PROJECT REPORT : DSO 560 TEXT ANALYTICS

TOXICITY PREDICTION & TOPIC MODELING

TEAM MEMBERS:
ABHINAV CHADHA
ARPAN SHRIVASTAVA
SHREE VIDYA RAVI KUMAR
VAIBHAV JHA

PROBLEM STATEMENT

- A tweet or a comment can have adverse affects on a person's national or international image.
- Geopolitics and political figures play an important role in world economy
- Public figures have a huge fan base and their actions can influence public in several ways

- **By understanding the toxic comments and classifying them, can we reduce the bias in misinformation and spreading them across globally?**

BUSINESS USE CASES

Use case 1:

An unverified information to a tweet containing toxic sentiment can be used to inform users that they should be careful consuming that piece of information. Moreover, it could be used in conjunction with the 'Report' button that users have access to, in order to validate that the tweet/comment is hurtful and delete it.

Use case 2:

With the help of topic modeling, we can determine the label for a tweet/comment and if the label suggests some adverse action and has a high toxicity score, the relevant authorities to handle a potential situation could be alerted. Having data on how many such tweets in the past has resulted in a bad action would be more helpful.

AGENDA

Toxicity Prediction:

We perform toxicity prediction on comment text to classify it as a toxic or non-toxic. If we classify the comment as toxic, we next try to predict the level of toxicity, on high medium and low scale

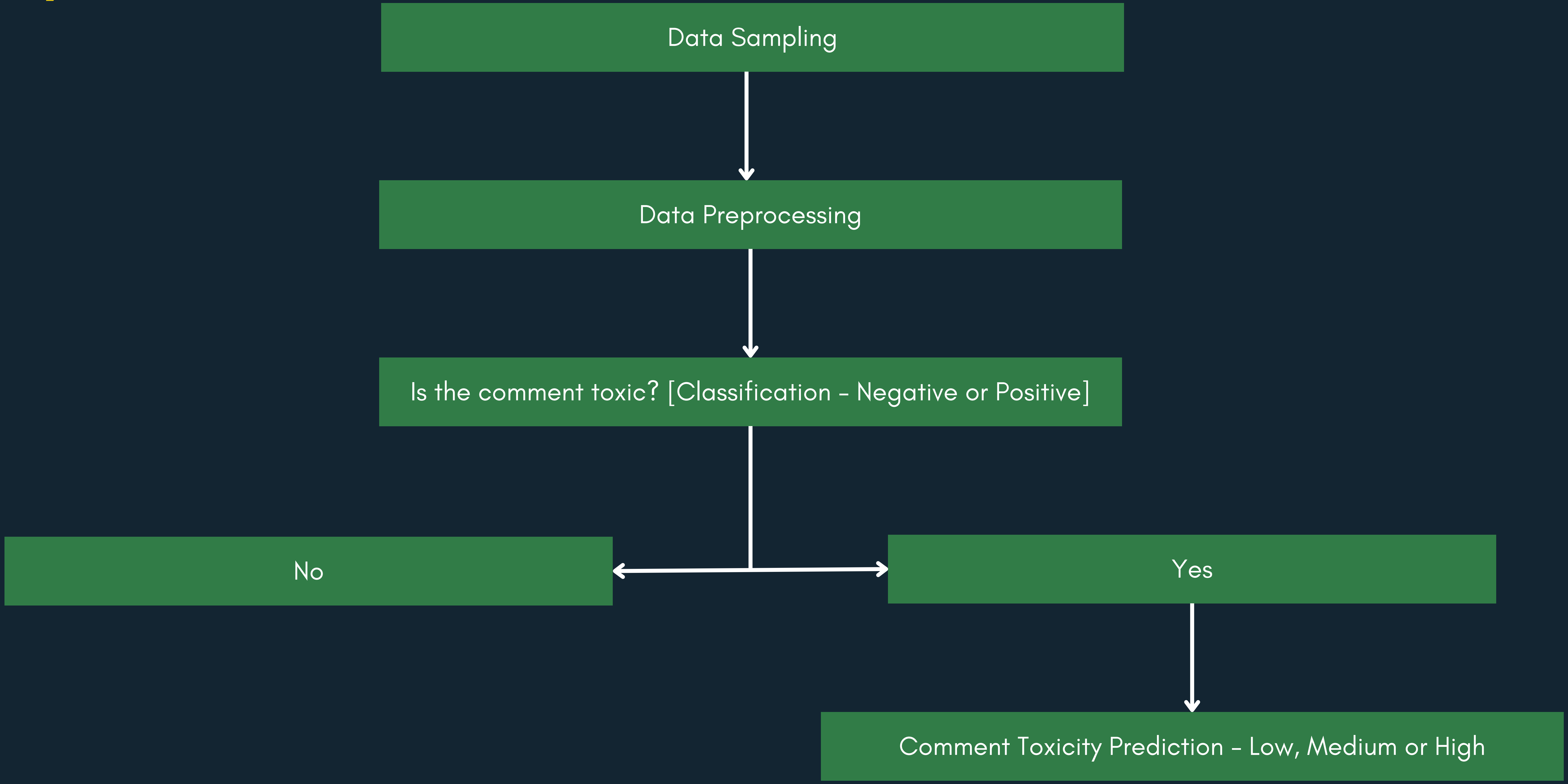
Topic Modeling:

Based on the level of toxicity, we again sample our data and conduct initial preprocessing and perform topic modeling using NMF model for dimensionality reduction to predict the toxic topics and how each comment can be associated with the topic

1. TOXICITY PREDICTION

MODEL STRUCTURE

Toxicity Prediction:



Data Sampling

Data Preprocessing

Is the comment toxic? [Classification - Negative or Positive]

No

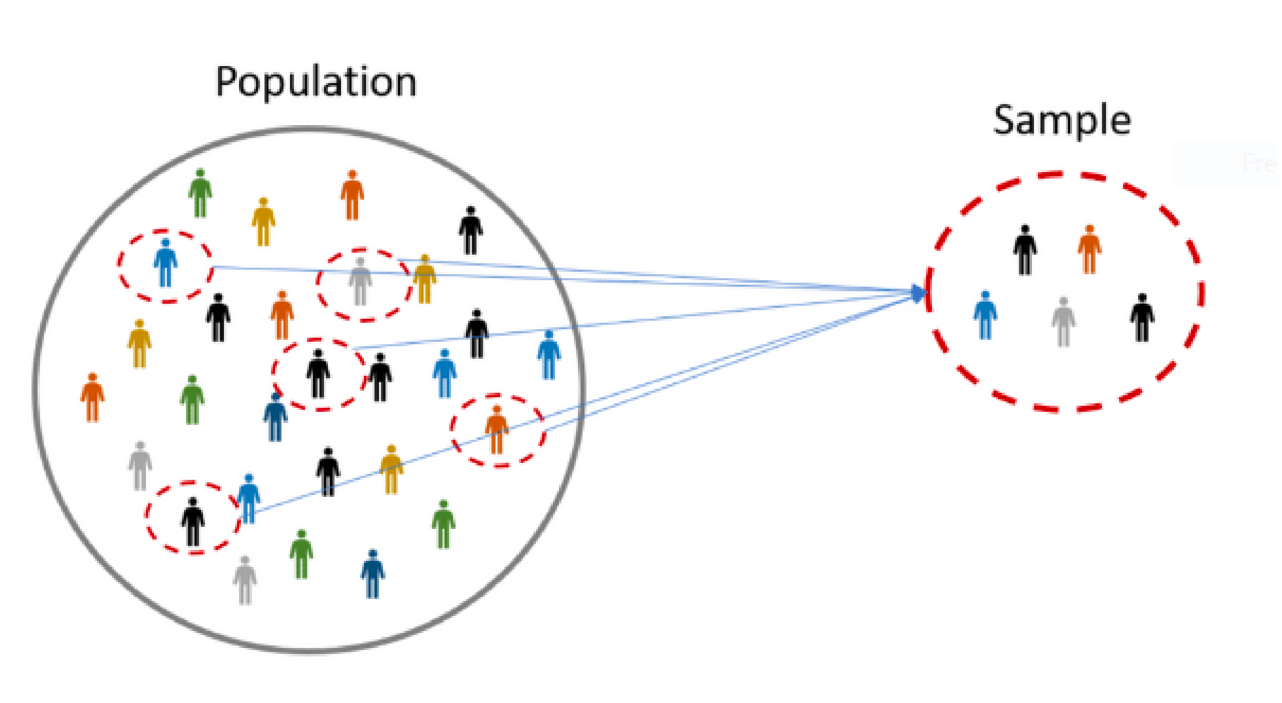
Yes

Comment Toxicity Prediction - Low, Medium or High

DATA DOWNLOAD AND DATA SAMPLING

Data used: https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data

- The dataset contains approximately 2 million rows of training and test dataset
- The features in the dataset are comment id, comment, toxicity (between 0 to 1), and certain metadata
- This metadata contains lots of missing information and for the sake of simplicity we will be not be working with any metadata



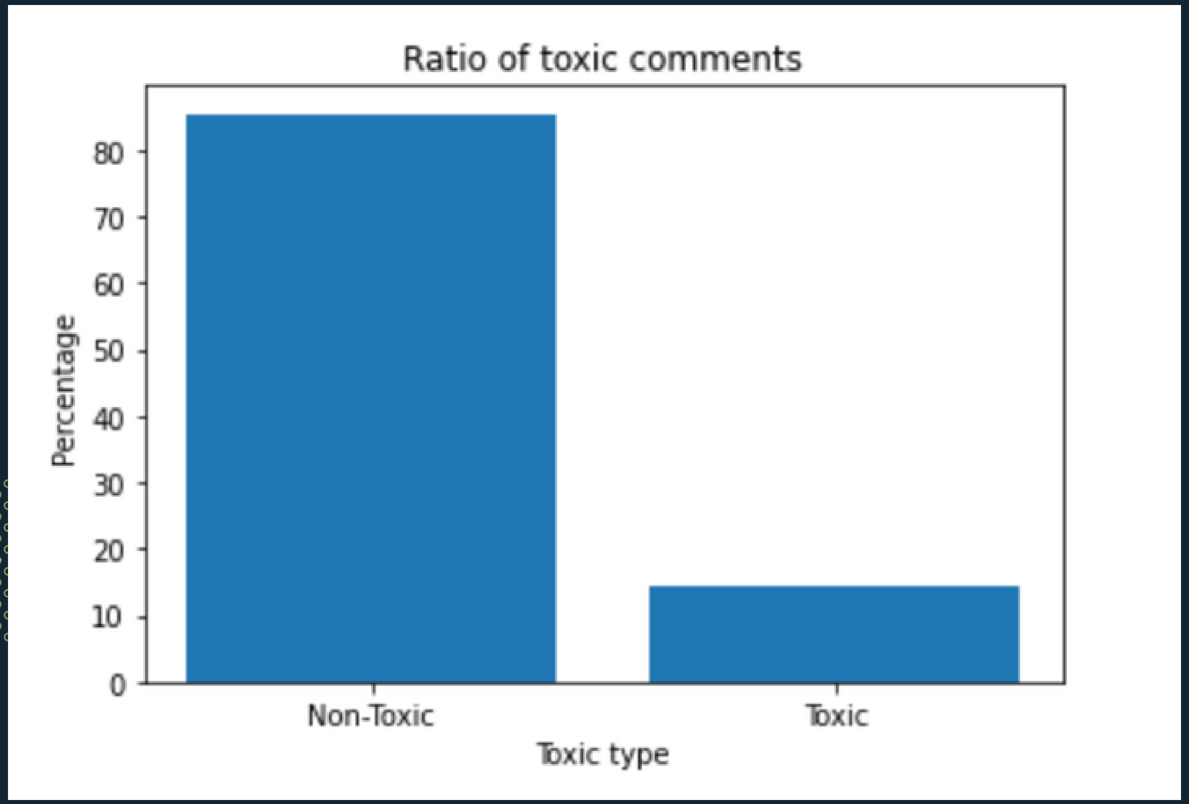
Data Sampling:

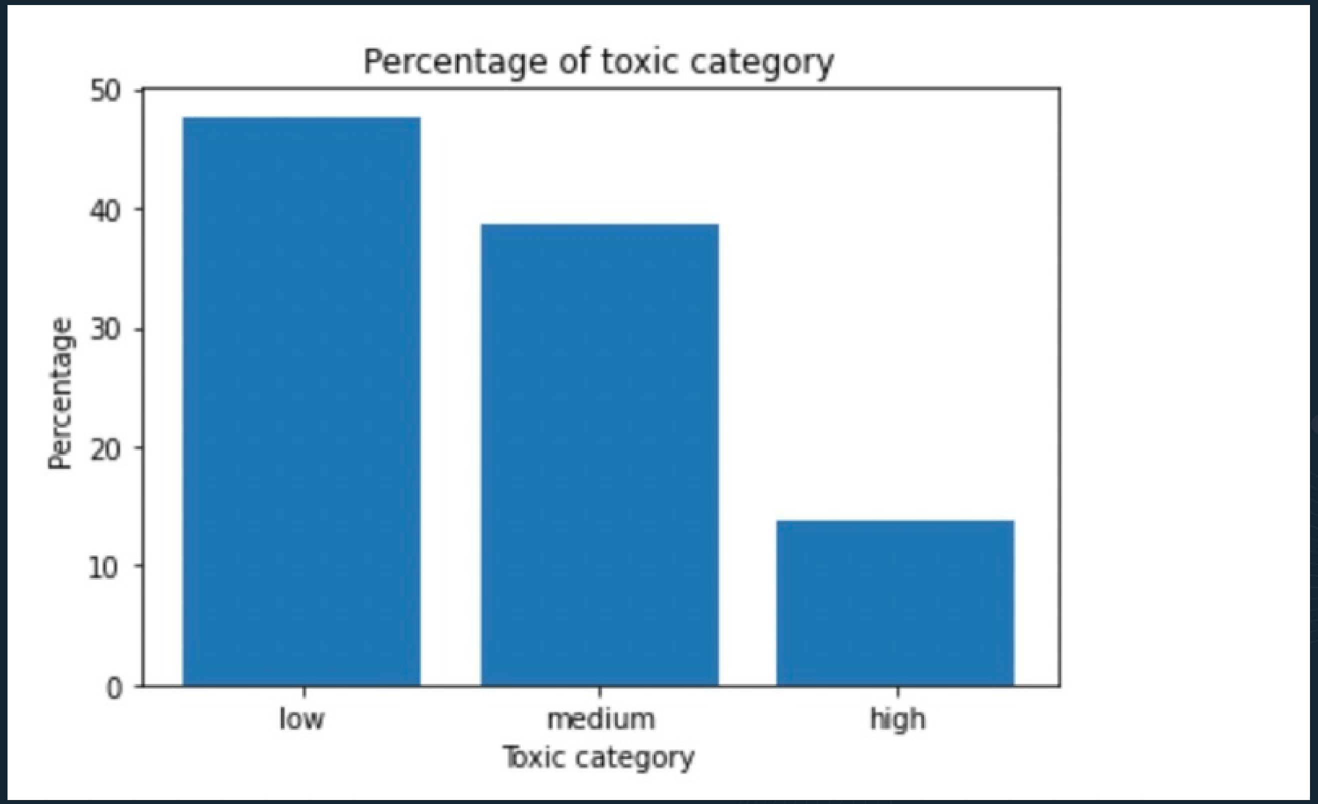
- Since, text processing using "spacy" and "textacy" takes huge computational time, the data is sampled to represent the complete picture
- The sampling is done on a 1:100 scale, therefore, for our analysis we will be using close to ~20K comments in a similar distribution as the original data

EXPLORATORY DATA ANALYSIS

- We have created bins in our dataset on the basis of toxicity to see the distribution of comments, and we find there is a huge class imbalance between toxic and non-toxic comments as expected. We create the following bins
  - Non-Toxic Comments [0 to 0.25]
  - Toxic Comments [0.25 to 1]

- For the toxic comments, it is important to understand the toxicity level and take necessary actions of content removal/address. therefore, we further divide these Toxic comments into three buckets
  - Low Toxic [0.25 to 0.5]
  - Medium Toxic [0.5 to 0.75]
  - High Toxic [0.75 to 1]





DATA PREPROCESSING

- We use Spacy, Regex and Texacy to do text preprocessing

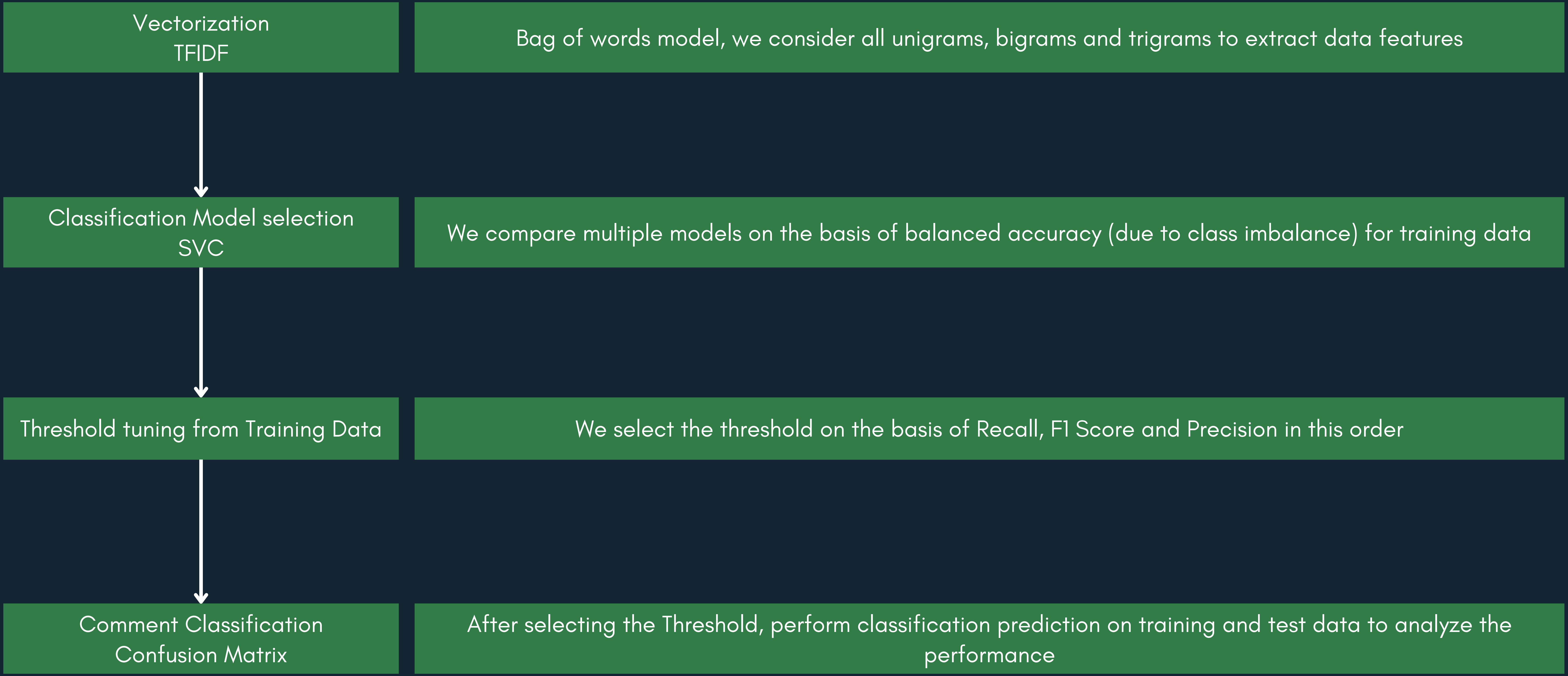
Text Deconstruction (e.g., won't to will not)

URL, Numbers, Currency, Hashtags, Email, Emoji and Punctuation

Named Entity Removal and Text Lemmatization

- Why is this required?
    - Most of the text data obtained is user-generated. As a result, it contains high noise and needs to be removed. preprocessing helps in obtaining clean data
    - Data such as HTML tags, URL, Currency etc does not add value to analyze a text either for its sentiment or for classification and can be brought down to a simple form using regex or in-built libraries
    - Once the text is preprocessed, we can use the data to solve various business cases such as topic modelling, sentiment analysis, text classification and translation etc.

COMMENT CLASSIFICATION



Vectorization
TFIDF

Bag of words model, we consider all unigrams, bigrams and trigrams to extract data features

Classification Model selection
SVC

We compare multiple models on the basis of balanced accuracy (due to class imbalance) for training data

Threshold tuning from Training Data

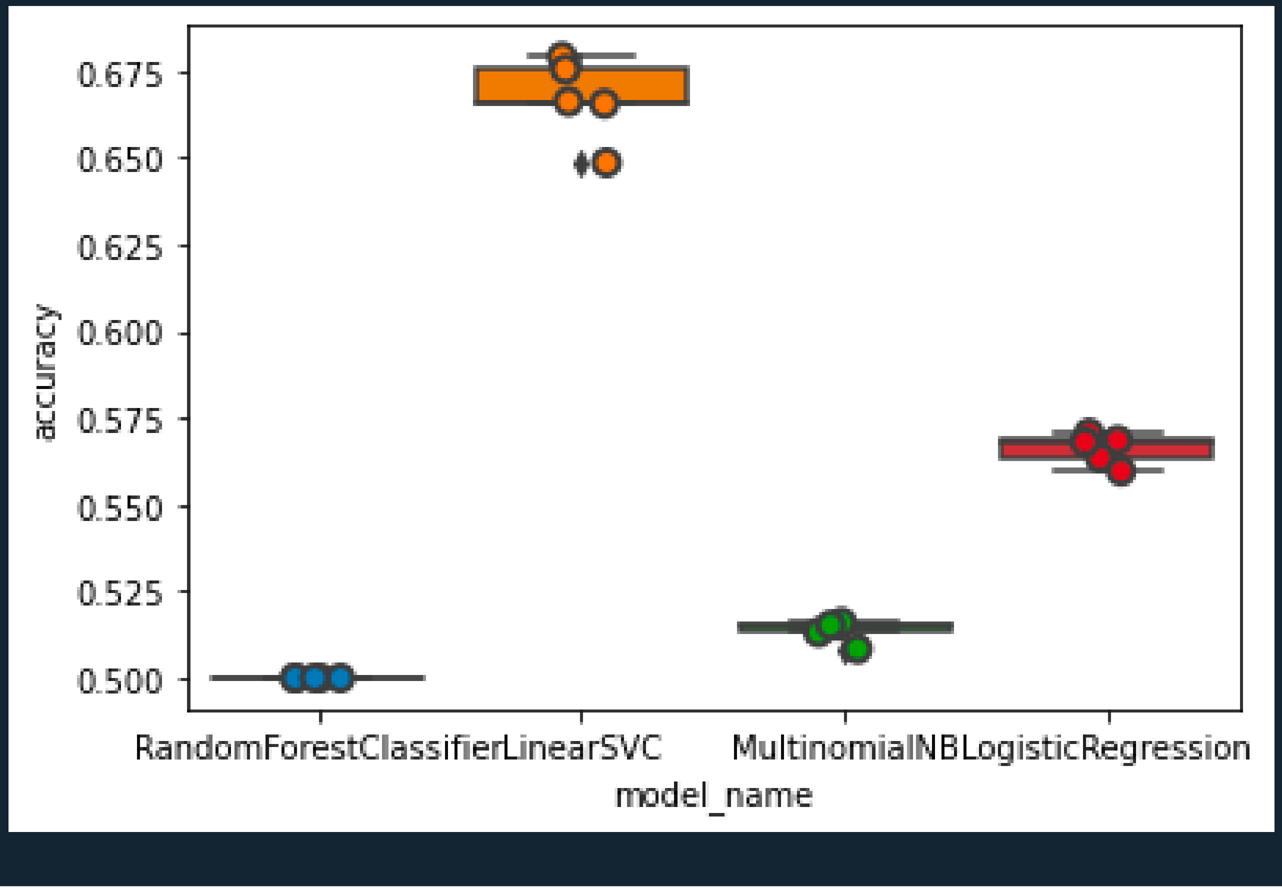
We select the threshold on the basis of Recall, F1 Score and Precision in this order

Comment Classification
Confusion Matrix

After selecting the Threshold, perform classification prediction on training and test data to analyze the performance

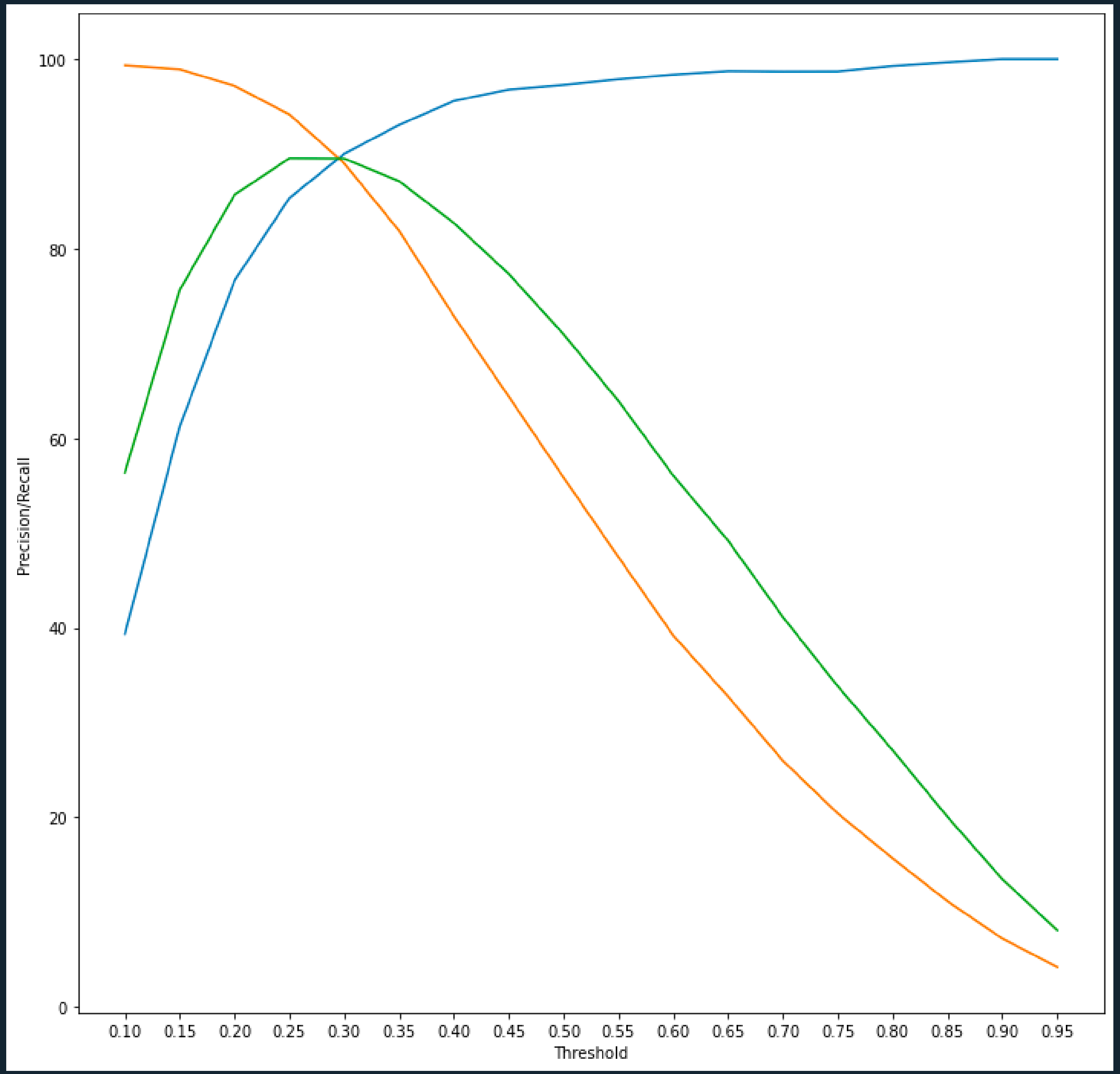
MODEL PERFORMANCE - PARAMETER TUNING

Classification Model selection
SVC



MODEL	BALANCED ACCURACY Mean
Random Forest Classifier	**0.5**
SVC	**0.67**
MultinomialNB	0.51
Logistic Regression	0.56

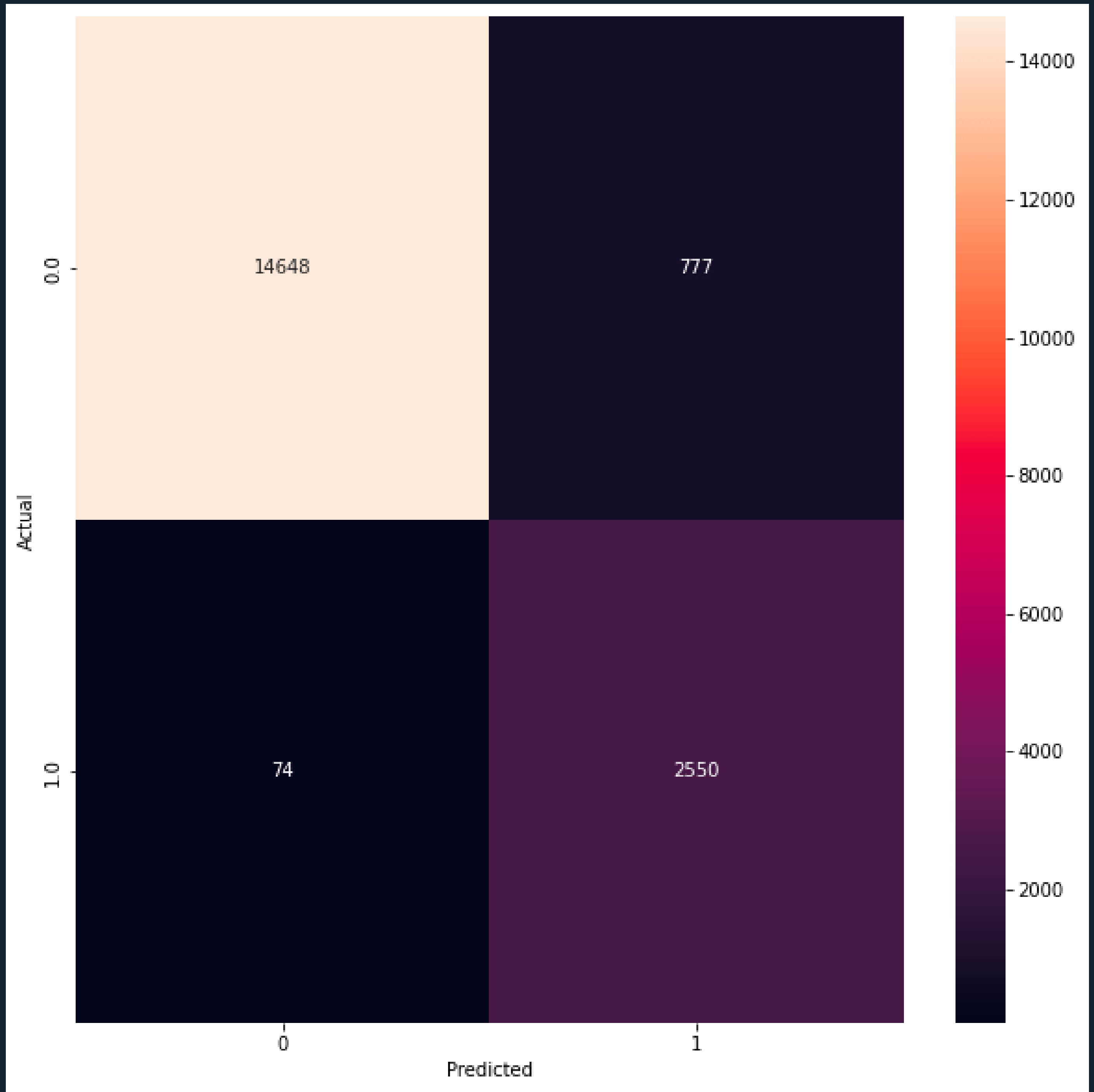
Threshold tuning from Training Data



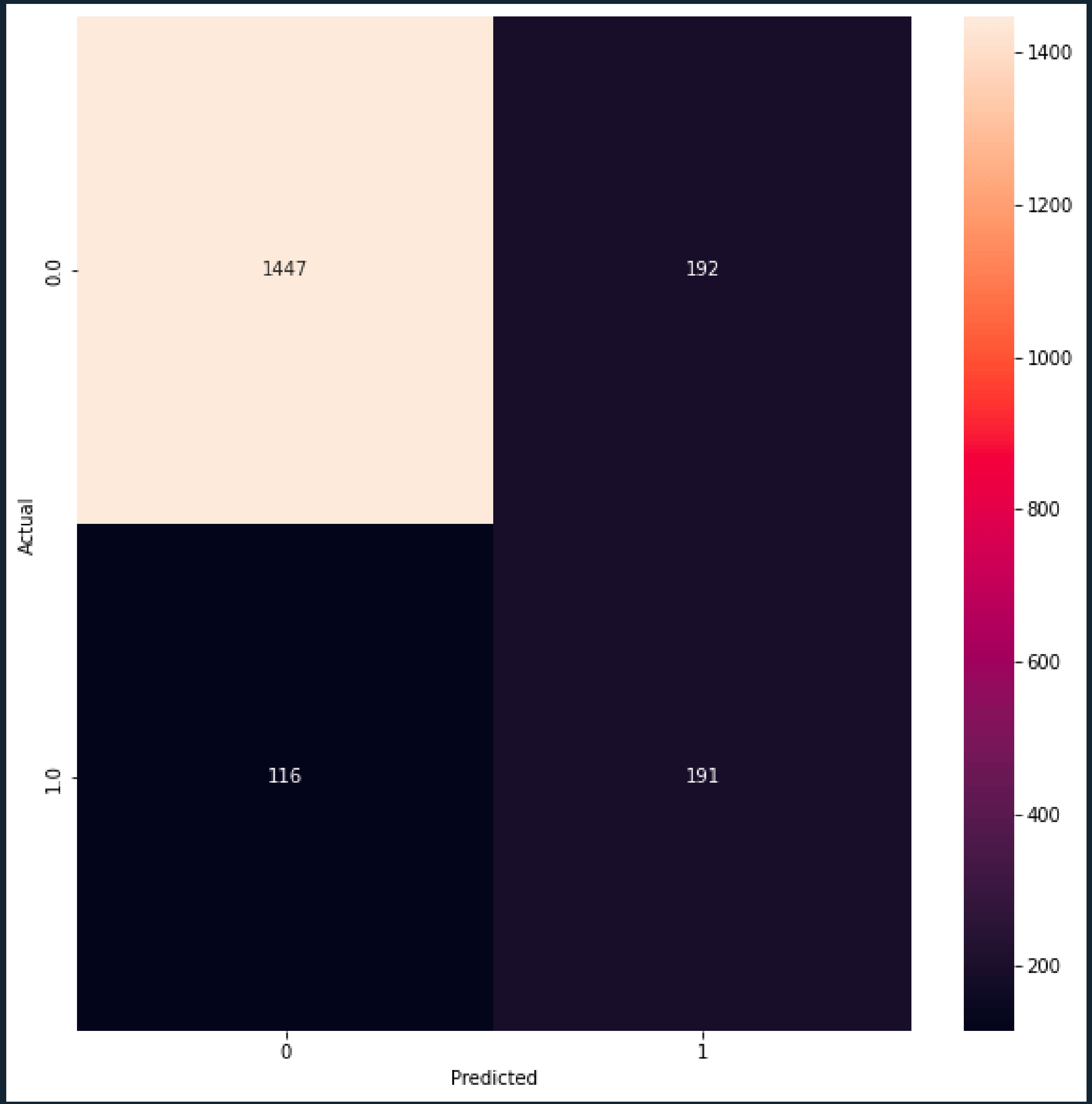
Selecting threshold as 0.2, since we need better recall than precision

MODEL PERFORMANCE

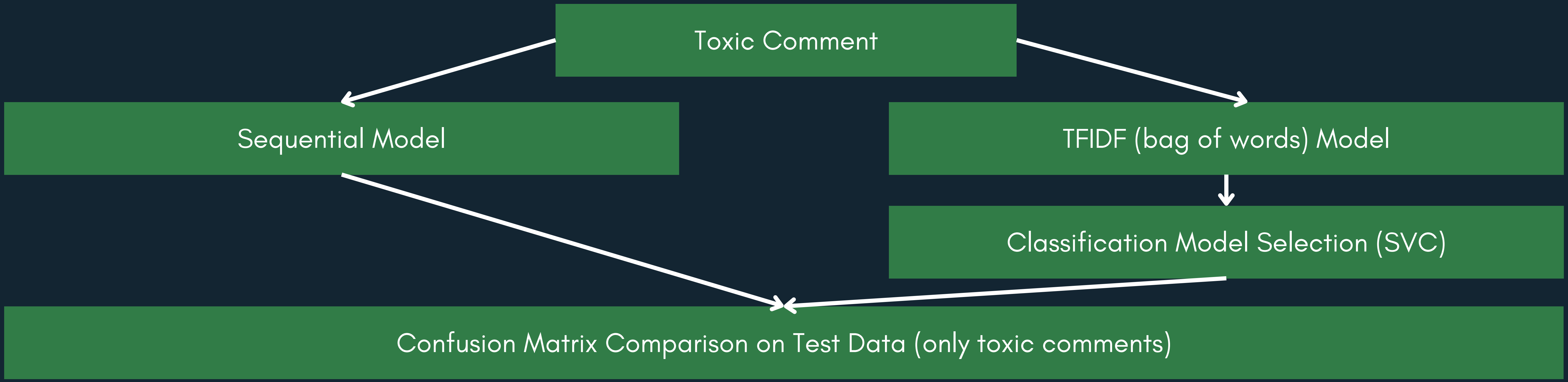
Model Performance on Training Data



Model Performance on Test Data



TOXICITY LEVEL CLASSIFICATION



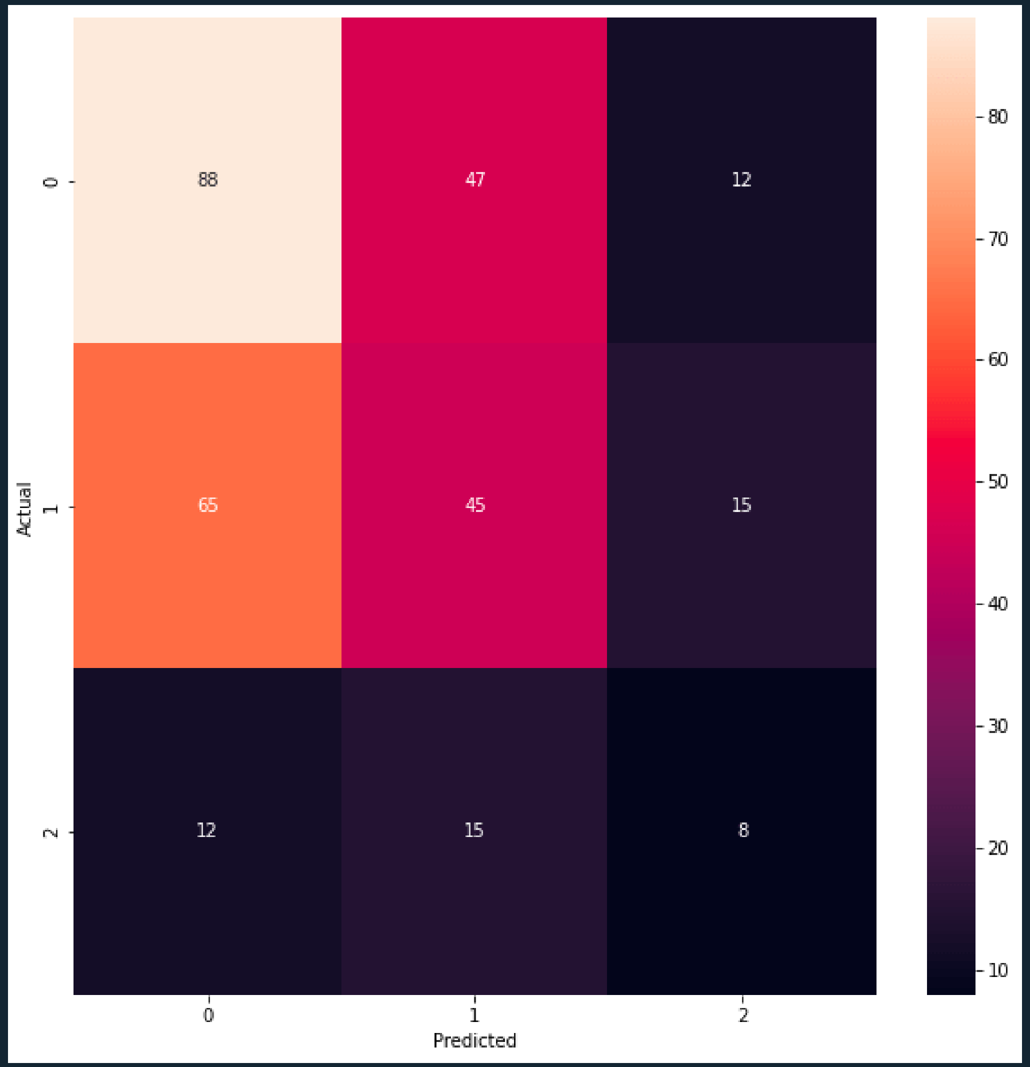
Toxic Comment

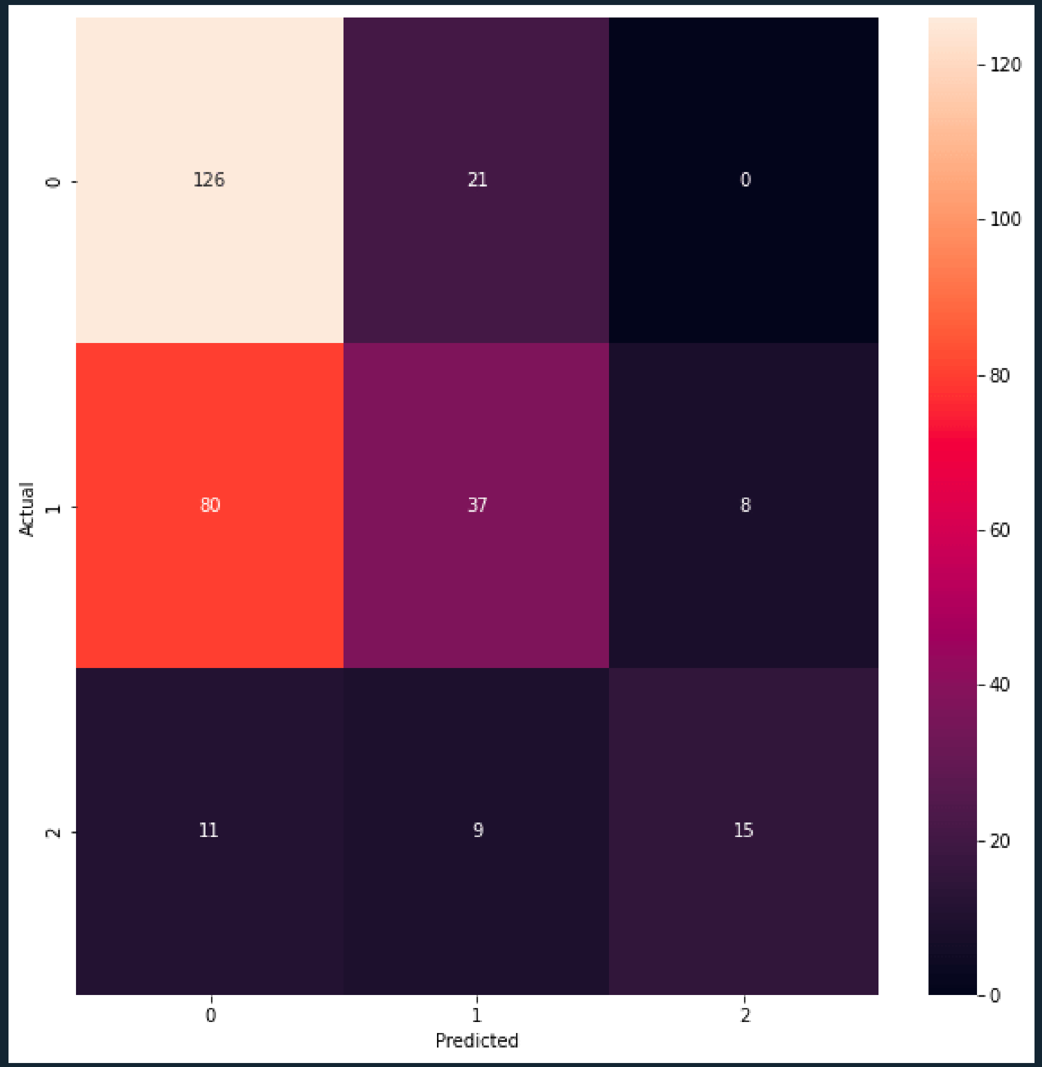
Sequential Model

TFIDF (bag of words) Model

Classification Model Selection (SVC)

Confusion Matrix Comparison on Test Data (only toxic comments)





MODEL INTERPRETATION

Toxic Comment Classification:

We selected TFIDF to perform classification. The use of 'balanced accuracy' as scoring helps us deal with class imbalance.
It is important to tune the parameter - the threshold to classify between Toxic and Non-Toxic due to the huge class imbalance. We look for a threshold that has a better threshold, because it will be more costly to have False Negatives than False positives

The model in itself does a good job in classifying Toxic and Non-toxic. And can definitely perform better if we further include customer review, meta data into our analysis

Toxicity Level Classification:

Since we have fewer toxic comments than non-toxic ones, we try to employ the Sequential model to predict the comment's toxicity level. However, due to the GPU memory shortage, we are not able to enhance model performance and therefore stick to TFIDF for our classification

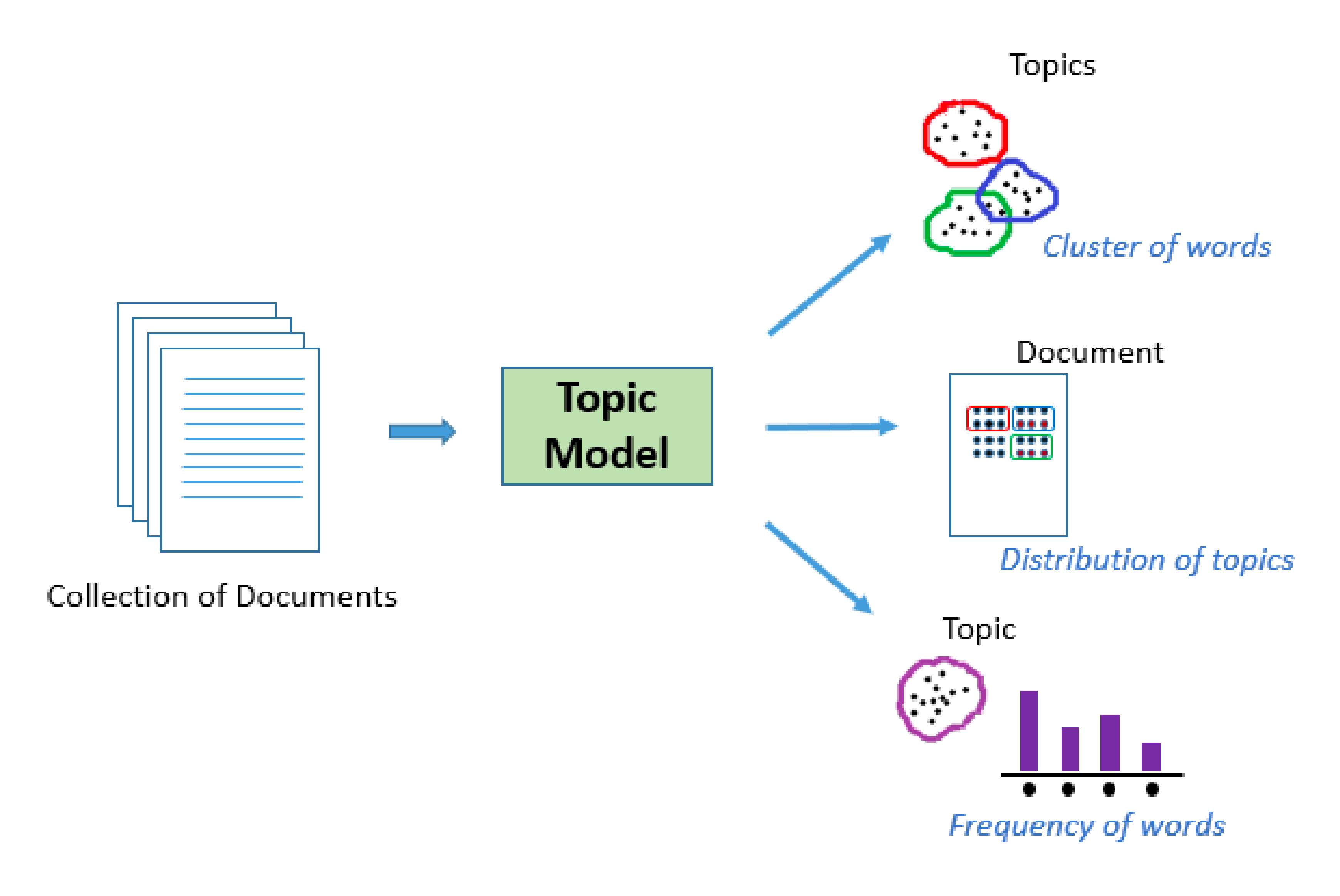
The model can further be hyper-tuned to perform better classification and along with customer report review, we can create an even better classifier

2. TOPIC MODELING

MOTIVATION FOR TOPIC MODELLING

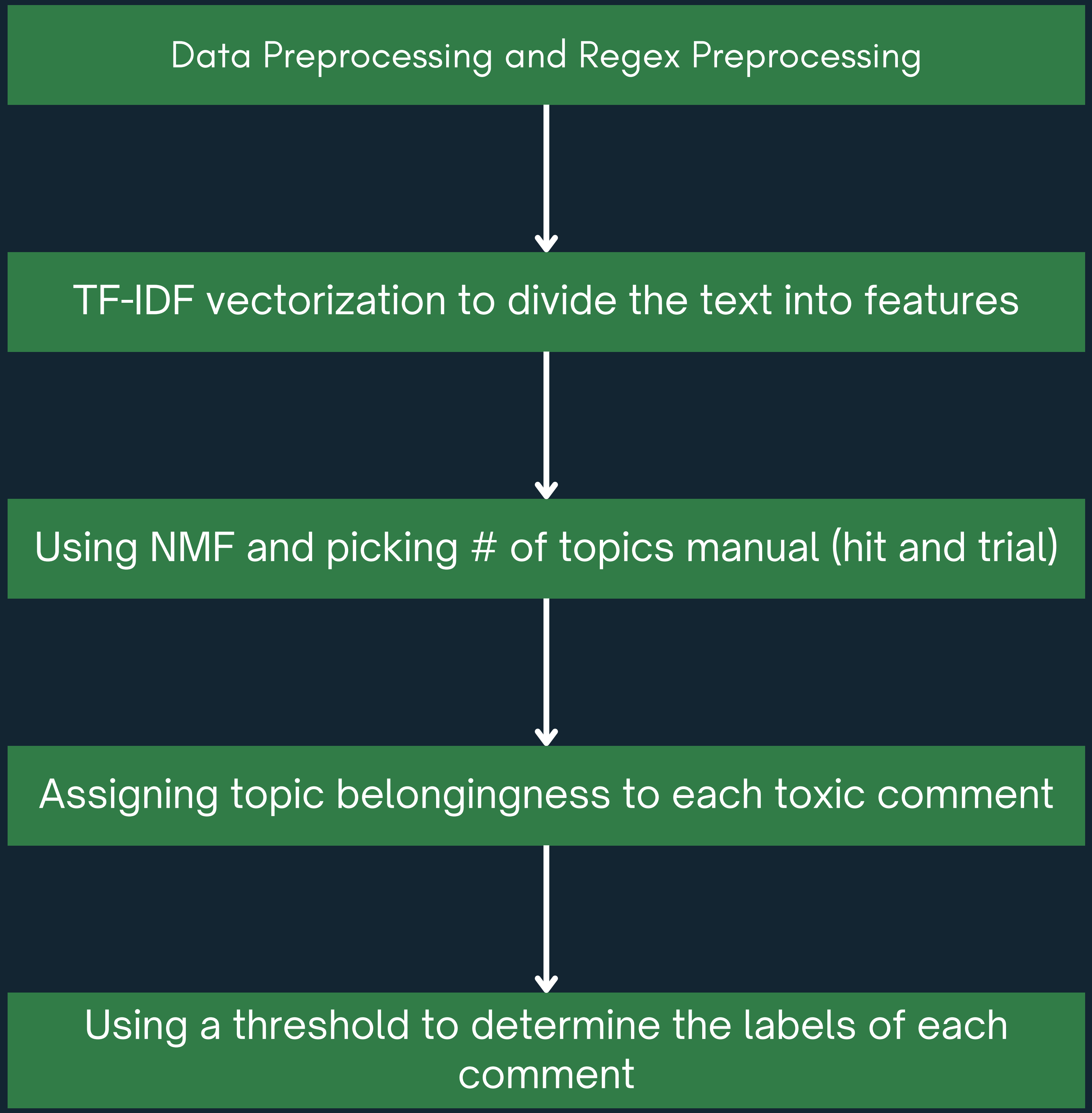
We wanted to see what are the dominating ideas and subjects that people are mentioning in their toxic comments.

The best way to determine this was through topic modeling. Therefore, after text preprocessing and data manipulation, we built a classification model which determined if the comment was toxic or not plus its toxicity level. Naturally, the next step was to determine what actually was the content of these comments. And if we can assign label to these comments



MODEL STRUCTURE

Topic Modeling:



Data Preprocessing and Regex Preprocessing

TF-IDF vectorization to divide the text into features

Using NMF and picking # of topics manual (hit and trial)

Assigning topic belongingness to each toxic comment

Using a threshold to determine the labels of each comment

DATA PREPROCESSING AND REGEX PREPROCESSING

- We use Spacy, Regex and Texacy to do text preprocessing

Text Deconstruction (e.g., won't to will not)

URL, Numbers, Currency, Hashtags, Email, Emoji and Punctuation

Named Entity Removal and Text Lemmatization

- We add an additional layer of regex cleaning to remove generic terms like URL, EMAIL, and NUMBER, etc. from our document corpus

TFIDF VECTORIZATION

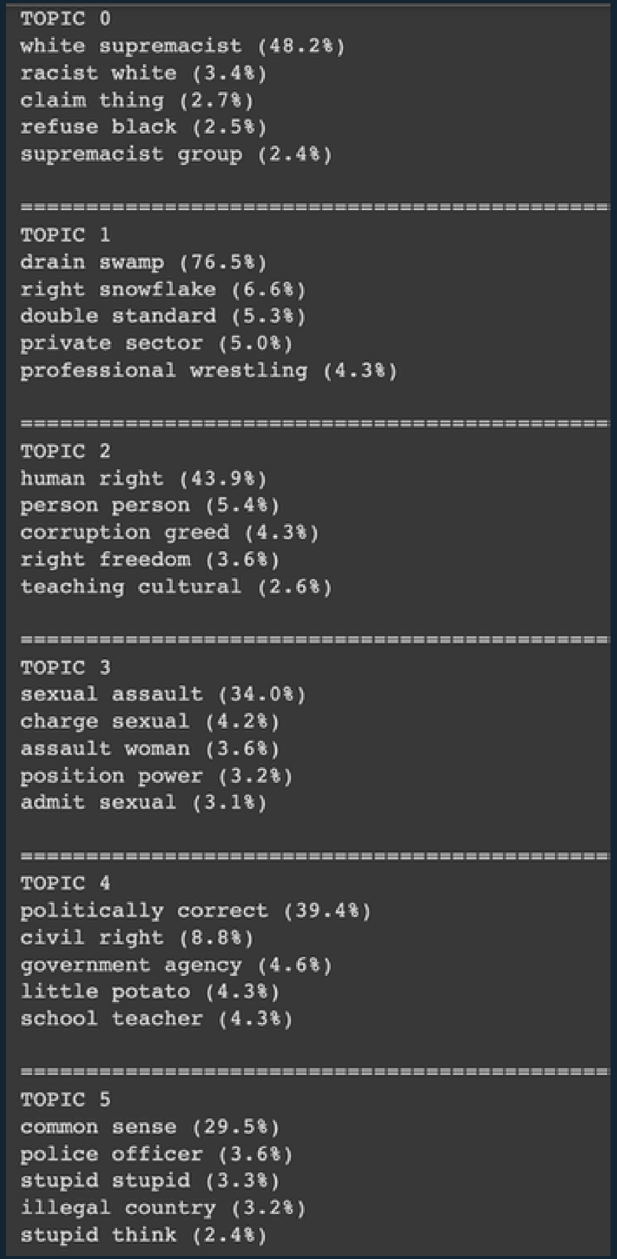
- We use TFIDF Vectorizer to vectorize our data using the following parameters
  - Ngrams = 2, 3, and 4 i.e., Bigrams, Trigrams and Quadgrams
  - min_df = 2 i.e., the token needs to be present in at least two documents
  - token_pattern = r'\b[A-Za-z]{5,}\b' , i.e. the pattern should be an english word of length 5 and above

- Why these parameters?
  - We tried using unigrams, but these are not good enough to distinguish between two different topics
  - Why length 5, certain words which are not considered stop words bias the data due to their regular occurrence, we needed some clean and clear indicators of each topics

TOPIC COUNT SELECTION

- We look at different topic counts to see any clear distinction that helps us label these documents on the basis of any specific indicators
- We start with n = 6 and decrease the count if there is an overlap or an unclear category

N = 6



```
TOPIC 0
white supremacist (48.2%)
racist white (3.4%)
claim thing (2.7%)
refuse black (2.5%)
supremacist group (2.4%)

=============================================
TOPIC 1
drain swamp (76.5%)
right snowflake (6.6%)
double standard (5.3%)
private sector (5.0%)
professional wrestling (4.3%)

=============================================
TOPIC 2
human right (43.9%)
person person (5.4%)
corruption greed (4.3%)
right freedom (3.6%)
teaching cultural (2.6%)

=============================================
TOPIC 3
sexual assault (34.0%)
charge sexual (4.2%)
assault woman (3.6%)
position power (3.2%)
admit sexual (3.1%)

=============================================
TOPIC 4
politically correct (39.4%)
civil right (8.8%)
government agency (4.6%)
little potato (4.3%)
school teacher (4.3%)

=============================================
TOPIC 5
common sense (29.5%)
police officer (3.6%)
stupid stupid (3.3%)
illegal country (3.2%)
stupid think (2.4%)
```

Racism

Politics

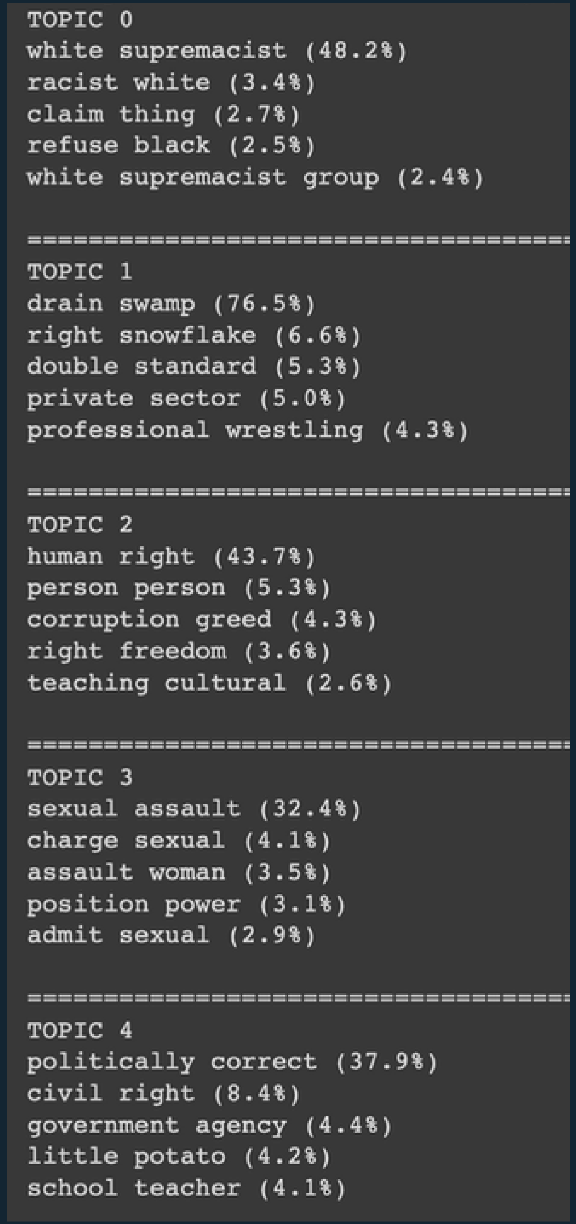
Human Rights

Sexual Abuse

Politics

Unclear

N = 5



```
TOPIC 0
white supremacist (48.2%)
racist white (3.4%)
claim thing (2.7%)
refuse black (2.5%)
white supremacist group (2.4%)

====================================
TOPIC 1
drain swamp (76.5%)
right snowflake (6.6%)
double standard (5.3%)
private sector (5.0%)
professional wrestling (4.3%)

====================================
TOPIC 2
human right (43.7%)
person person (5.3%)
corruption greed (4.3%)
right freedom (3.6%)
teaching cultural (2.6%)

====================================
TOPIC 3
sexual assault (32.4%)
charge sexual (4.1%)
assault woman (3.5%)
position power (3.1%)
admit sexual (2.9%)

====================================
TOPIC 4
politically correct (37.9%)
civil right (8.4%)
government agency (4.4%)
little potato (4.2%)
school teacher (4.1%)
```

Racism

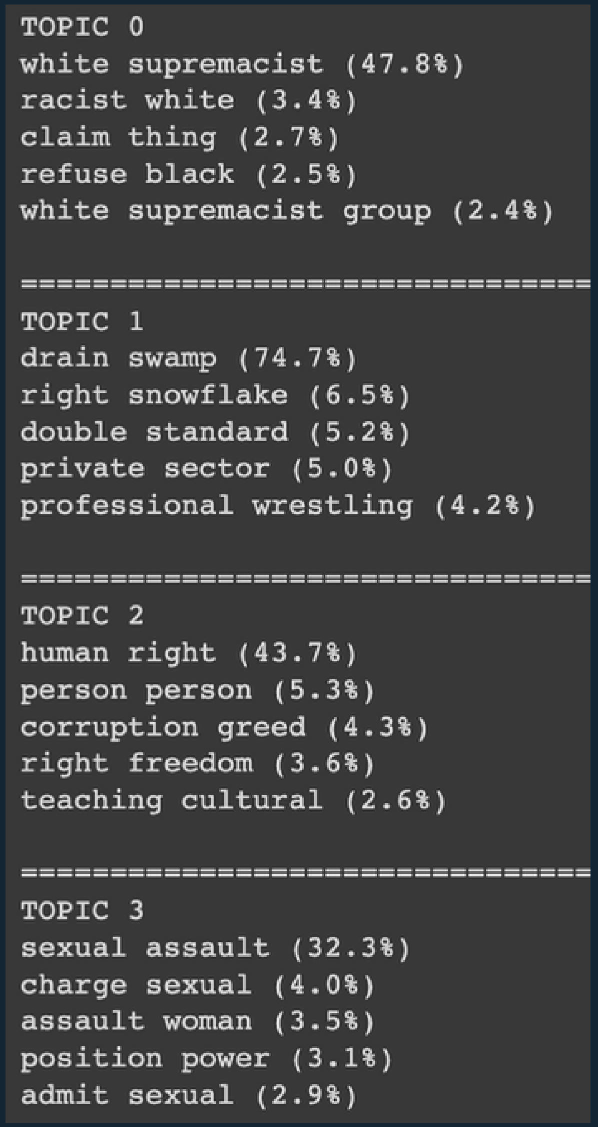
Politics

Human Rights

Sexual Abuse

Politics

N = 4



```
TOPIC 0
white supremacist (47.8%)
racist white (3.4%)
claim thing (2.7%)
refuse black (2.5%)
white supremacist group (2.4%)

================================
TOPIC 1
drain swamp (74.7%)
right snowflake (6.5%)
double standard (5.2%)
private sector (5.0%)
professional wrestling (4.2%)

================================
TOPIC 2
human right (43.7%)
person person (5.3%)
corruption greed (4.3%)
right freedom (3.6%)
teaching cultural (2.6%)

================================
TOPIC 3
sexual assault (32.3%)
charge sexual (4.0%)
assault woman (3.5%)
position power (3.1%)
admit sexual (2.9%)
```

Racism

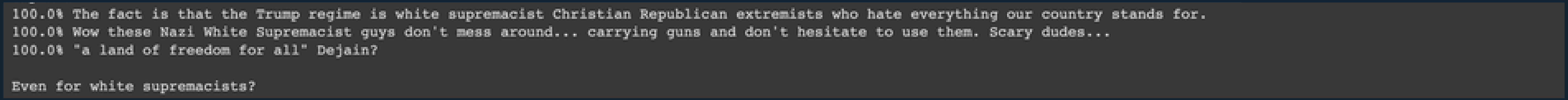
Politics

Human Rights

Sexual Abuse

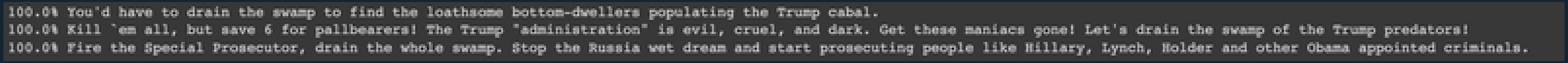
TOP COMMENTS FOR EACH TOPIC CATEGORY

Racism



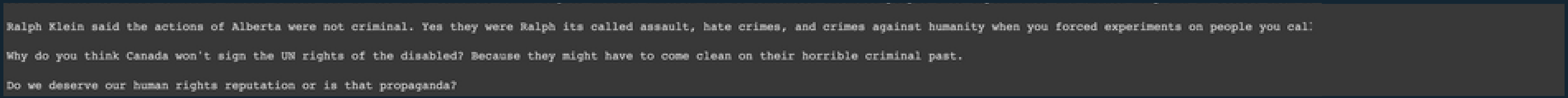
```
100.0% The fact is that the Trump regime is white supremacist Christian Republican extremists who hate everything our country stands for.
100.0% Wow these Nazi White Supremacist guys don't mess around... carrying guns and don't hesitate to use them. Scary dudes...
100.0% "a land of freedom for all" Dejain?

Even for white supremacists?
```

Politics



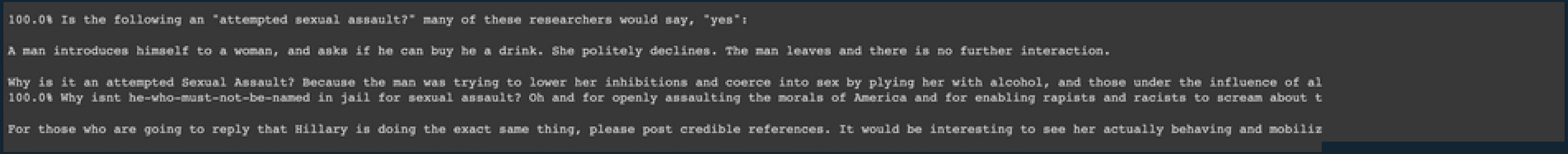
```
100.0% You'd have to drain the swamp to find the loathsome bottom-dwellers populating the Trump cabal.
100.0% Kill `em all, but save 6 for pallbearers! The Trump "administration" is evil, cruel, and dark. Get these maniacs gone! Let's drain the swamp of the Trump predators!
100.0% Fire the Special Prosecutor, drain the whole swamp. Stop the Russia wet dream and start prosecuting people like Hillary, Lynch, Holder and other Obama appointed criminals.
```

Human Rights



```
Ralph Klein said the actions of Alberta were not criminal. Yes they were Ralph its called assault, hate crimes, and crimes against humanity when you forced experiments on people you cal

Why do you think Canada won't sign the UN rights of the disabled? Because they might have to come clean on their horrible criminal past.

Do we deserve our human rights reputation or is that propaganda?
```

Sexual Abuse



```
100.0% Is the following an "attempted sexual assault?" many of these researchers would say, "yes":

A man introduces himself to a woman, and asks if he can buy he a drink. She politely declines. The man leaves and there is no further interaction.

Why is it an attempted Sexual Assault? Because the man was trying to lower her inhibitions and coerce into sex by plying her with alcohol, and those under the influence of al
100.0% Why isnt he-who-must-not-be-named in jail for sexual assault? Oh and for openly assaulting the morals of America and for enabling rapists and racists to scream about t

For those who are going to reply that Hillary is doing the exact same thing, please post credible references. It would be interesting to see her actually behaving and mobiliz
```

*refer to the python file for complete text

ASSIGNING LABELS TO EACH COMMENT

- There are many examples in which we see a single comment can cover many labels of toxicity, therefore it does not make sense to tag every comment with only one label
- Therefore, we devise a threshold assignment, in which we see if a particular comment belongs to multiple topics (based on its NMF probability) we compare it with the threshold and assign multiple labels
- For this exercise, we use a threshold of 0.3

Comment:

Here I'll fix this Leftist sophomore literary crap once and for all
#WHITELIVESMATTER
Now STFU communist marxist evil
If you respond in any way that denies me my right to protect my whiteness
you are f'in racist

Assigned Labels:
Racism
Human Rights

MODEL INTERPRETATION

We are able to correctly tag these comments on the basis of their content into one of those 4 topics. This can help the company analyze the comments data to see,

- Spike in a specific topic category for a particular region
- Most prominent toxic topics during a specific time window or geographic location
- Flag such comments easily and disable these users from the product/ service

Using more sophisticated techniques, we will be able to enhance these topic modeling approaches, such as using a LDA model

REFERENCES

- https://www.kaggle.com/code/thrillanalysis/amazon-reviews-analysis
- https://www.kaggle.com/c/jigsaw-unintended-bias-in-toxicity-classification/data
- https://towardsdatascience.com/simple-method-of-targeted-tf-idf-topic-modeling-using-yelp-open-dataset-298e019d6c09
- https://www.machinelearningplus.com/nlp/topic-modeling-gensim-python/#15visualizethetopicskeywords
- https://stackoverflow.com/questions/44060759/how-to-implement-latent-dirichlet-allocation-to-give-bigrams-trigrams-in-topics
- https://michael-fuchs-python.netlify.app/2021/05/25/nlp-text-pre-processing-ii-tokenization-and-stop-words/

THANK YOU