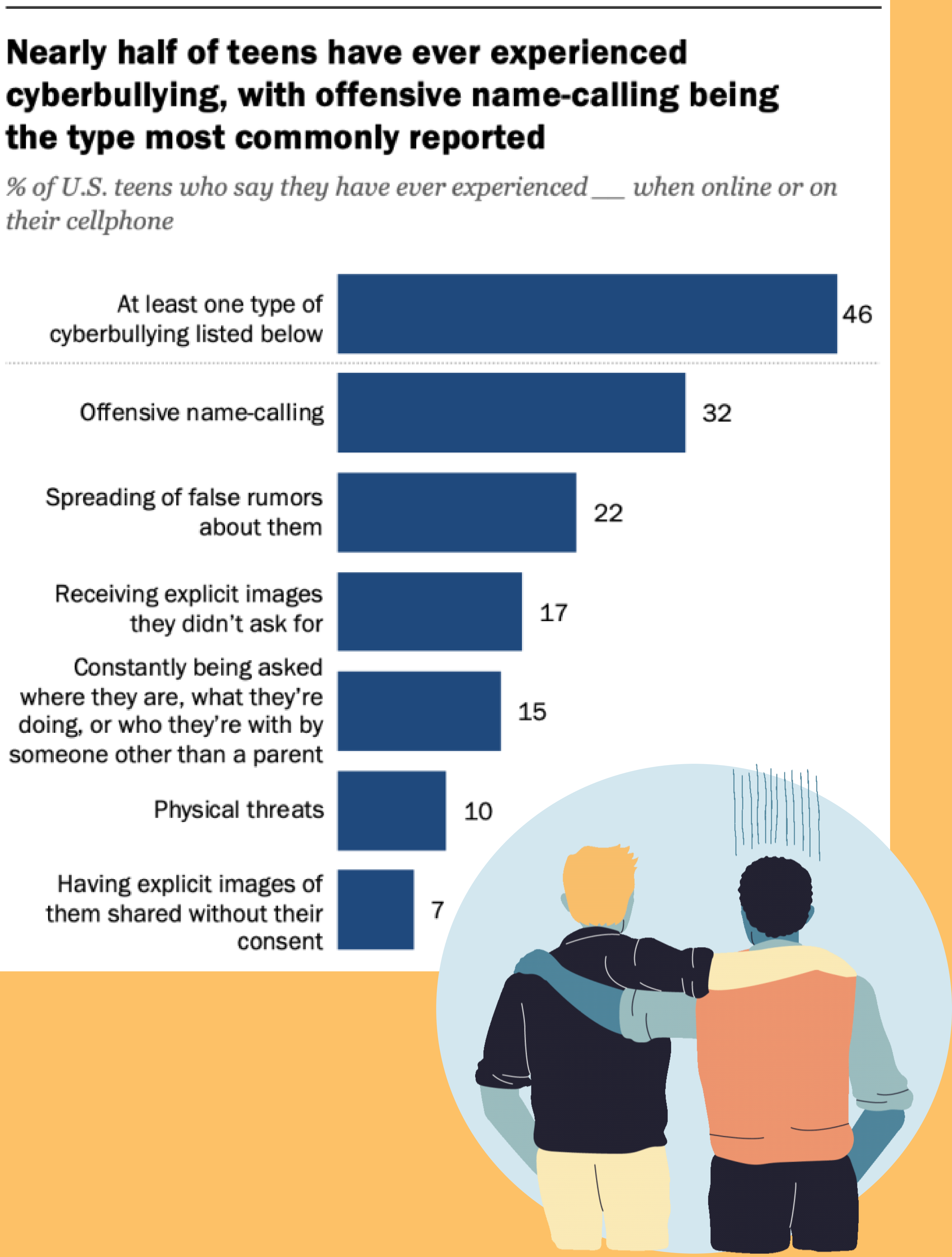
Natural Language Process (NLP) Model for Cyber bullying

The Importance of identifying pattern of Cyberbullying

Social media usage: As of 2021, there are 4.33 billion social media users worldwide, and the number is projected to increase to 4.41 billion by 2025. (Source: Statista)

- Cyberbullying prevalence: In a survey of American teenagers aged 13 to 17 years old, 59% reported experiencing some form of cyberbullying. (Source: Pew Research Center)

- Types of cyberbullying: According to a study by the Cyberbullying Research Center, the most common forms of cyberbullying include name-calling (27.9%), spreading rumors (26.3%), and posting embarrassing pictures or videos (25.6%).



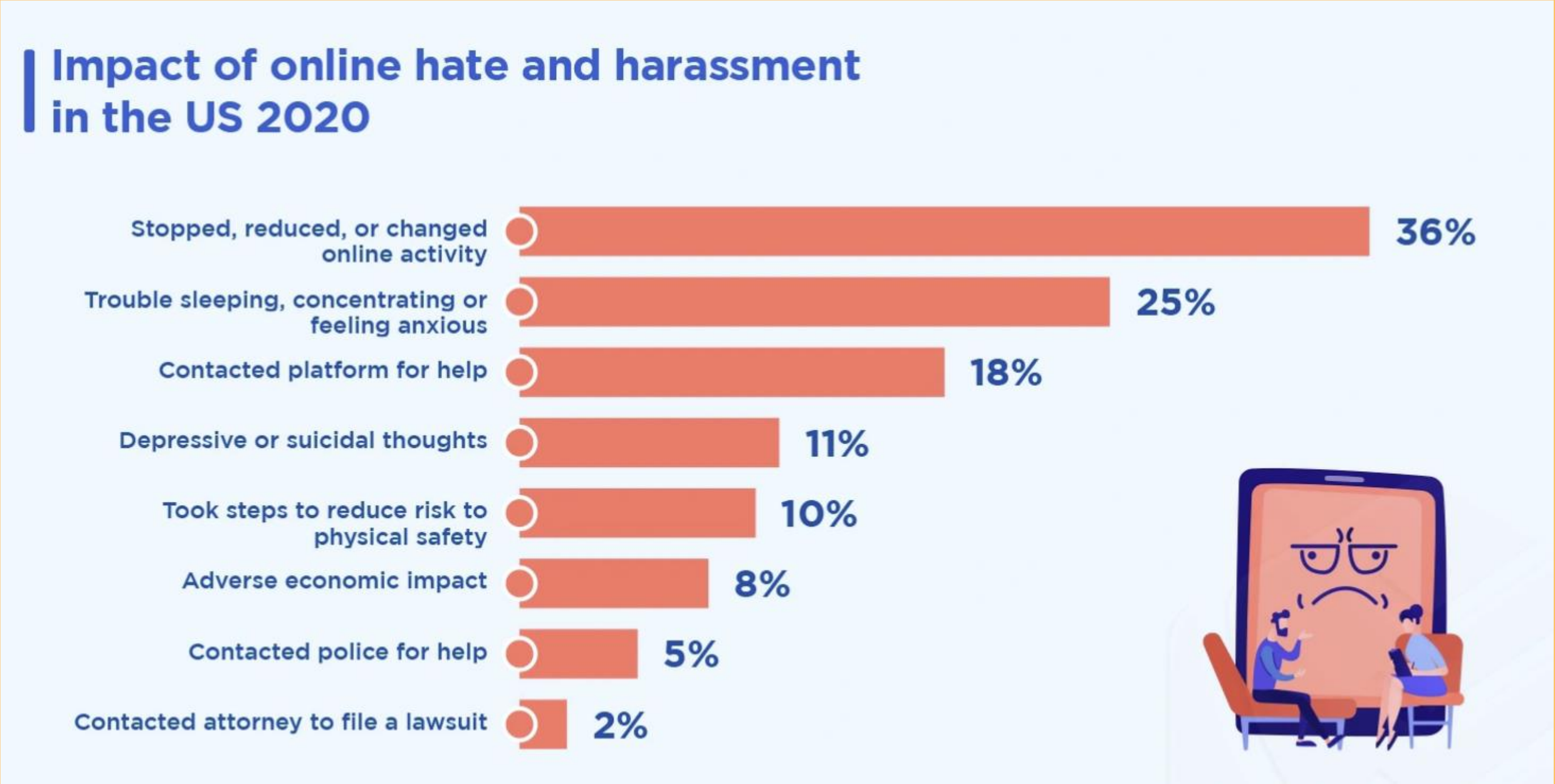
Nearly half of teens have ever experienced cyberbullying, with offensive name-calling being the type most commonly reported

% of U.S. teens who say they have ever experienced ___ when online or on their cellphone

The impact of Cyberbullying

The study published in the journal JAMA Pediatrics found that youth who were cyberbullied were more than twice as likely to attempt suicide than those who were not cyberbullied. Specifically, the study found that:

1. 15.9% of youth who reported being cyberbullied had attempted suicide, compared to 6.1% of youth who had not been cyberbullied.
2. The association between cyberbullying and suicide attempts remained significant even after controlling for other factors that can contribute to suicide risk, such as depression, substance use, and previous suicide attempts.



Impact of online hate and harassment in the US 2020

Investigating Cyber Bullying through Social Media Analytics

Purpose:

Our team's purpose is to develop an NLP model to identify patterns of cyberbullying in user comments and posts on Reddit. This model can help social media companies improve their content moderation policies and protect their users from the harmful effects of cyberbullying.

Significance for practitioners and business:

- Provides a way for social media companies to improve content moderation policies and protect users from cyberbullying
- Helps to create a safer and more inclusive online environment for all users
- Enhances brand reputation and trust with users by demonstrating a commitment to responsible and ethical social media practices

How NLP Models Can Help Combat Cyberbullying on Social Media

- **Automated detection:** Detect potentially problematic content in user comments and posts on Reddit, flagging abusive language and common cyberbullying phrases.

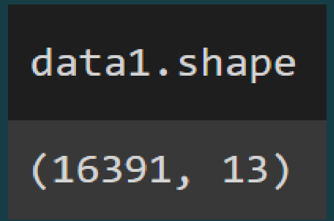
- **Identification of victims and perpetrators:** Analyze language patterns to identify victims and perpetrators of cyberbullying, based on negative language frequency and specific words used.

- **Real-time response:** NLP models enable real-time monitoring of social media to detect cyberbullying and send alerts for swift action to Saves time and resources.

- **Customization:** Ensure accurate cyberbullying detection across different social media platforms, communities, and languages by accounting for nuances in language use.

Dataset Summary

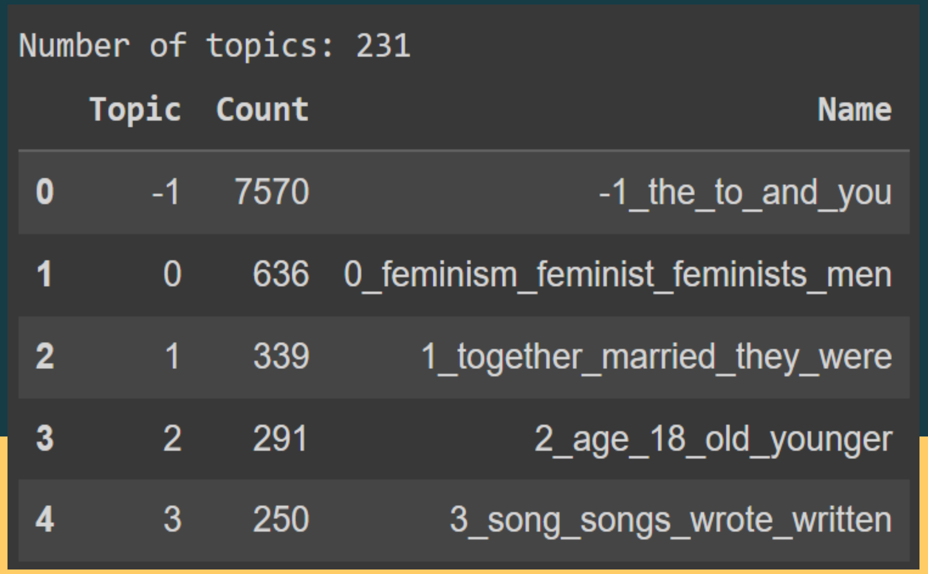
- Data Source: initial subreddits considered are r/Vent, r/Fauxmoi, r/antifeminists, r/fakedisordercringe,and r/MadeMeSmile
- Subreddits : r/fakedisordercringe and r/MadeMeSmile are removed since potential topics were not found.
- Data shape:



```
data1.shape
```

```
(16391, 13)
```

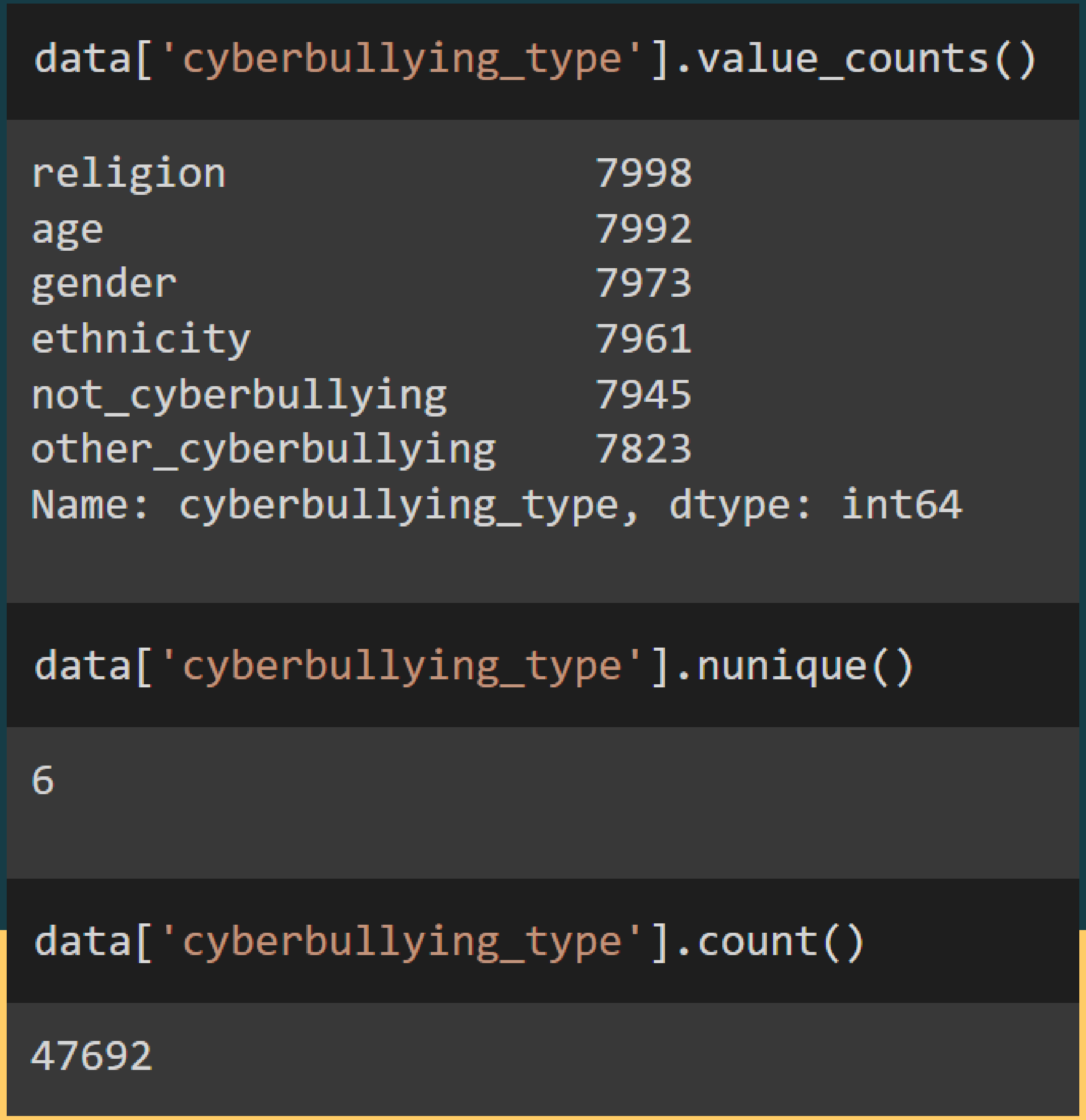
- Topic extraction:



Number of topics: 231

	Topic	Count	Name
0	-1	7570	-1_the_to_and_you
1	0	636	0_feminism_feminist_feminists_men
2	1	339	1_together_married_they_were
3	2	291	2_age_18_old_younger
4	3	250	3_song_songs_wrote_written

- Types of cyberbullying considered:



```
data['cyberbullying_type'].value_counts()
```

```
religion               7998
age                    7992
gender                 7973
ethnicity              7961
not_cyberbullying      7945
other_cyberbullying    7823
Name: cyberbullying_type, dtype: int64
```

```
data['cyberbullying_type'].nunique()
```

```
6
```

```
data['cyberbullying_type'].count()
```

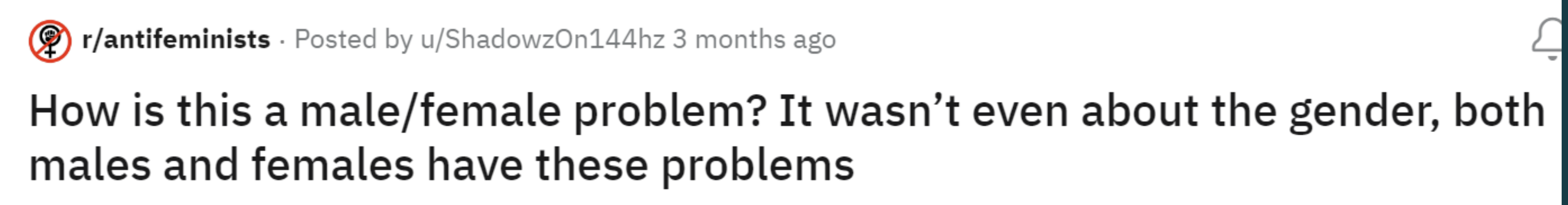
```
47692
```

Data Acquisition

- Datasets were collected related to Cyberbullying subreddits using PRAW (Python Reddit API Wrapper) which is a Python library that provides interface for accessing the Reddit API.

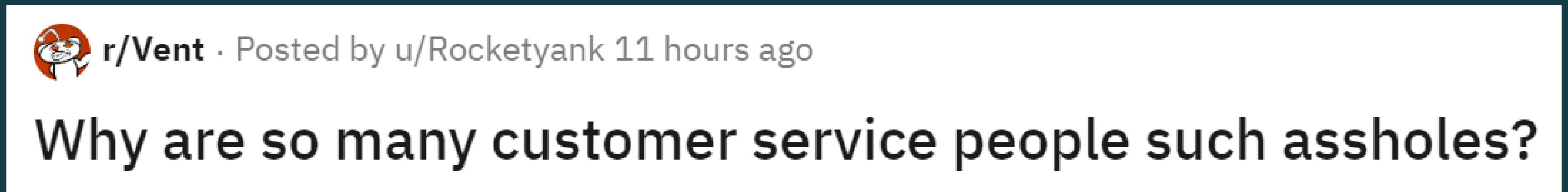
- Scraped over 25,000++ comments from 5 subreddits.

- Focused on 16,391 comments from the 3 subreddits (r/Vent, r/Fauxmoi, r/antifeminists).



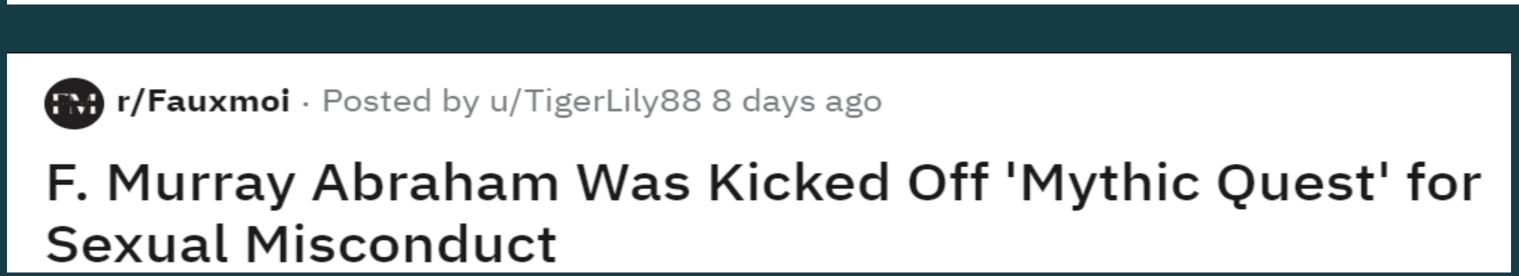
r/antifeminists · Posted by u/ShadowzOn144hz 3 months ago

How is this a male/female problem? It wasn't even about the gender, both males and females have these problems



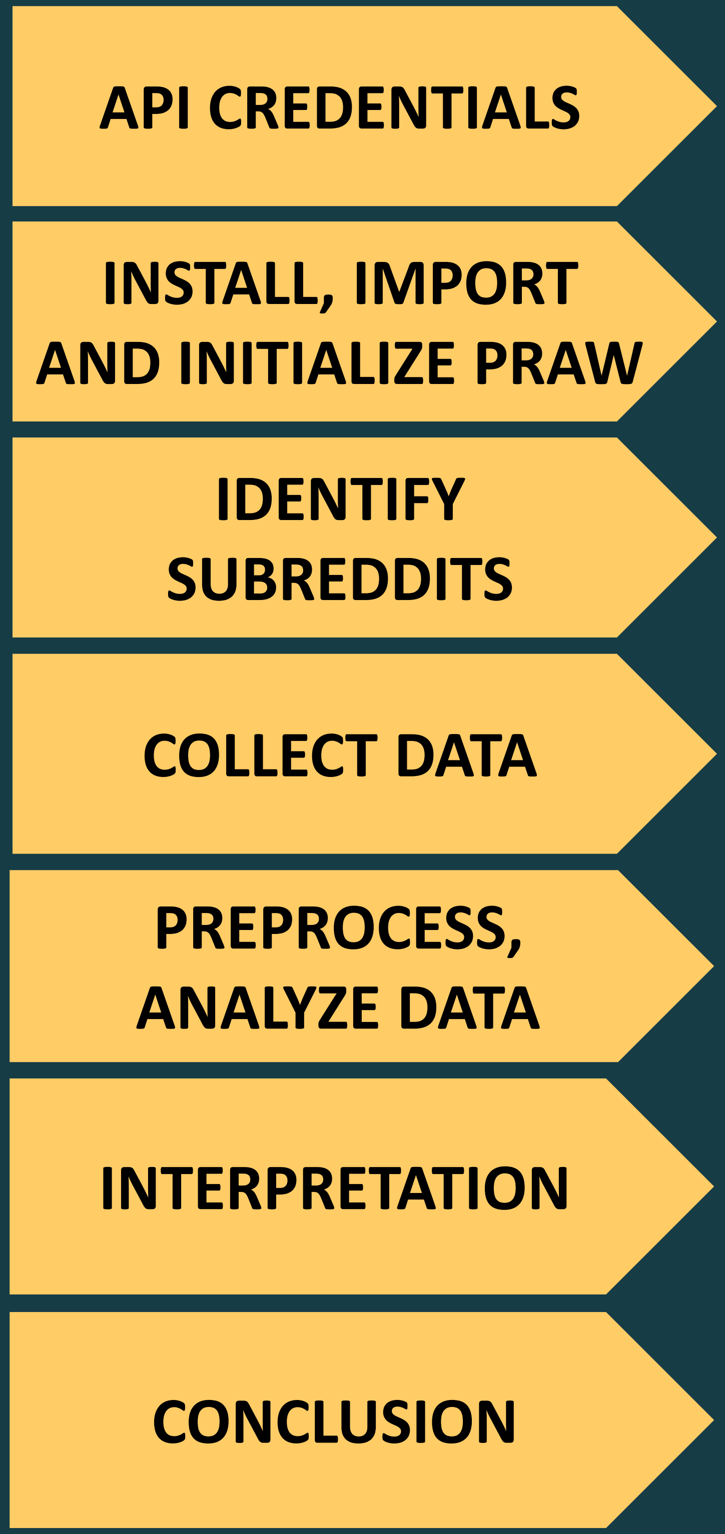
r/Vent · Posted by u/Rocketyank 11 hours ago

Why are so many customer service people such assholes?



r/Fauxmoi · Posted by u/TigerLily88 8 days ago

F. Murray Abraham Was Kicked Off 'Mythic Quest' for Sexual Misconduct



API CREDENTIALS

INSTALL, IMPORT AND INITIALIZE PRAW

IDENTIFY SUBREDDITS

COLLECT DATA

PREPROCESS, ANALYZE DATA

INTERPRETATION

CONCLUSION

Data mining Methods

For extracting relevant Topics

- We have used BERTopic package for topic modeling on a dataset containing comments from different subreddits.
- The BERTopic with UMAP is used for topic modeling and probability calculation.
- The resulting topics and their corresponding probabilities are saved, and the top words for each topic are displayed using the get_topic function from the BERTopic package.

For our Cyberbullying detection model

- Involve tokenizing input text data for model training using the BERT AutoTokenizer from the transformers library to tokenize the text data.
- The BERT (Bidirectional Encoder Representations from Transformers) model is a pre-trained natural language processing model which uses Huggingface's Transformers library to perform variety of tasks while fine-tuning the model.

Overview of how our project can be used to help identify pattern of Cyberbullying

- **Data sources:** Collecting and preprocessing text data from various online platforms. This diverse data pool can ensure a comprehensive analysis.
- **BERTopic:** Utilizing BERTopic for topic modelling to uncover hidden patterns in the data. This helps to identify common phrases, words, and themes indicative of cyberbullying behavior.
- **BERT integration:** Leveraging BERT's natural language processing capabilities for text analysis, helping identify potential instances of cyberbullying.
- **Pattern detection:** The combination of BERT and BERTopic allows your project to detect patterns in cyberbullying behavior, such as frequently used offensive language, targeted harassment, and other abusive actions.

Steps of how our project can be used to help identify pattern of Cyberbullying

Data Preparation :

- Load, concatenate, clean, and preprocess datasets.

Topic Modeling with BERTopic:

- Create BERTopic model and fit it on cleaned text data.

Classification using BERT:

- Tokenize input text using BERT's AutoTokenizer.
- Create custom neural network model with BERT as base model.
- Train model, validate, and evaluate using classification report.

Applying the Model:

- Tokenize headlines, predict cyberbullying type, and decode predictions using label encoder.
- Add predictions to dataset and save as CSV file.

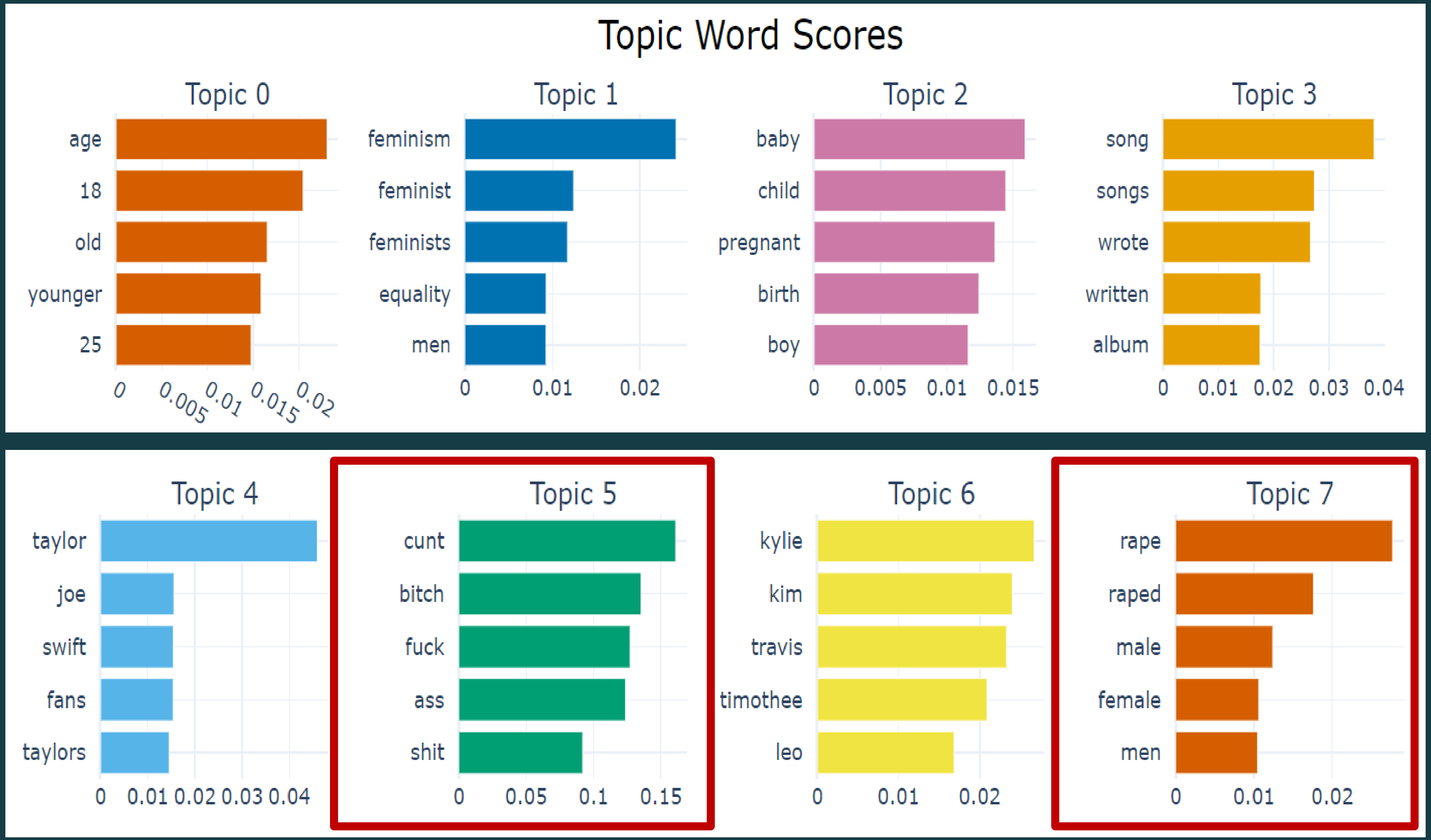
Testing the Model with New Input Text:

- Tokenize the input text using BERT's AutoTokenizer.
- Predict the cyberbullying type using the trained model.
- Decode the predicted label and classify the input text.

Warning: After this slide, there will be profane language. It is to just show the real-world scenarios and not intended to hurt anyone.

Data Analysis

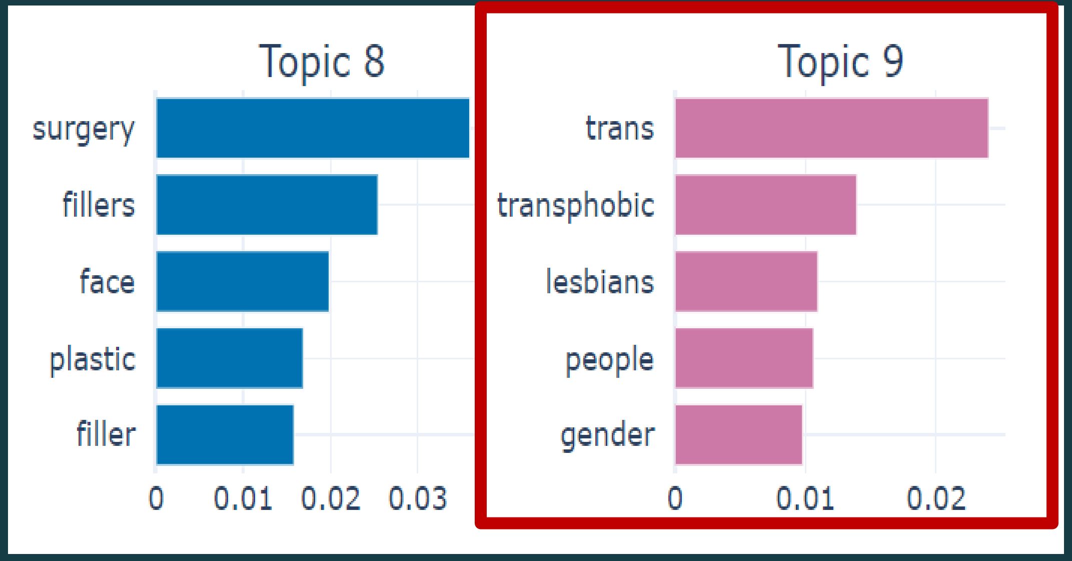
- Collected **16,391** messages from the **3** subreddits including, **Antifeminist, Fauxmoi and vent.**
- Split the message into sentences.
- Identified **10 Topic Terms, 4 subset and 3 clusters** of topics using **BERTopic.**



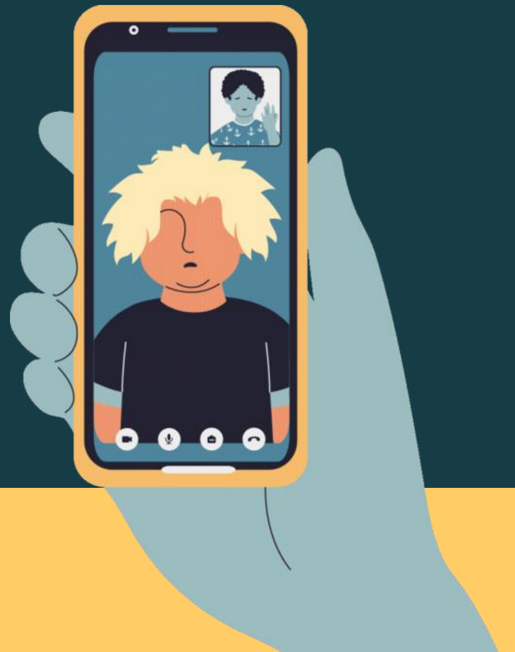
10 Topic Terms summarize

These Topics Word Scores Visualization helps in gaining more insight about relevant words of each topic.

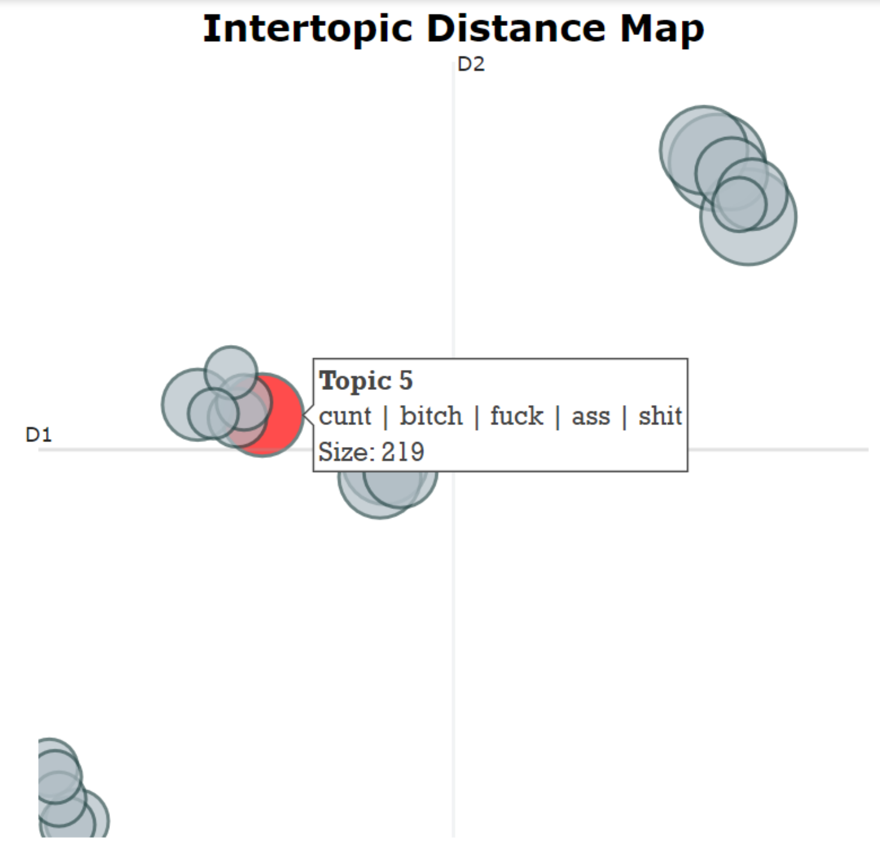
According to these 10 Bar charts, the potential topic that relates to Cyberbullying is **Topic 5,7 and 9.**

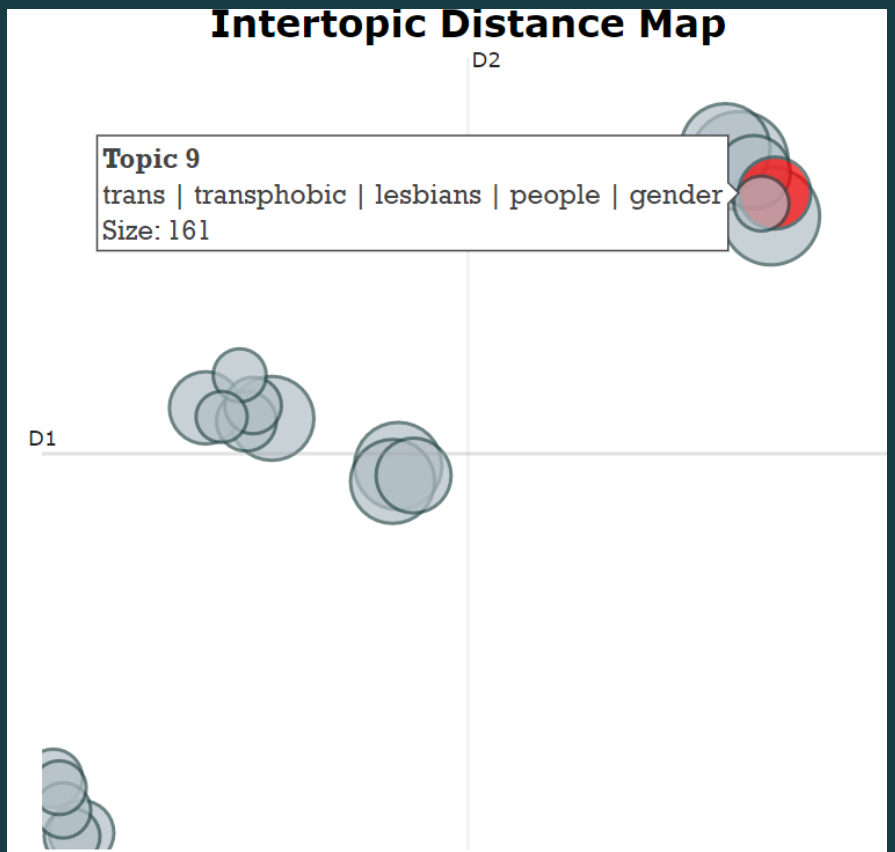


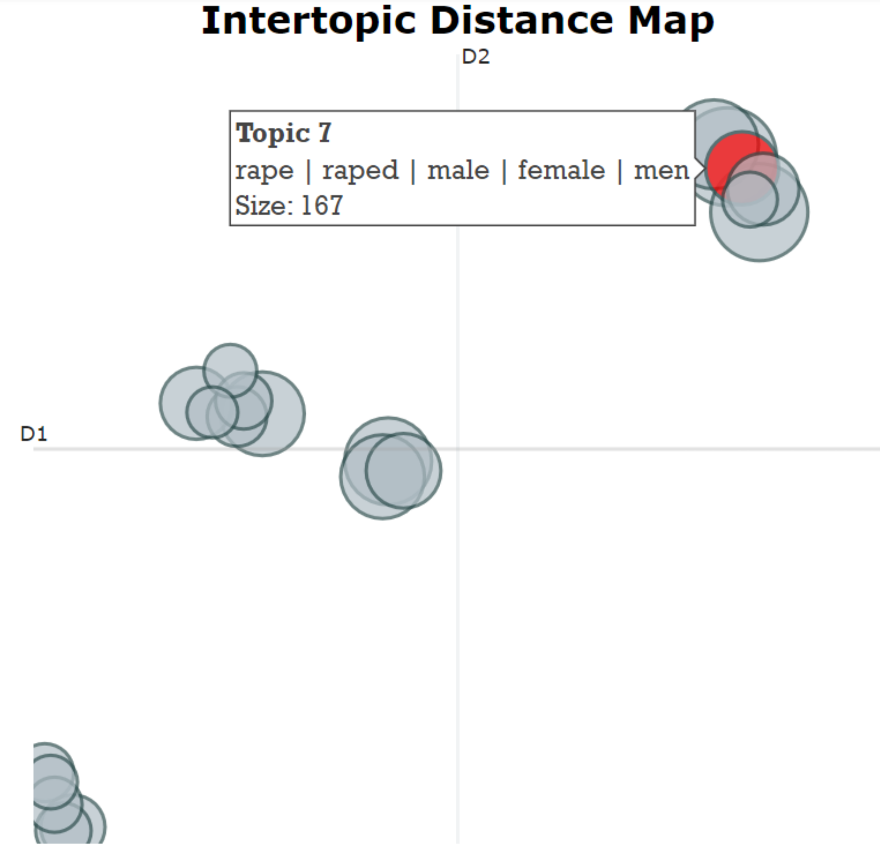
The c-TF-IDF score

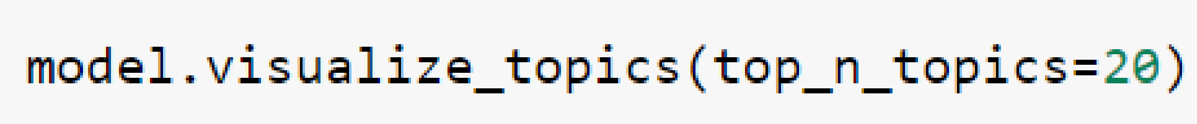


4 subsets of Intertopic Distance Map summarize





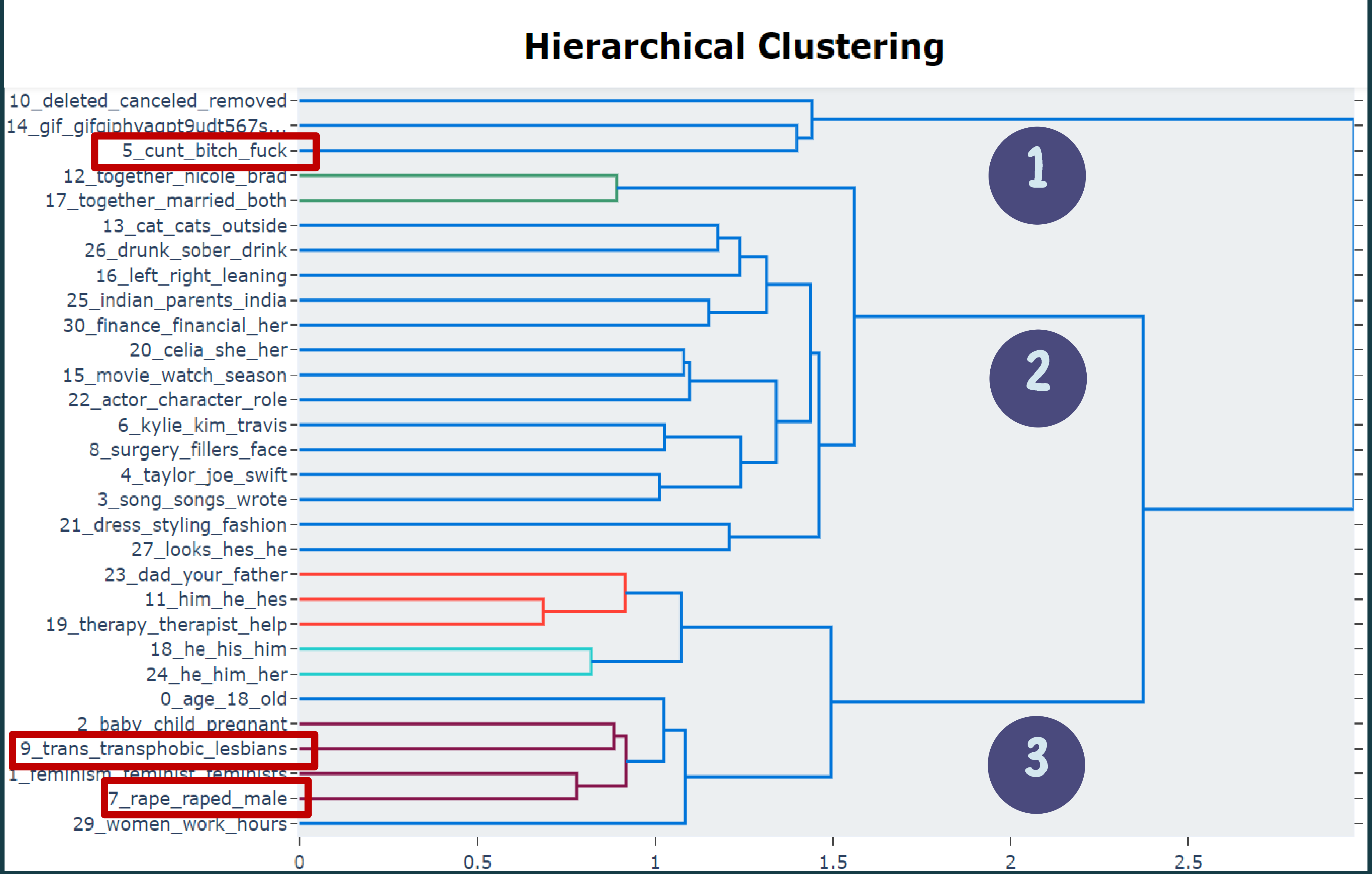




```
model.visualize_topics(top_n_topics=20)
```

- **This** is a visualization used in topic modeling to display the relationships between topics.
- It is a two-dimensional map that represents the distance between topics based on their semantic similarity.
- Topics that are closer together on the map are more similar to each other, while topics that are farther apart are more dissimilar.
- Topic 5 may represent more general vulgar language, while topic 7 and 9 may be more specific to discrimination against gender especially Transgender individuals.

3 clusters of topic summarize



```
model.visualize_hierarchy(top_n_topics=30)
```

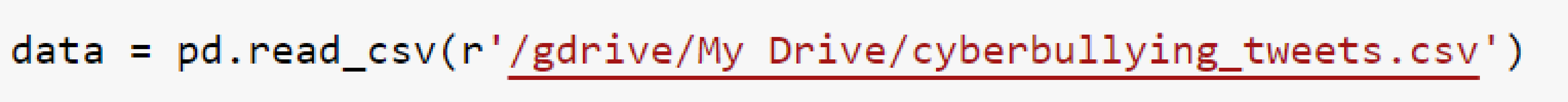
After organizing the topics hierarchically to help reduce the number of topics, we can reduce 30 topics to only 3 topics.

1. Related to general vulgar language
2. Related to people typical conversation such as gossiping and activities
3. Related to gender and sexuality

**However, since Reddit has rules and regulations against cyberbullying that prohibit anyone from making derogatory comments about individuals. Hence, our findings are limited and do not provide sufficient information for us to flag instances of cyberbullying.

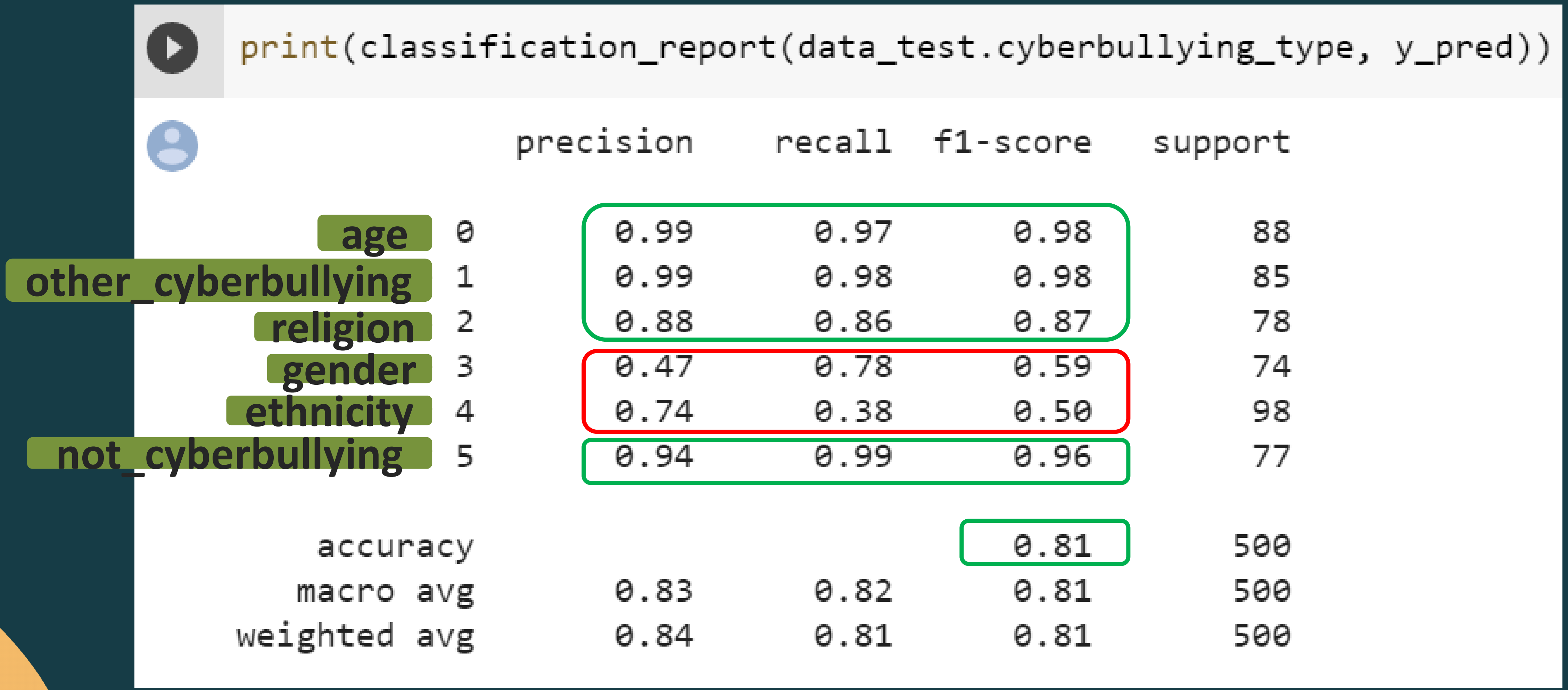
Result summary from our Cyberbullying detection model

1 From ‘cyberbullying_tweets’ data source



```
data = pd.read_csv(r'/gdrive/My Drive/cyberbullying_tweets.csv')
```

2 Using BERT, a pre-trained language model, this is the classification report which contains accuracy, precision, recall, and f1-score of our model.



```
print(classification_report(data_test.cyberbullying_type, y_pred))
```

```
                           precision    recall  f1-score   support

                age  0       0.99      0.97      0.98        88
other_cyberbullying  1       0.99      0.98      0.98        85
           religion  2       0.88      0.86      0.87        78
             gender  3       0.47      0.78      0.59        74
          ethnicity  4       0.74      0.38      0.50        98
 not_cyberbullying   5       0.94      0.99      0.96        77

           accuracy                              0.81       500
          macro avg          0.83      0.82      0.81       500
       weighted avg          0.84      0.81      0.81       500
```

According to the metrics from cyberbullying type 0,1,2,5, and overall accuracy, the results are very promising. However, since our benchmark is at 0.8, there are rooms for our model to be developed.

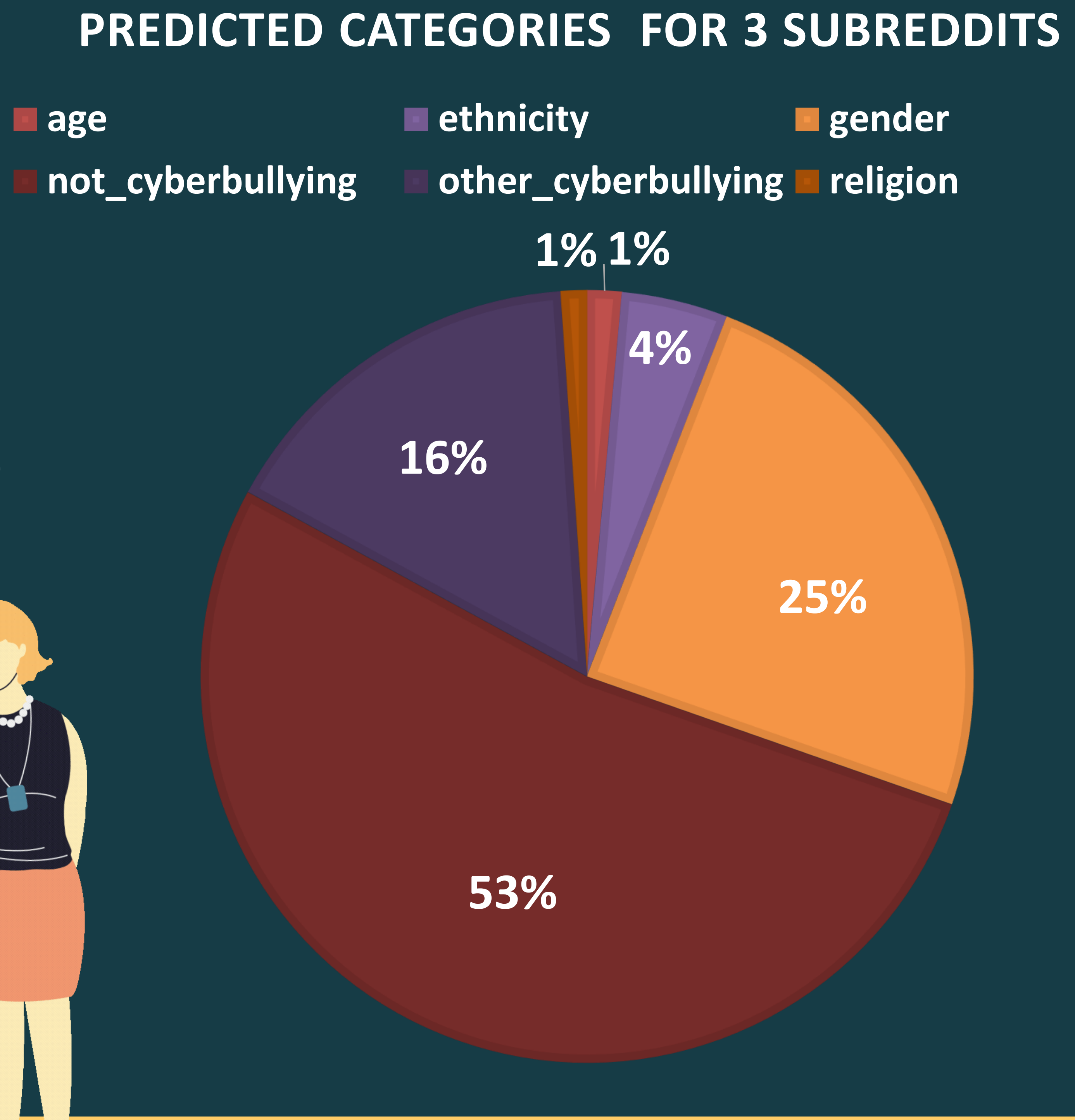
How our model can drive effective cyberbullying prevention strategies

•**Real-time detect cyberbullying**: Allowing individuals or organizations to take immediate action and prevent further harm.

•**Empowering individuals**: By providing individuals with the ability to detect and prevent cyberbullying, our model can empower them to take control of their online safety and well-being.

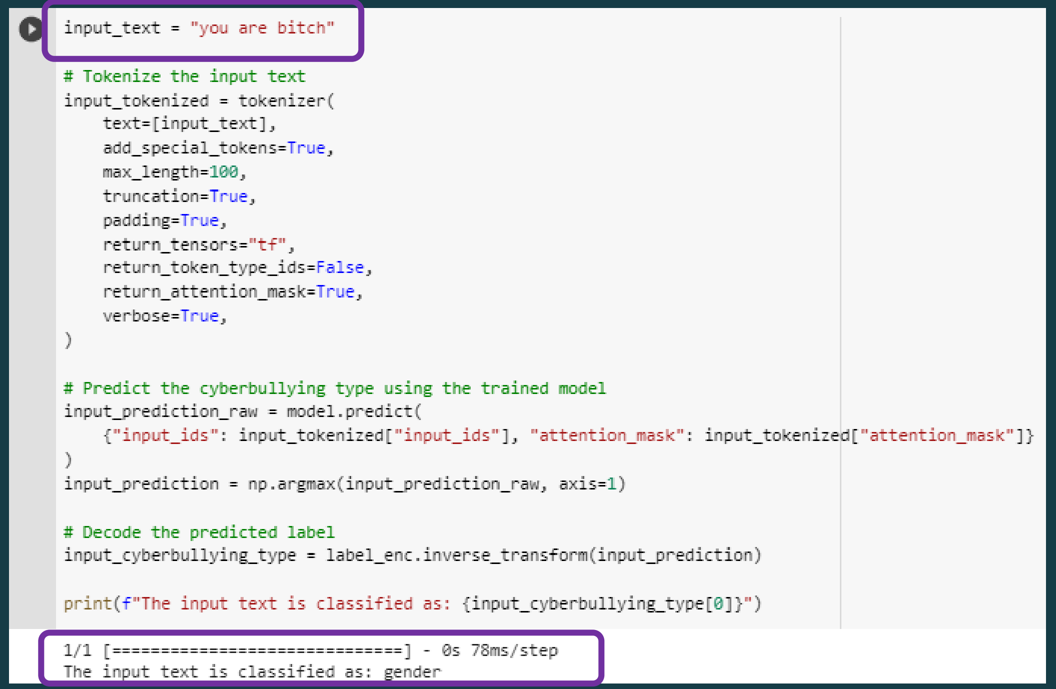
•**Raising awareness**: Our model can raise awareness about the prevalence and harmful effects of cyberbullying, encouraging individuals and organizations to take proactive steps to prevent it.

•**Data insights for policy**: Our model's data insights can inform policies and advocacy efforts aimed at preventing cyberbullying at the community, state, or national level, ultimately contributing to reducing rates of suicide.



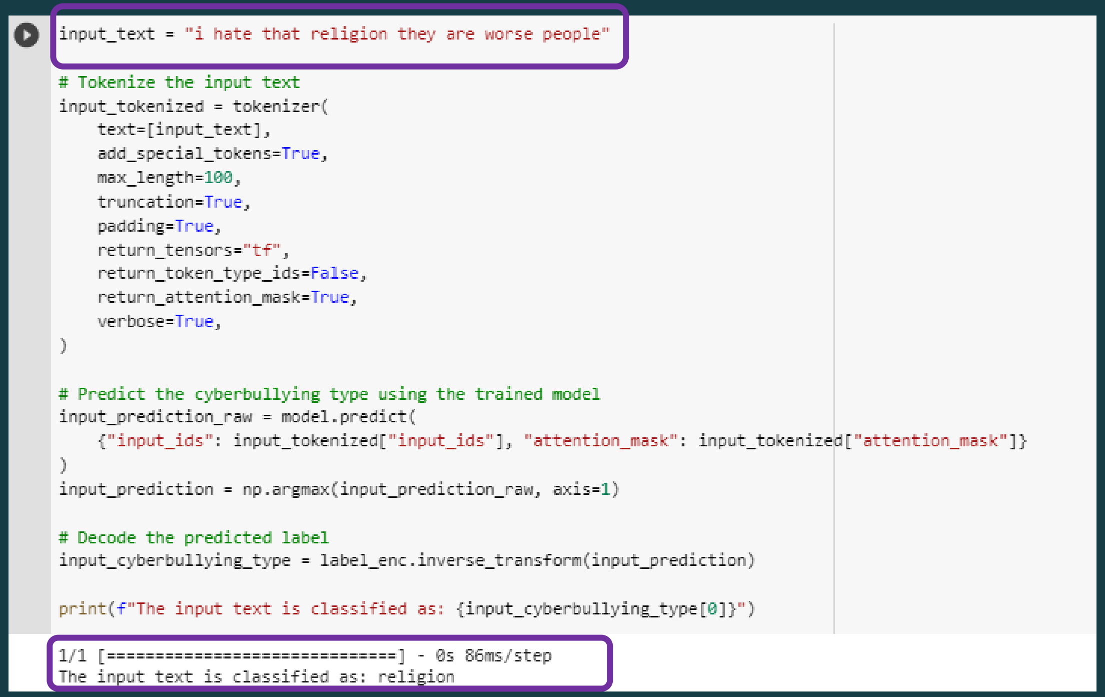
Conclusion

- Collected 16,391 messages from the 3 subreddits (Antifeminist, Fauxmoi, vent).
- Topics were extracted and reduced using BERTopic, resulting in only 3 key themes related.
- Our Cyberbullying detection model not only flags potential sentences when typed in, but also categorizes them into five related Cyberbullying types with an overall accuracy of 81%.



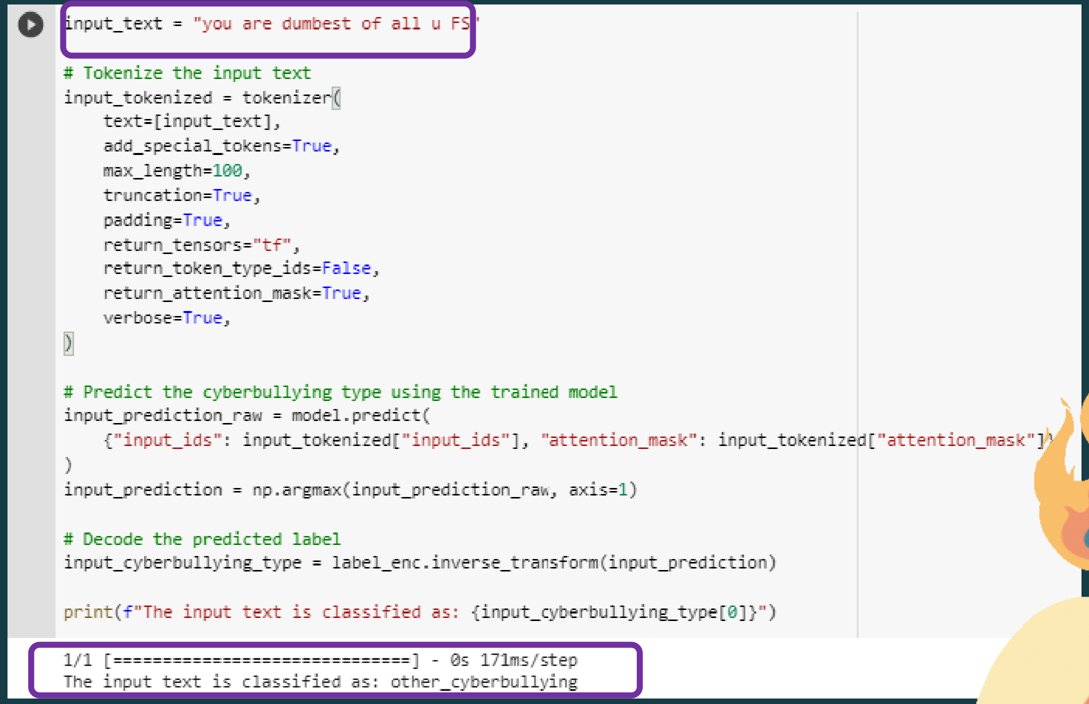
```
input_text = "you are bitch"

# Tokenize the input text
input_tokenized = tokenizer(
    text=[input_text],
    add_special_tokens=True,
    max_length=100,
    truncation=True,
    padding=True,
    return_tensors="tf",
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True,
)

# Predict the cyberbullying type using the trained model
input_prediction_raw = model.predict(
    {"input_ids": input_tokenized["input_ids"], "attention_mask": input_tokenized["attention_mask"]}
)
input_prediction = np.argmax(input_prediction_raw, axis=1)

# Decode the predicted label
input_cyberbullying_type = label_enc.inverse_transform(input_prediction)

print(f"The input text is classified as: {input_cyberbullying_type[0]}")
```

```
1/1 [==============================] - 0s 78ms/step
The input text is classified as: gender
```



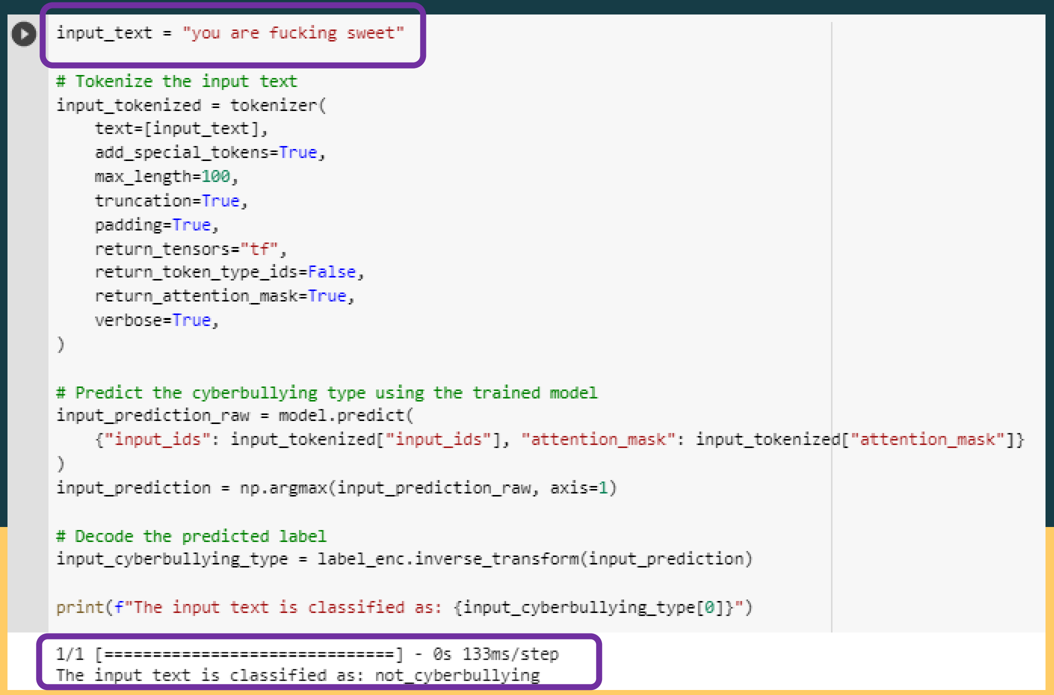
```
input_text = "i hate that religion they are worse people"

# Tokenize the input text
input_tokenized = tokenizer(
    text=[input_text],
    add_special_tokens=True,
    max_length=100,
    truncation=True,
    padding=True,
    return_tensors="tf",
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True,
)

# Predict the cyberbullying type using the trained model
input_prediction_raw = model.predict(
    {"input_ids": input_tokenized["input_ids"], "attention_mask": input_tokenized["attention_mask"]}
)
input_prediction = np.argmax(input_prediction_raw, axis=1)

# Decode the predicted label
input_cyberbullying_type = label_enc.inverse_transform(input_prediction)

print(f"The input text is classified as: {input_cyberbullying_type[0]}")
```

```
1/1 [==============================] - 0s 86ms/step
The input text is classified as: religion
```



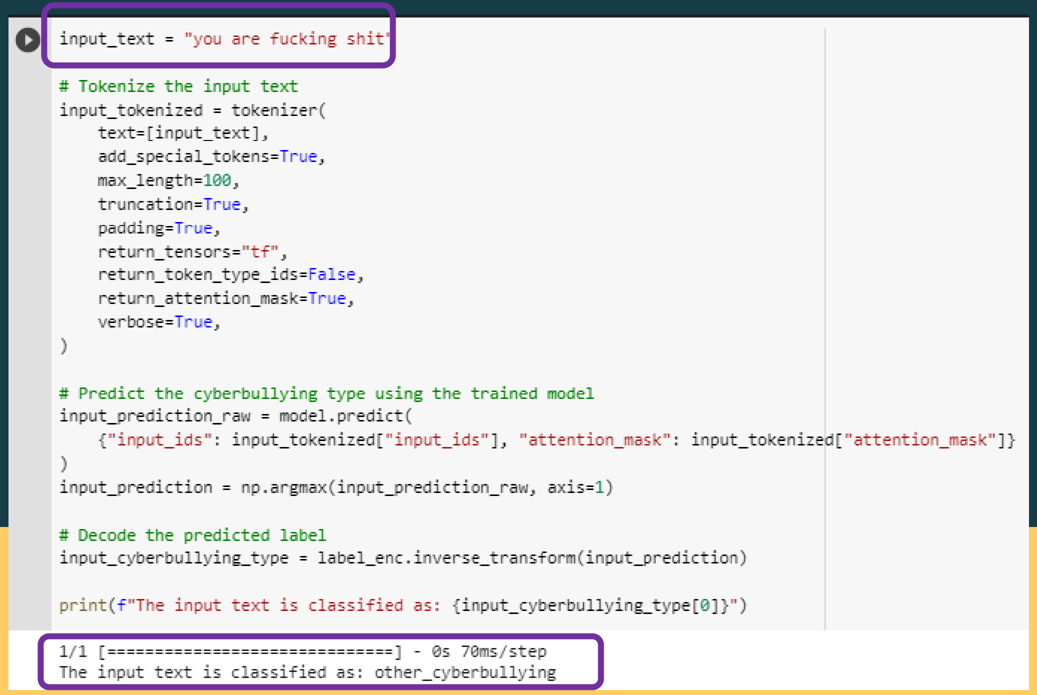
```
input_text = "you are dumbest of all u FS"

# Tokenize the input text
input_tokenized = tokenizer(
    text=[input_text],
    add_special_tokens=True,
    max_length=100,
    truncation=True,
    padding=True,
    return_tensors="tf",
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True,
)

# Predict the cyberbullying type using the trained model
input_prediction_raw = model.predict(
    {"input_ids": input_tokenized["input_ids"], "attention_mask": input_tokenized["attention_mask"]}
)
input_prediction = np.argmax(input_prediction_raw, axis=1)

# Decode the predicted label
input_cyberbullying_type = label_enc.inverse_transform(input_prediction)

print(f"The input text is classified as: {input_cyberbullying_type[0]}")
```

```
1/1 [==============================] - 0s 171ms/step
The input text is classified as: other_cyberbullying
```



```
input_text = "you are fucking sweet"

# Tokenize the input text
input_tokenized = tokenizer(
    text=[input_text],
    add_special_tokens=True,
    max_length=100,
    truncation=True,
    padding=True,
    return_tensors="tf",
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True,
)

# Predict the cyberbullying type using the trained model
input_prediction_raw = model.predict(
    {"input_ids": input_tokenized["input_ids"], "attention_mask": input_tokenized["attention_mask"]}
)
input_prediction = np.argmax(input_prediction_raw, axis=1)

# Decode the predicted label
input_cyberbullying_type = label_enc.inverse_transform(input_prediction)

print(f"The input text is classified as: {input_cyberbullying_type[0]}")
```

```
1/1 [==============================] - 0s 133ms/step
The input text is classified as: not_cyberbullying
```



```
input_text = "you are fucking shit"

# Tokenize the input text
input_tokenized = tokenizer(
    text=[input_text],
    add_special_tokens=True,
    max_length=100,
    truncation=True,
    padding=True,
    return_tensors="tf",
    return_token_type_ids=False,
    return_attention_mask=True,
    verbose=True,
)

# Predict the cyberbullying type using the trained model
input_prediction_raw = model.predict(
    {"input_ids": input_tokenized["input_ids"], "attention_mask": input_tokenized["attention_mask"]}
)
input_prediction = np.argmax(input_prediction_raw, axis=1)

# Decode the predicted label
input_cyberbullying_type = label_enc.inverse_transform(input_prediction)

print(f"The input text is classified as: {input_cyberbullying_type[0]}")
```

```
1/1 [==============================] - 0s 70ms/step
The input text is classified as: other_cyberbullying
```

Q and A

Thank you