

# **Online Toxicity Analysis: Comparative Evaluation of Text Classification and Topic Modeling Techniques**

**Text Mining and Search**  
**A.Y. 2025/2026**

Garofalo Kevin 930804

Lonati Giulio 924924

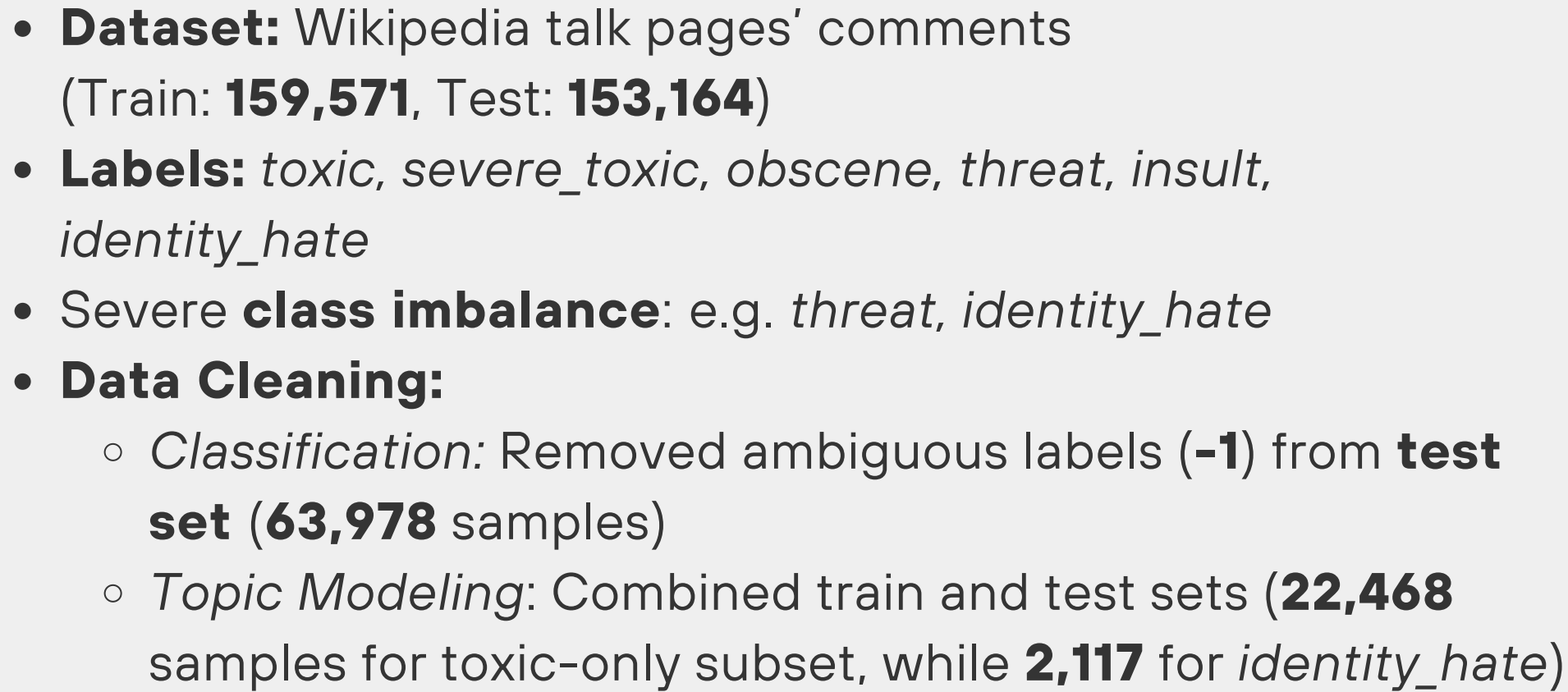
Siano Vincenzo 934168

# Introduction



- Online platforms face increasing challenges in moderating **toxic and abusive** user-generated content at scale
- Toxicity detection goes beyond sentiment analysis, requiring understanding of **context, sarcasm, and identity-based hate**
- Manual moderation is **not scalable**, motivating automated Text Mining and NLP-based solutions
- This project compares **classical machine learning** and **deep learning approaches** for:
  - **Text Classification** (multi-label, supervised learning)
  - **Topic Modeling** (unsupervised learning)

## 02



Label	Count	Percentage
Toxic	15,294	9.58%
Severe Toxic	1,595	1.00%
Obscene	8,449	5.29%
Threat	478	0.30%
Insult	7,877	4.94%
Identity Hate	1,405	0.88%
Clean (non-toxic)	143,346	89.83%



- Distinct vocabularies for **toxic vs. clean** comments
- Overlapping **length distributions** for both toxic and non-toxic
- Some labels (e.g. *threat, identity\_hate*) are extremely rare

# Preprocessing

## General cleaning:

1. HTML tags and URLs
2. User mentions
3. IP addresses
4. Repeated characters
5. Redundant whitespace



## Statistical models:

1. Tokenization
2. Stop-word removal
3. Lemmatization



## Deep Learning models:

no further cleaning to  
preserve context

# Text Classification

## Methods

- **Text Representation**
  - **Bag-of-Words** for statistical classifiers
  - **Contextual word embeddings** for Transformer-based models
- **Machine Learning classifiers**
  - *Naive Bayes, Logistic Regression, LinearSVM, SGD*
  - **One-vs-Rest** strategy
- **Deep Learning**
  - Fine-tuned **DistilBERT** for multi-label prediction
  - **Binary Cross-Entropy loss** and **AdamW optimizer**
  - **Dynamic padding** (*max length = 256*)



# Results

Evaluation

Model	ROC-AUC
Naive Bayes (TF)	0.8720
Logistic Regression (TF-IDF)	0.9806
LinearSVC (TF-IDF)	0.9688
SGD (Log Loss, TF-IDF)	0.9769
SGD (Hinge Loss, TF-IDF)	0.9797
DistilBERT	0.9922

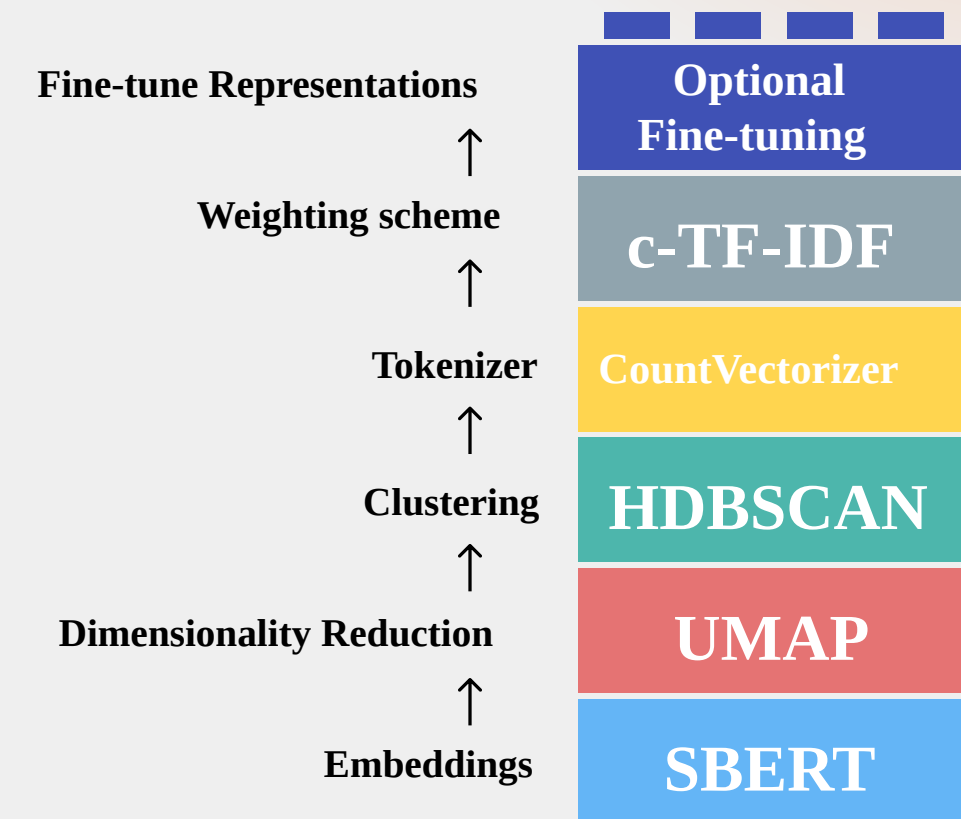
Test set

Model	ROC-AUC	F1 Macro
Logistic Regression	0.9618	0.50
DistilBERT	0.9865	0.61

# Topic Modeling

## Methods

- **Bag-of-Words approaches**
  - **Latent Dirichlet Allocation (LDA):** term-frequency (*TF*) representation
  - **Non-negative Matrix Factorization (NMF):** *TF-IDF* representation
- **Contextual embeddings (BERTopic)**
  - Neural topic modeling via **sentence embeddings**
  - **Clustering-based** topic discovery
    - Density-based (**HDBSCAN**)
    - Centroid-based (**k-means**)



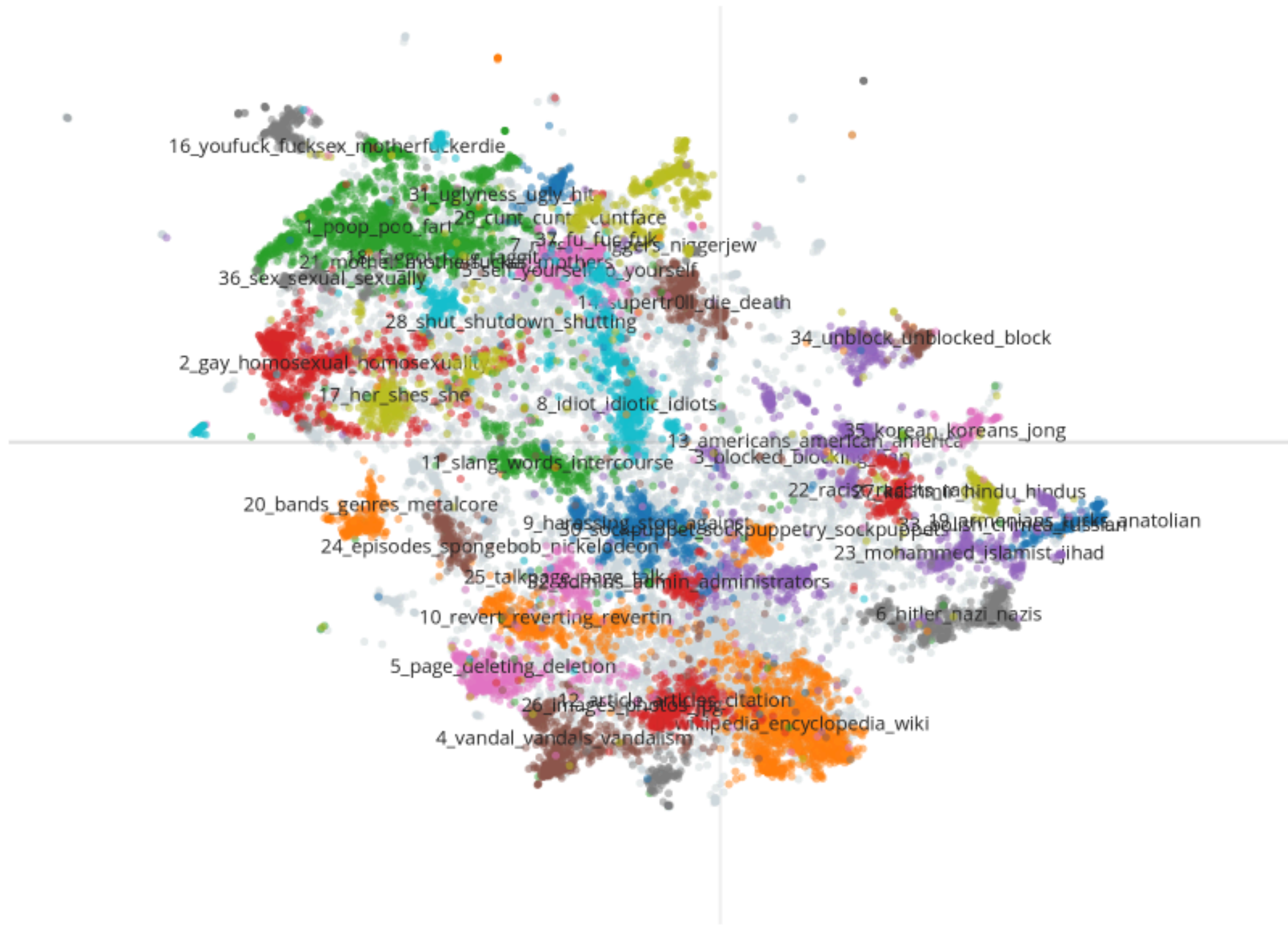
# Topic Modeling

## Results: Toxic-only

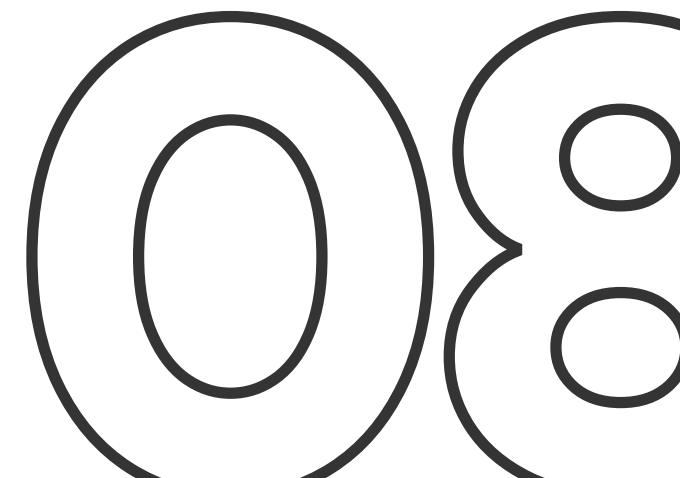
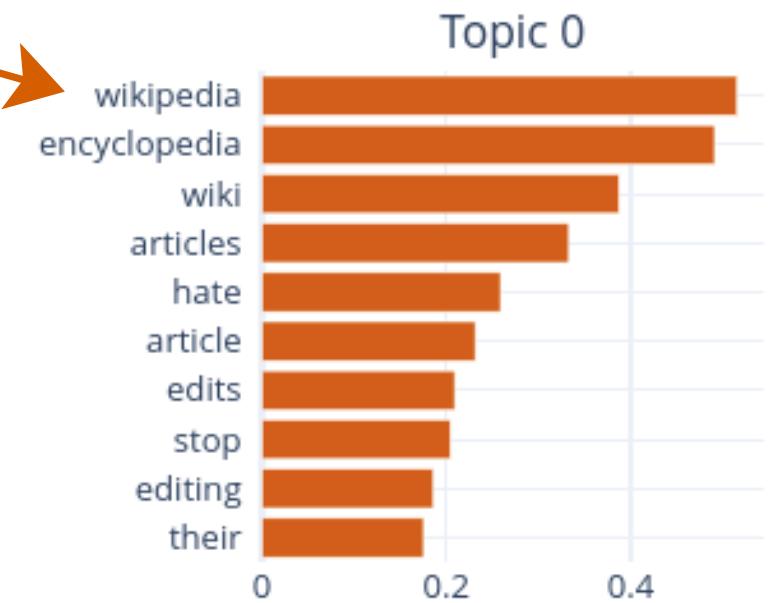
Model	Coherence	Lexical Diversity	Perplexity	Reconstruction Error	Semantic Diversity
LDA (k = 5)	0.514	0.96	0,0010	-	-
LDA (k = 10)	0.515	0.97	0,0011	-	-
NMF (k = 5)	0.490	0.92	-	147.1227	-
NMF (k = 10)	0.457	0.69	-	145.6265	-
BERTopic with HDBSCAN	0.367	-	-	-	0.528
BERTopic with k-means (k = 5)	0.454	-	-	-	0.276
BERTopic with k-means (k = 10)	0.455	-	-	-	0.329

# Topic Modeling

## 👤 BERTopic with HDBSCAN



- 0\_wikipedia\_encyclopedia\_wiki
- 1\_poop\_poo\_fart
- 2\_gay\_homosexual\_homosexuality
- 3\_blocked\_blocking\_ban
- 4\_vandal\_vandals\_vandalism
- 5\_page\_deleting\_deletion
- 6\_hitler\_nazi\_nazis
- 7\_nigger\_niggers\_niggerjew
- 8\_idiot\_idiotic\_idiots
- 9\_harassing\_stop\_against
- 10\_revert\_reverting\_revertin
- 11\_slang\_words\_intercourse
- 12\_article\_articles\_citation
- 13\_americans\_american\_america
- 14\_supertr0ll\_die\_death
- 15\_self\_yourselfgo\_yourself
- 16\_youfuck\_fucksex\_motherfuckerdie
- 17\_her\_shes\_she
- 18\_faggot\_fagg\_faggit
- 19\_armenians\_turks\_anatolian
- 20\_bands\_genres\_metalcore
- 21\_mother\_motherfucker\_mothers
- 22\_racist\_racists\_racial
- 23\_mohammed\_islamist\_jihad
- 24\_episodes\_spongebob\_nickelodeon
- 25\_talkpage\_page\_talk
- 26\_images\_photos\_jpg
- 27\_kashmir\_hindu\_hindus
- 28\_shut\_shutdown\_shutting
- 29\_cunt\_cunts\_cuntface



# Topic Modeling

 **Results: Identity Hate**

Model	Coherence	Semantic Diversity
HDBSCAN	0.3399	0.4352
k-means (k = 10)	0.3778	0.3913
k-means (k = 5)	0.3670	0.3341



# Topic Modeling

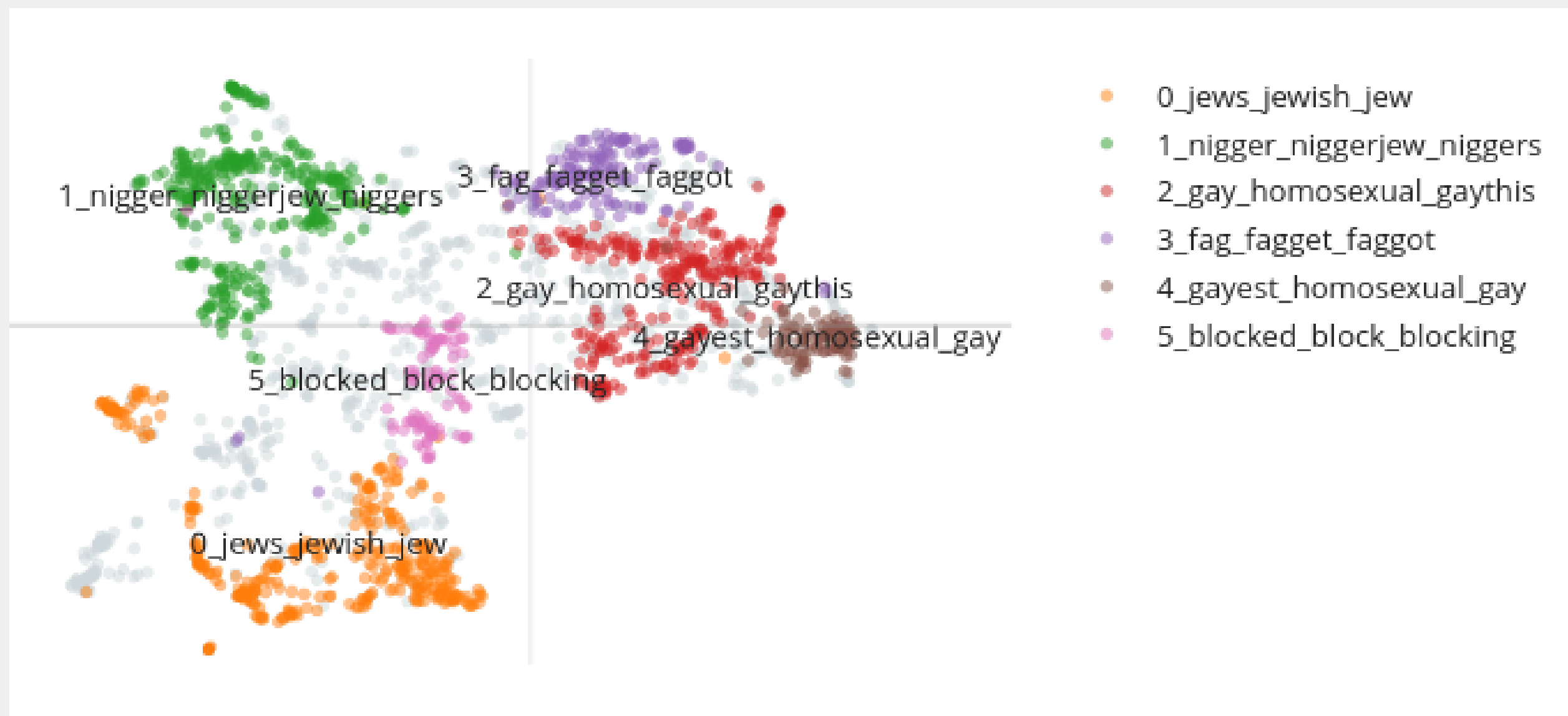
## BERTopic with HDBSCAN (Identity Hate)

Topic Word Scores



# Topic Modeling

## BERTopic with HDBSCAN (Identity Hate)



# Strengths and Limitations

## Text Classification:

- Tuned **Logistic Regression** and **SGD** emerged as the strongest statistical models
  - SGD offered faster training, while Logistic Regression achieved the best overall performance
- **DistilBERT** consistently outperformed classical models, especially on context-dependent and **rare toxicity labels**



- **Limitation:** single-model approaches were used

## Topic Modeling:

- **LDA** and **NMF** produced relatively decent topics but struggled with **semantic overlap**
- **BERTopic** better captured nuanced, diverse forms of toxic language
- **Limitation:** embedding-based models are more sensitive to noise and produce many outliers



# Conclusions



## Classification:

- **Deep Learning vs. Traditional:** contextual embeddings consistently outperformed lexical representations
- **Addressed Challenges:** successfully managed **multi-label** structure and **severe class imbalance**



## Topic Modeling:

- Revealed **recurring themes and nuanced patterns** in toxic comments beyond classification



## Future Developments:

- **Optimization:** ensemble methods and model stacking
- **Hybrid approaches:** combine classification with topic modeling
- **Scaling:** extension to multilingual datasets

**Thank you**