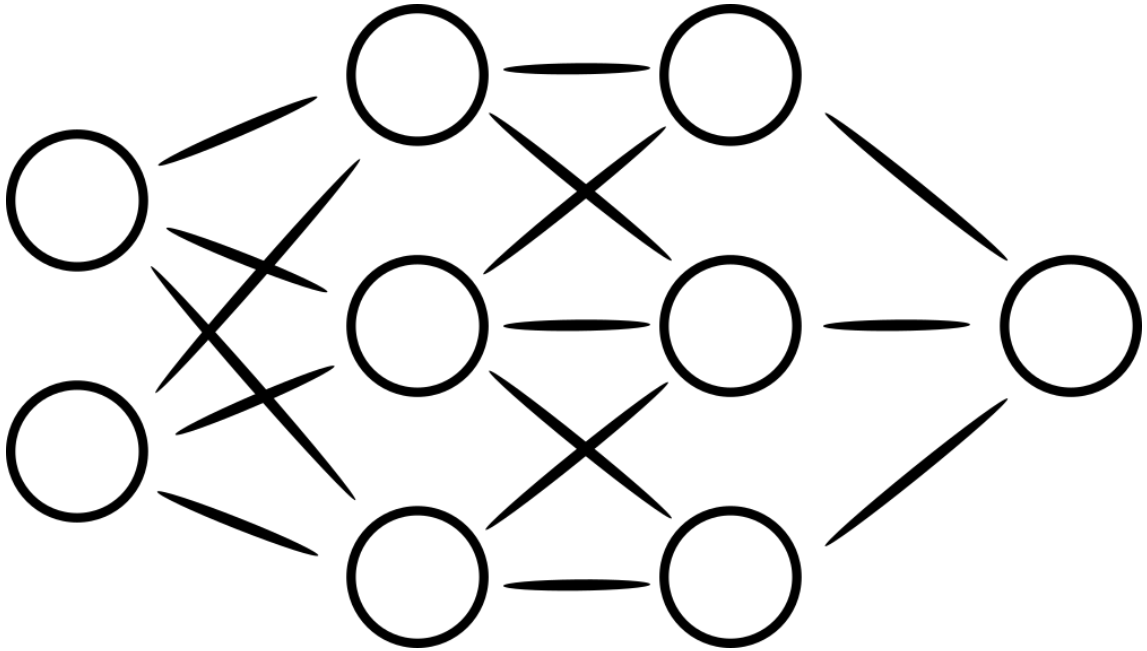
## *We tested some simpler models*

- Logistic Regression
- Multinomial NB
- Random Forest Classifier
- Stochastic Gradient Descent Classifier
- Decision Tree Classifier
- XGBoost Classifier

# *Models*

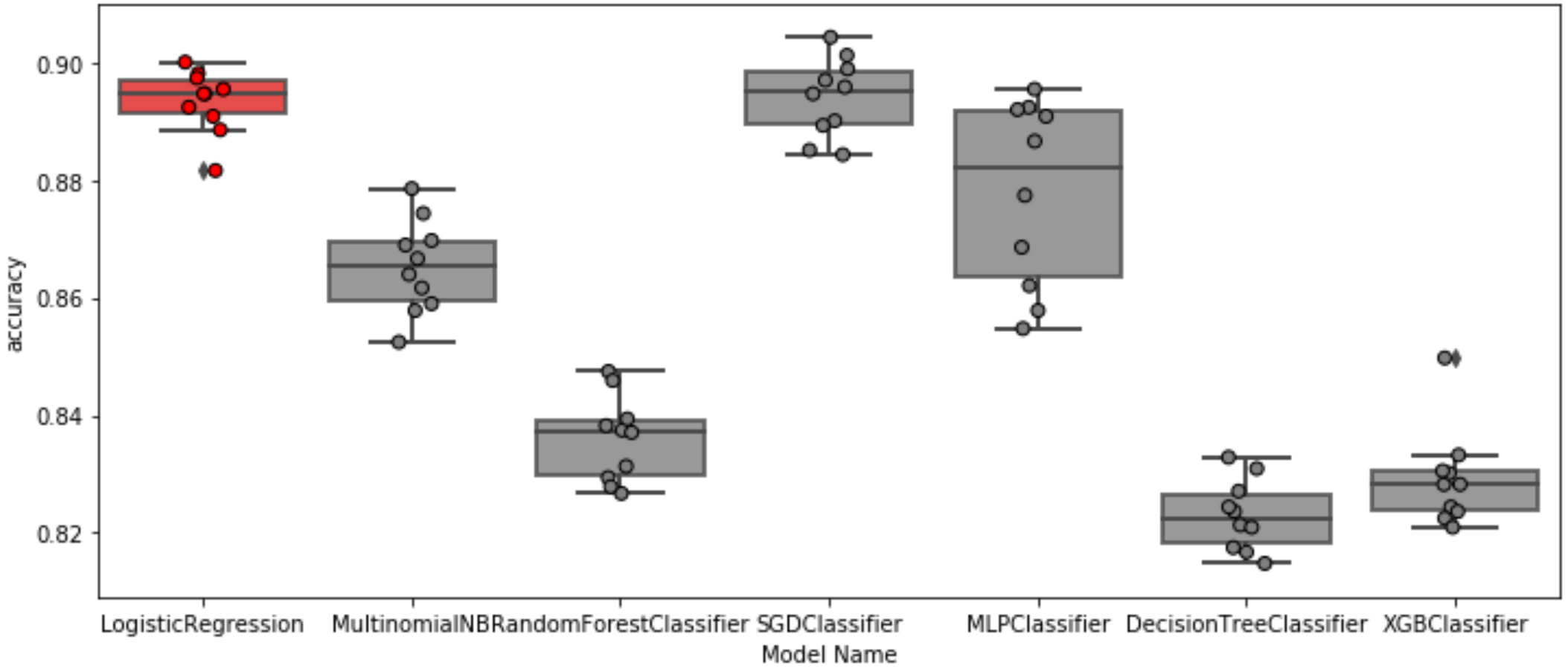
## *And some more complicated - NN*

- MLP
- Deep Learning applications using Keras



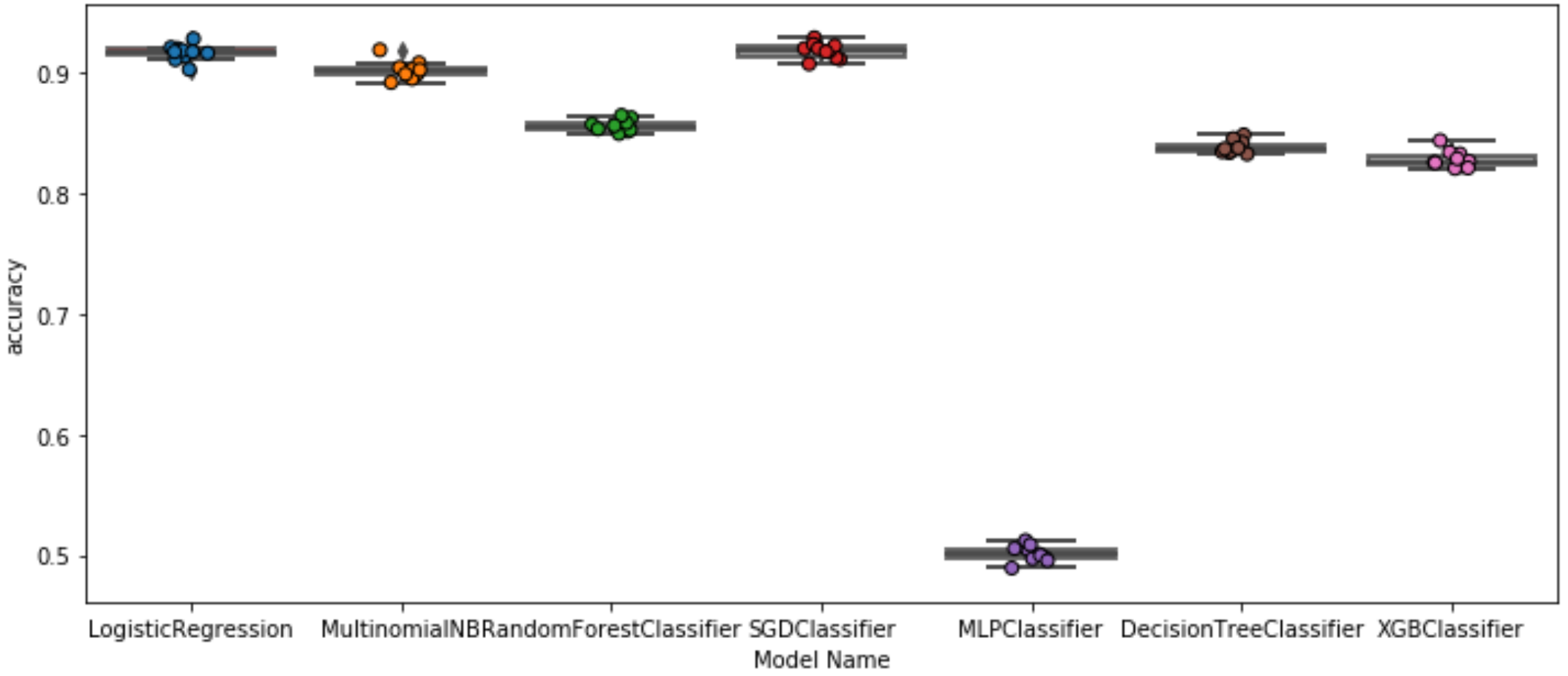
# Models

## Without feature selection



# Models

## With feature selection



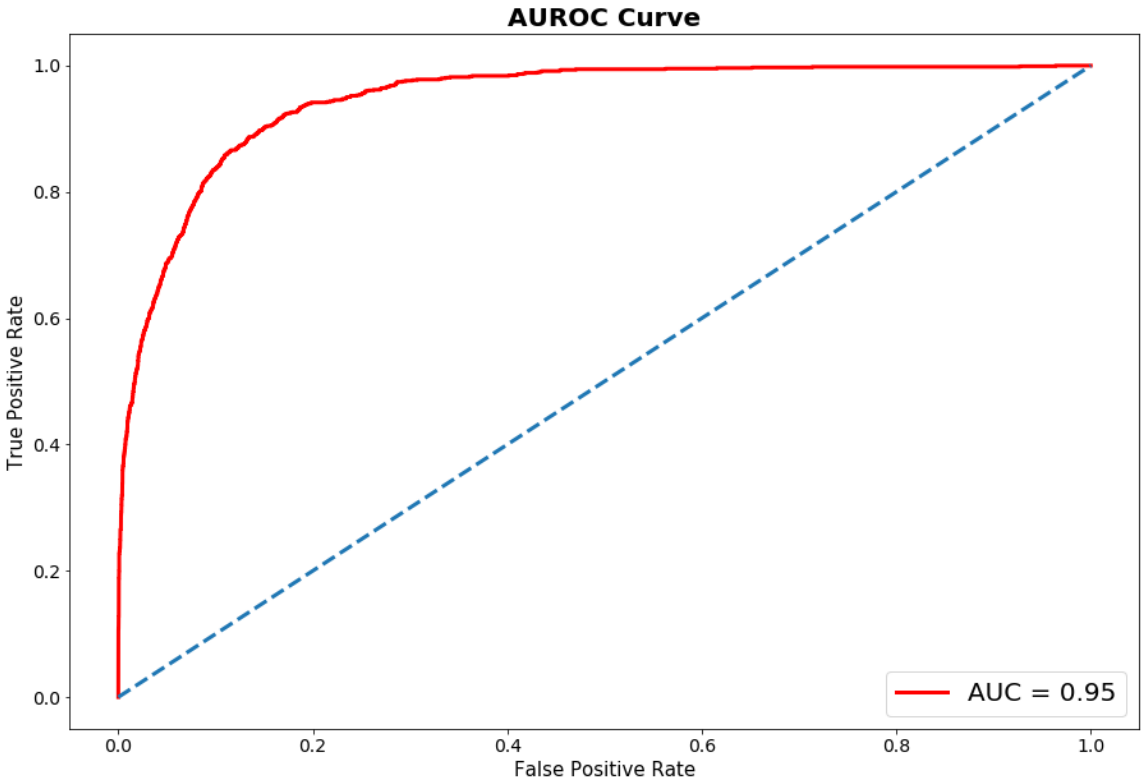
# *Models*

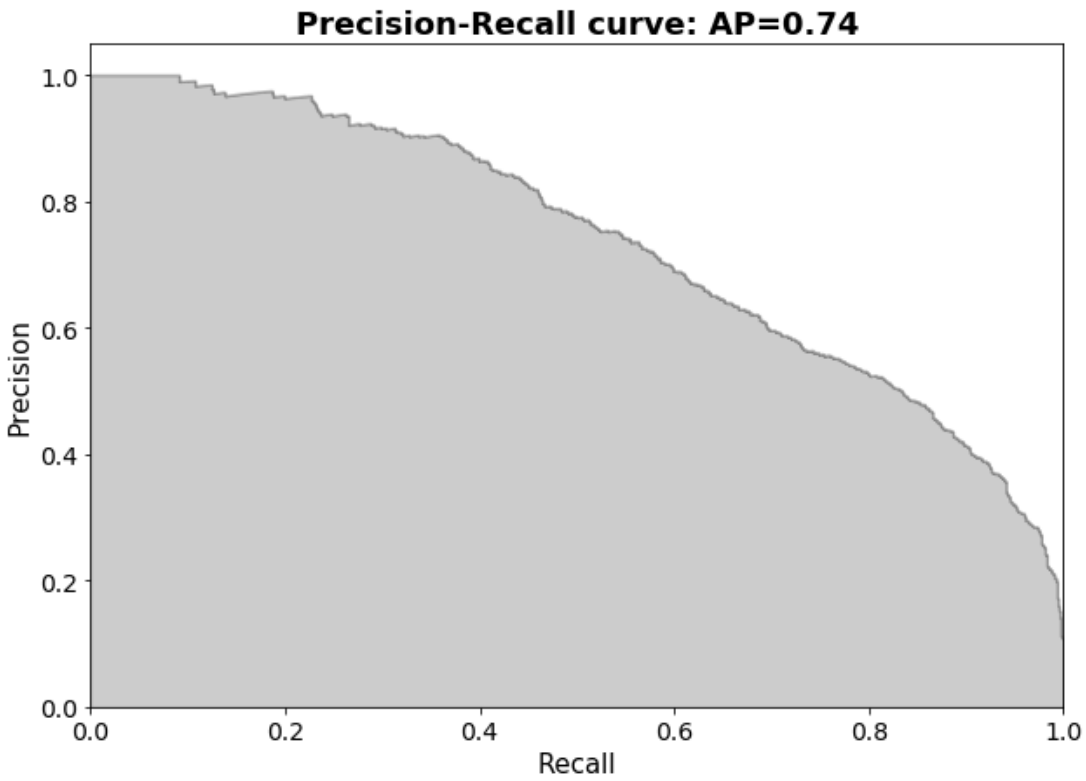
## *Final Model: SGDClassifier with grid search*

| | Precision | Recall | F1 | Support |
|---|---|---|---|---|
| **0** | 0.99 | 0.85 | 0.91 | 8954 |
| **1** | 0.41 | 0.90 | 0.56 | 1046 |
| **Weighted AVG** | 0.93 | 0.85 | 0.87 | 10000 |

Accuracy: 0.852

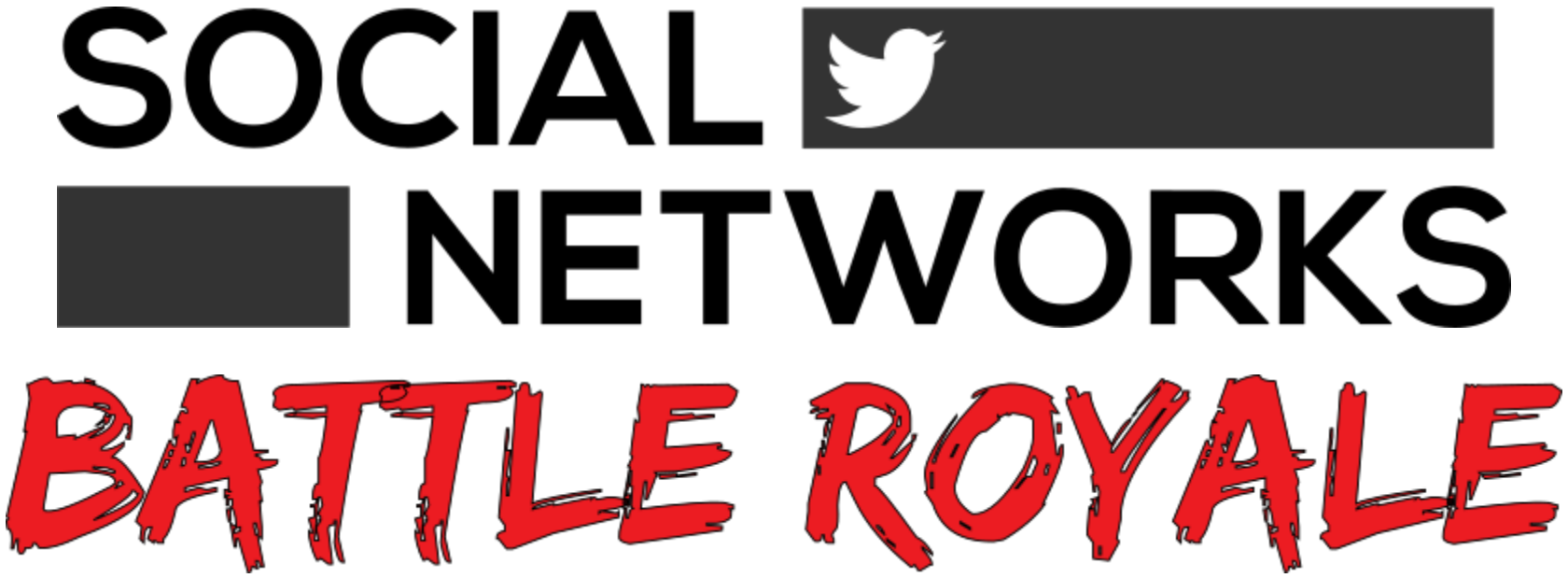
# Models


AUROC Curve
True Positive Rate
False Positive Rate
AUC = 0.95
0.0
0.2
0.4
0.6
0.8
1.0

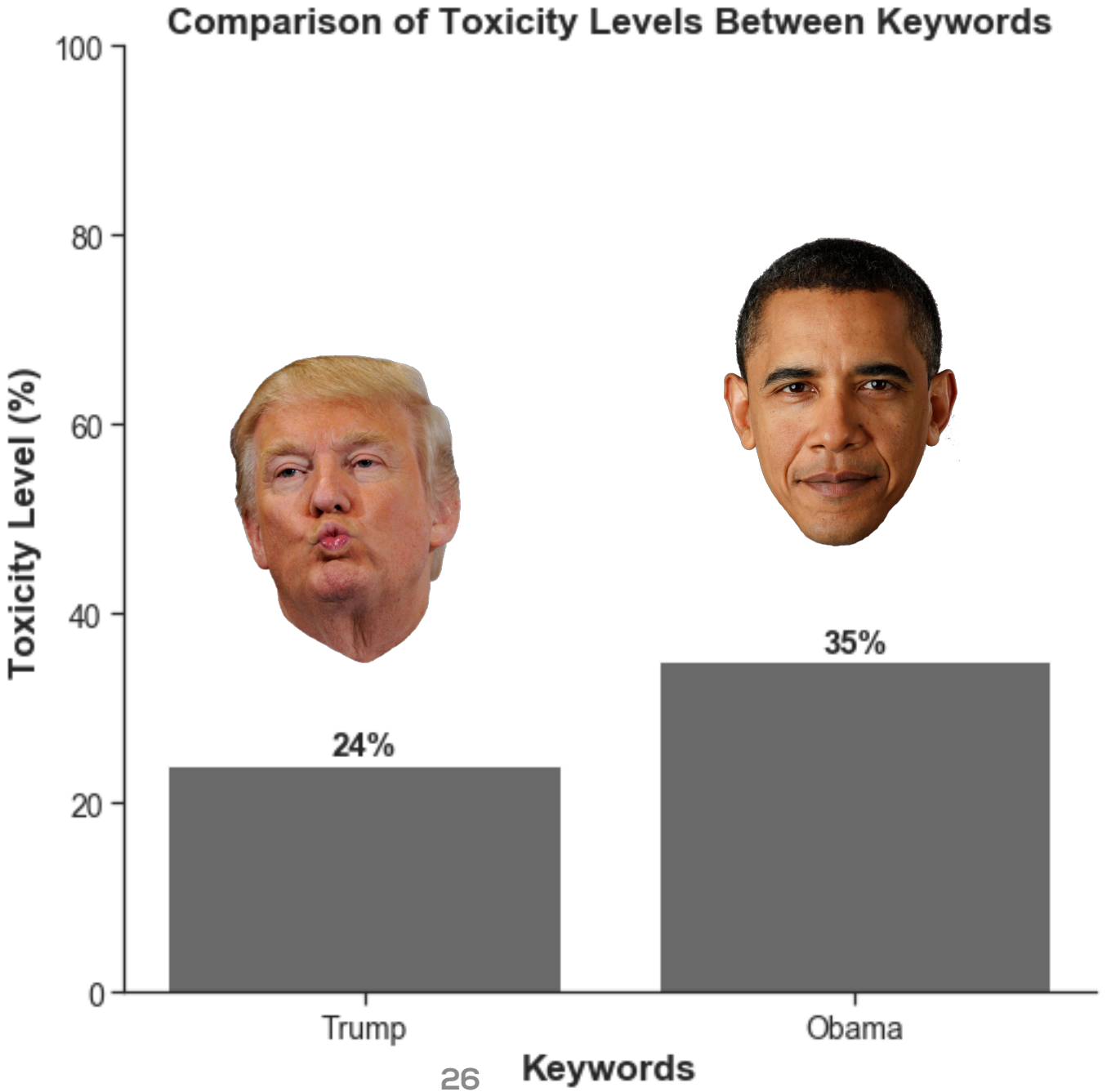

Precision-Recall curve: AP=0.74
Precision
Recall
0.0
0.2
0.4
0.6
0.8
1.0

# *Applications*

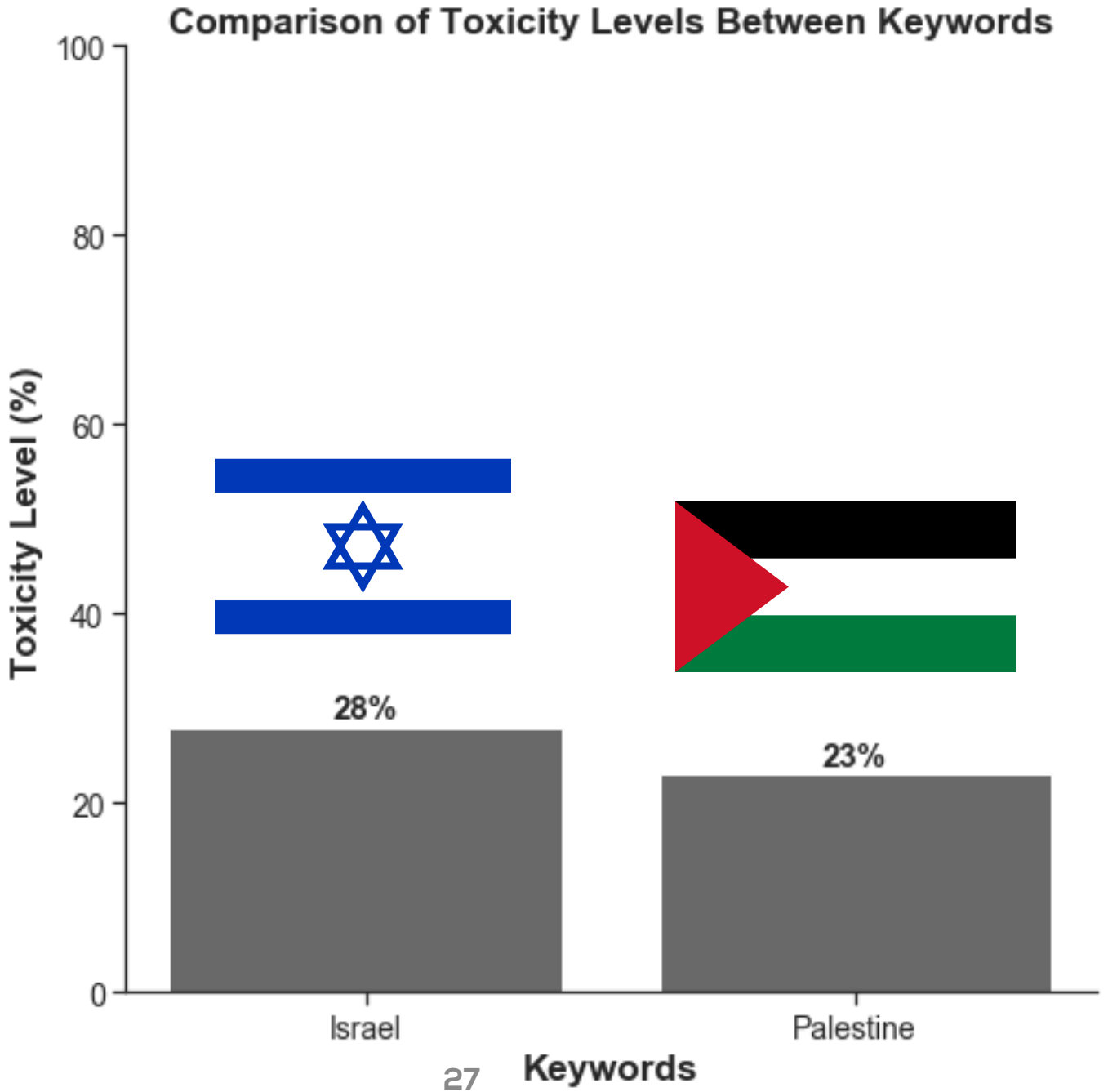
# *Applications*



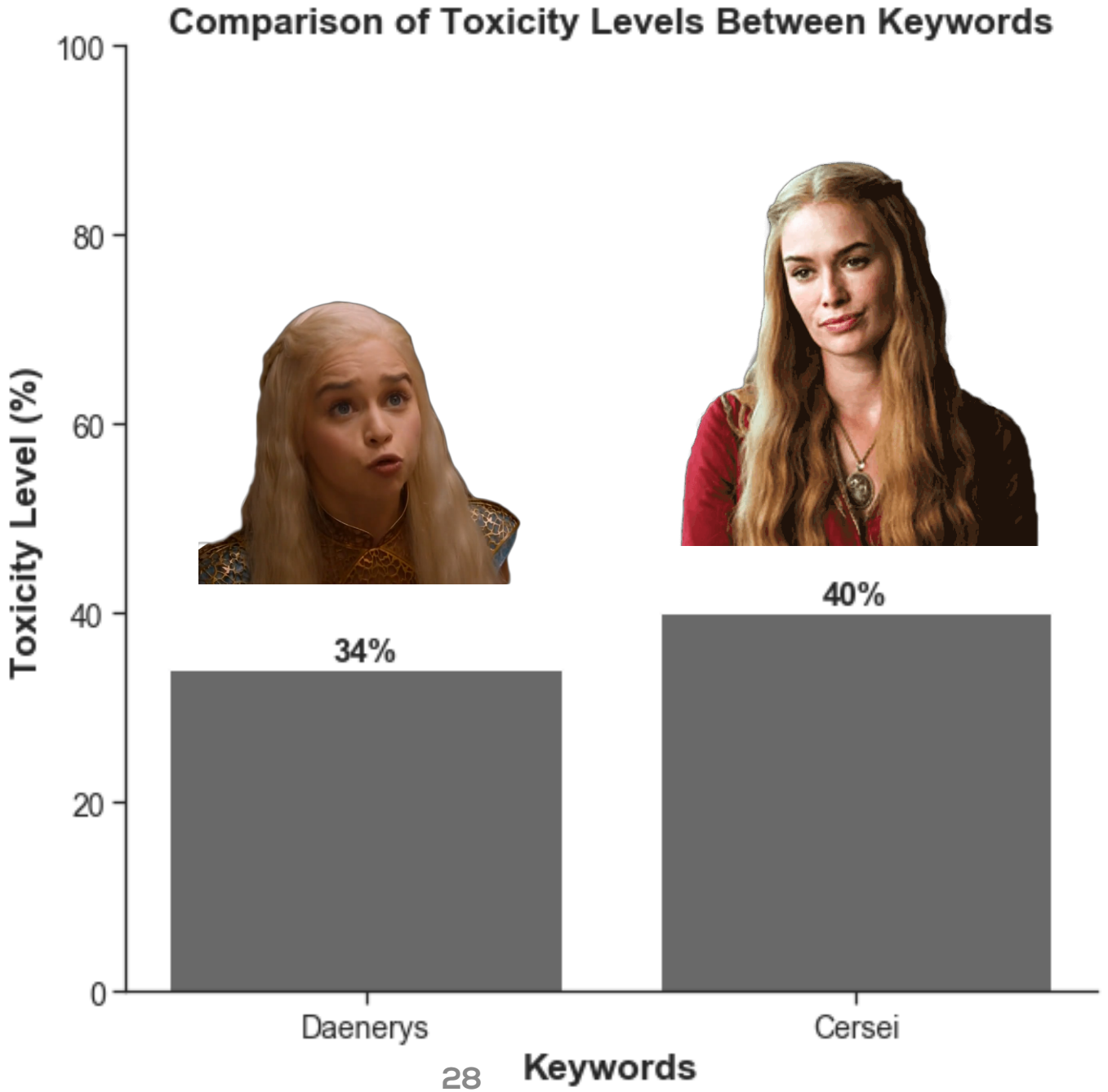
# *Applications*



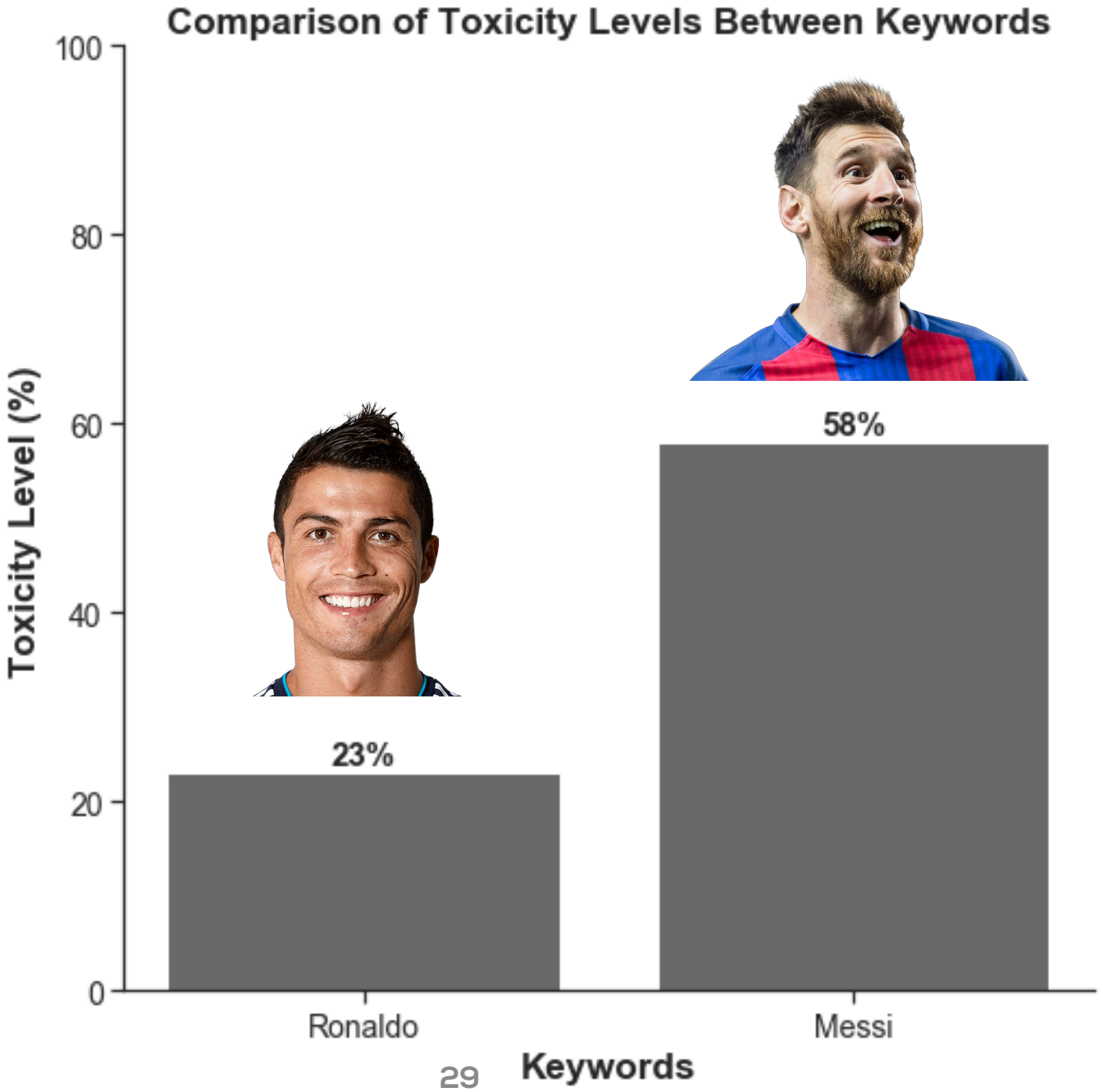
# Applications



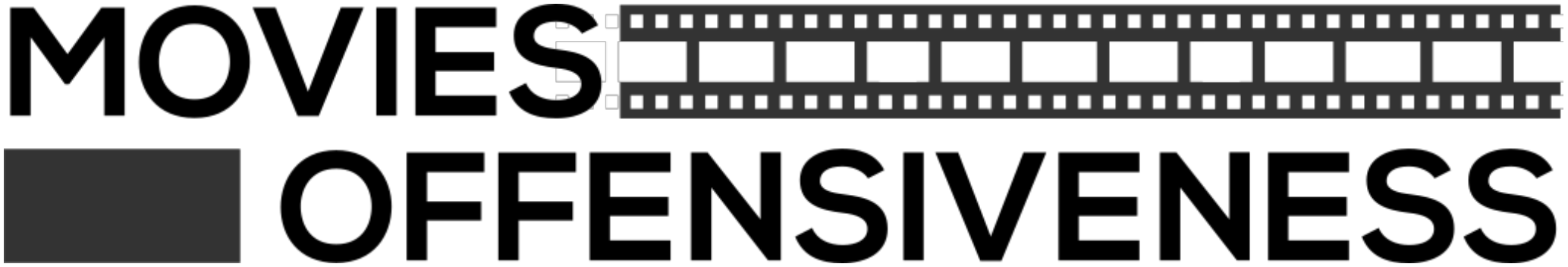
# *Applications*



# *Applications*



# *Applications*



IMSDb
The web's largest movie script resource!

**Search IMSDb**

Go!

**Alphabetical**

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| # | A | B | C | D | E | F | G | H |
| I | J | K | L | M | N | O | P | Q |
| R | S | T | U | V | W | X | Y | Z |

**Genre**

| | | |
|---|---|---|
| Action | Adventure | Animation |
| Comedy | Crime | Drama |
| Family | Fantasy | Film-Noir |
| Horror | Musical | Mystery |
| Romance | Sci-Fi | Short |
| Thriller | War | Western |

**Sponsor**

**TV Transcripts**

Futurama
Seinfeld
South Park
Stargate SG-1
Lost
The 4400

**International**

French scripts

**Movie Software**



## All Movie Scripts on IMSDb (A-Z)

10 Things I Hate About You (1997-11 Draft)
*Written by Karen McCullah Lutz,Kirsten Smith,William Shakespeare*

12 (Undated Draft)
*Written by Lawrence Bridges*

12 and Holding (2004-04 Draft)
*Written by Anthony Cipriano*

12 Monkeys (1994-06 Draft)
*Written by David Peoples,Janet Peoples*

12 Years a Slave (Undated Draft)
*Written by John Ridley*

127 Hours (Undated Draft)
*Written by Simon Beaufoy,Danny Boyle*

1492: Conquest of Paradise (1991-09 Draft)
*Written by Roslyne Bosch*

15 Minutes (Undated Draft)
*Written by John Hertzfield*

17 Again (2007-10 Draft)
*Written by Jason Filardi*

187 (1996-11 Draft)
*Written by Scott Yagemann*

2001: A Space Odyssey (1989-02 Draft)
*Written by Stanley Kubrick,Arhur C. Clarke*

2012 (2008-02 Second draft)
*Written by Roland Emmerich,Harald Kloser*

25th Hour (2001-04 Draft)
*Written by David Benioff*

30 Minutes or Less (2009-12 Draft)
*Written by Michael Diliberti,Matthew Sullivan*

42 (2012-07 Revised draft)
*Written by Brian Helgeland*

44 Inch Chest (Undated Draft)
*Written by Louis Mellis,David Scinto*

48 Hrs. (Undated Draft)
*Written by Steven E. De Souza,Walter Hill,Roger Spottiswoode,Larry Gross,Jeb Stuart*



Source: https://www.imsdb.com/all%20scripts/

"PULP FICTION"

By

Quentin Tarantino & Roger Avary

PULP [pulp] n.

1. A soft, moist, shapeless mass or matter.

2. A magazine or book containing lurid subject matter and being characteristically printed on rough, unfinished paper.

American Heritage Dictionary: New College Edition

**INT. COFFEE SHOP - MORNING**

A normal Denny's, Spires-like coffee shop in Los Angeles. It's about 9:00 in the morning. While the place isn't jammed, there's a healthy number of people drinking coffee, munching on bacon and eating eggs.

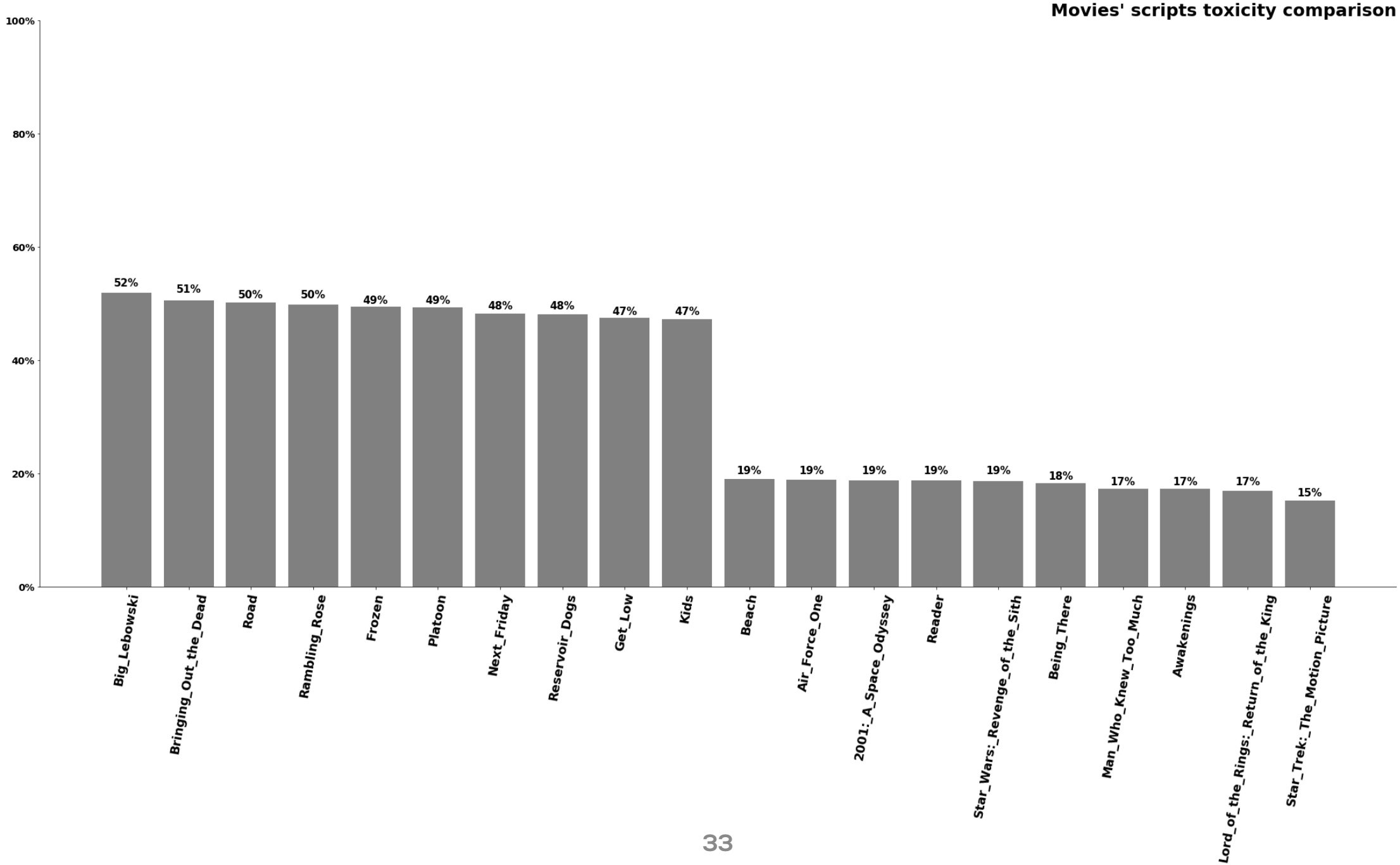
Two of these people are a YOUNG MAN and a YOUNG WOMAN. The Young Man has a slight working-class English accent and, like his fellow countryman, smokes cigarettes like they're going out of style.

It is impossible to tell where the Young Woman is from or how old she is; everything she does contradicts something she did. The boy and girl sit in a booth. Their dialogue is to be said in a rapid pace "HIS GIRL FRIDAY" fashion.

**YOUNG MAN**
No, forget it, it's too risky. I'm through doin' that shit.



Source: https://www.imsdb.com/all%20scripts/


Movies' scripts toxicity comparison
100%
80%
60%
40%
20%
0%
52%
51%
50%
50%
49%
49%
48%
48%
47%
47%
19%
19%
19%
19%
19%
18%
17%
17%
17%
15%
Big_Lebowski
Bringing_Out_the_Dead
Road
Rambling_Rose
Frozen
Platoon
Next_Friday
Reservoir_Dogs
Get_Low
Kids
Beach
Air_Force_One
2001:_A_Space_Odyssey
Reader
Star_Wars:_Revenge_of_the_Sith
Being_There
Man_Who_Knew_Too_Much
Awakenings
Lord_of_the_Rings:_Return_of_the_King
Star_Trek:_The_Motion_Picture

# *Thank you*


| | |
|---|---|
| *Rita Franco* | *20180081* |
| *Rodrigo Umbelino* | *20180060* |
| *Vitor Manita* | *20180054* |