

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



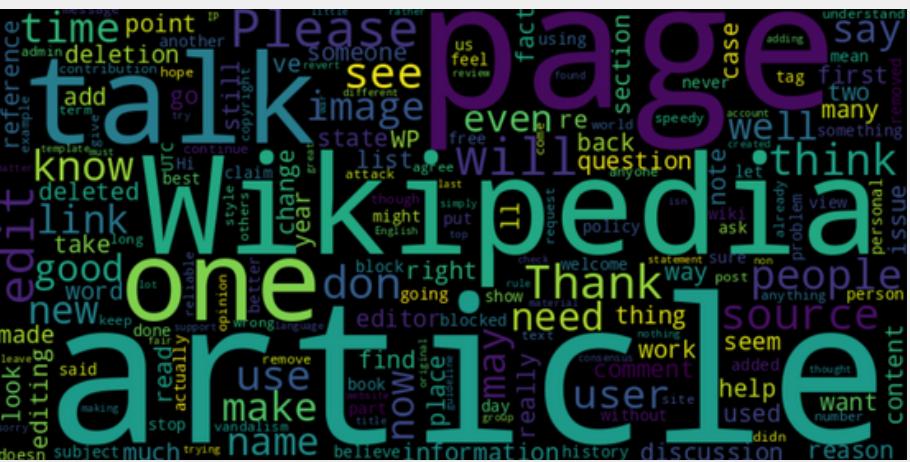
- Online platforms face increasing challenges in moderating **toxic and abusive** user-generated content at scale
- Toxicity detection goes beyond sentiment analysis, requiring understanding of **context, sarcasm, and identity-based hate**
- Manual moderation is **not scalable**, motivating automated 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)

Introduction

Dataset and EDA



- **Dataset:** Wikipedia talk pages' comments
(Train: **159,571**, Test: **153,164**)
 - **Labels:** *toxic, severe_toxic, obscene, threat, insult, identity_hate*
 - Severe **class imbalance**: e.g. *threat, identity_hate*
 - **Data Cleaning:**
 - *Classification*: Removed ambiguous labels (**-1**) from **test set (63,978 samples)**
 - *Topic Modeling*: Combined train and test sets (**22,468** samples for toxic-only subset, while **2,117** for *identity_hate*)



- 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

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%

Preprocessing

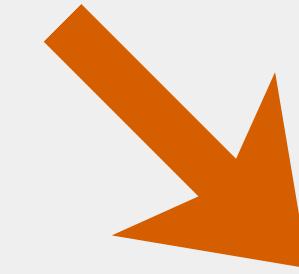
General cleaning:

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

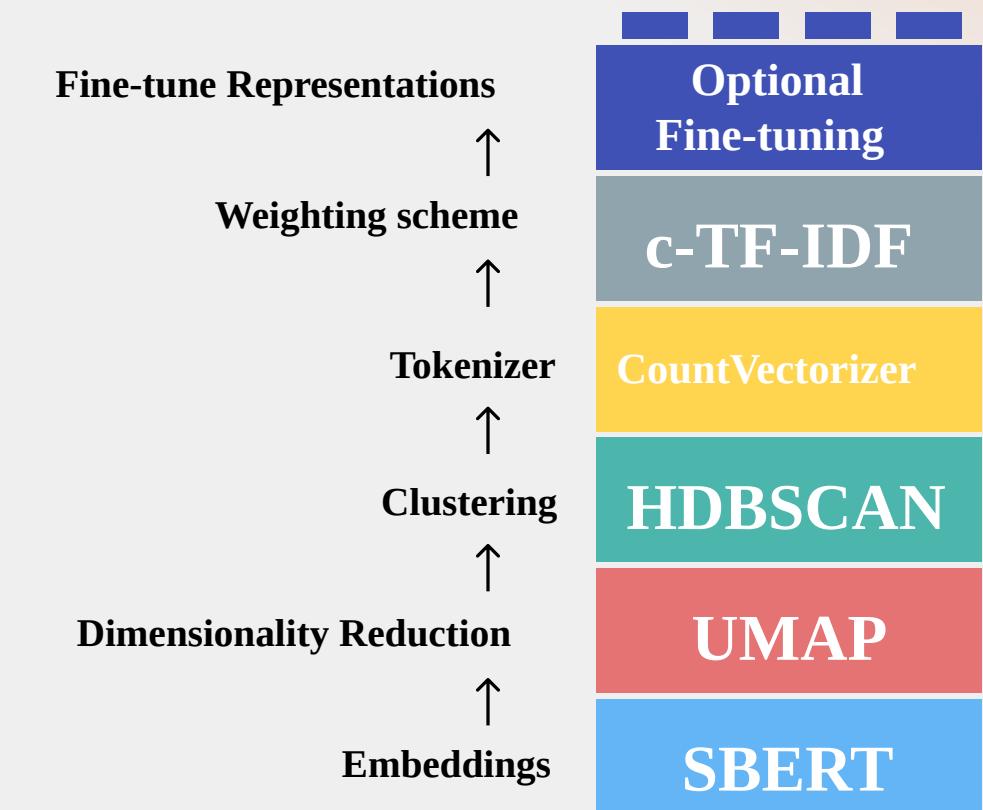
Test set

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**)



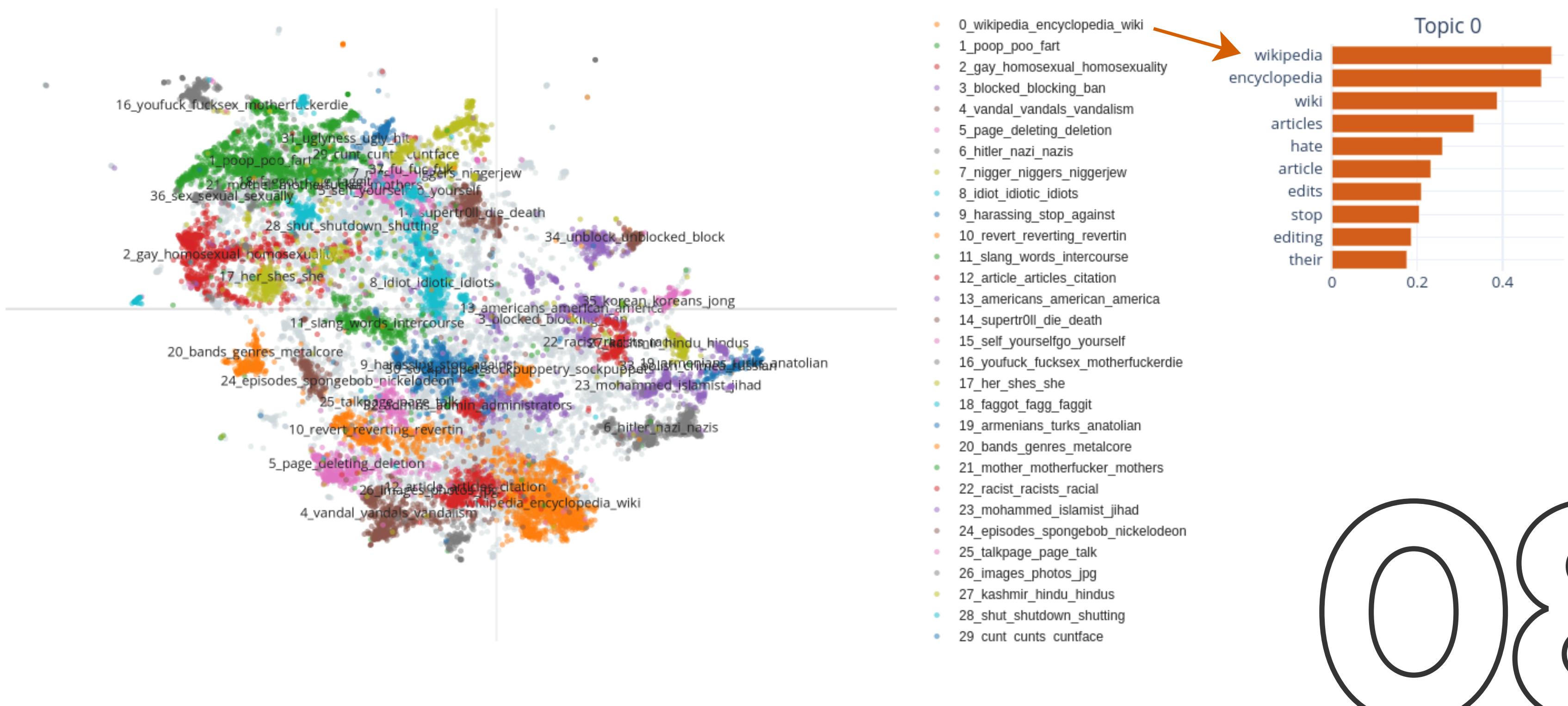
Topic Modeling

☠️ 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

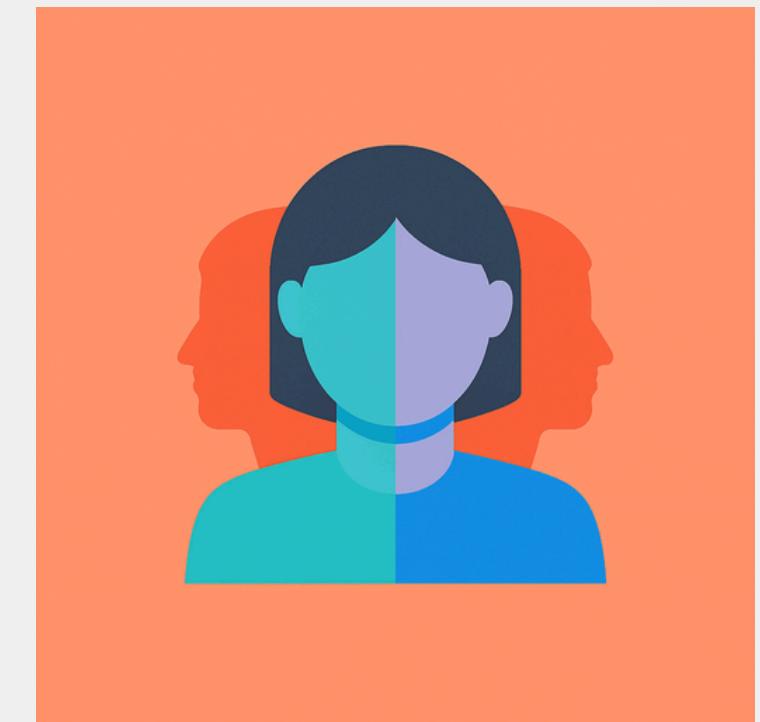
BERTopic with HDBSCAN



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

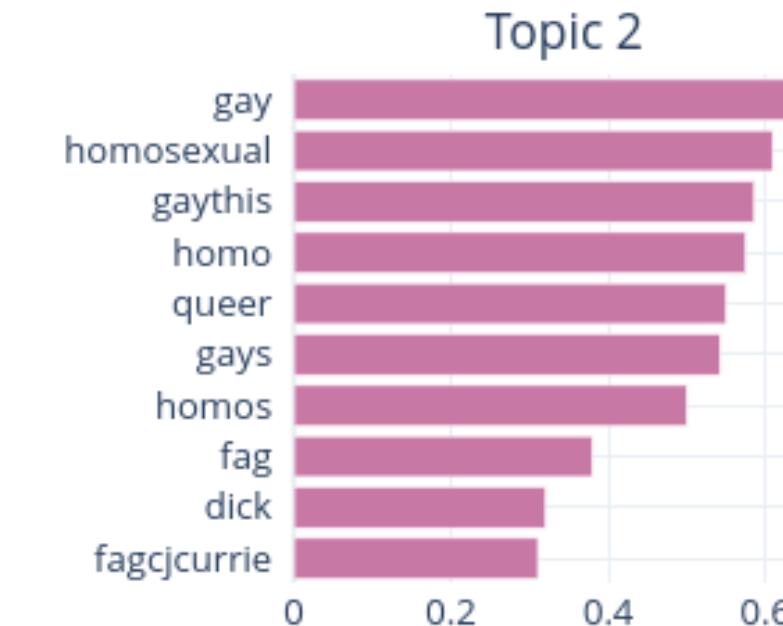
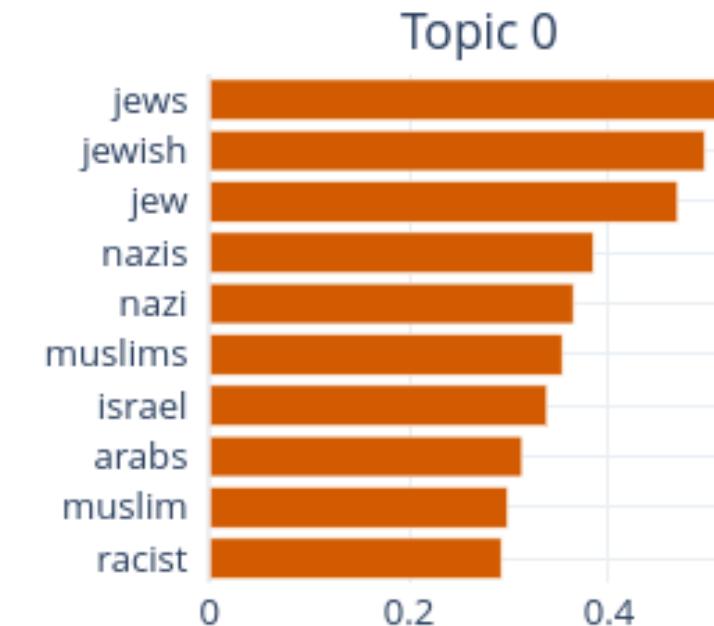


09

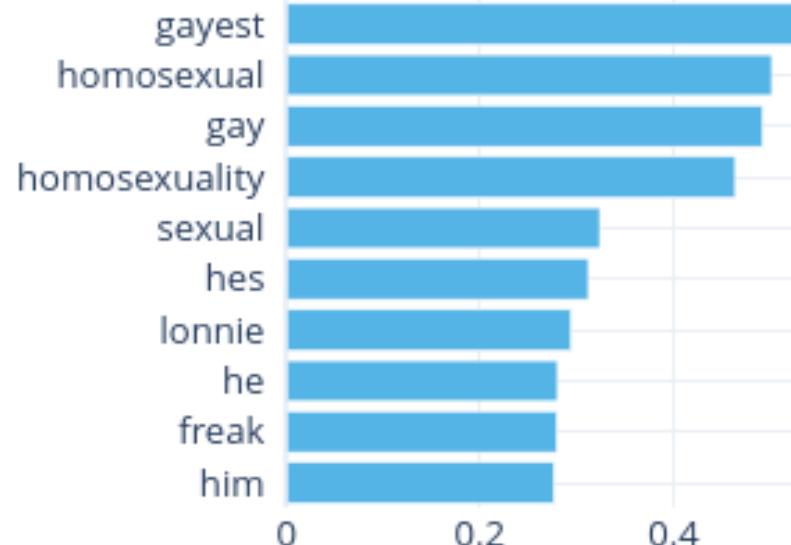
Topic Modeling

BERTopic with HDBSCAN (Identity Hate)

Topic Word Scores



Topic 4

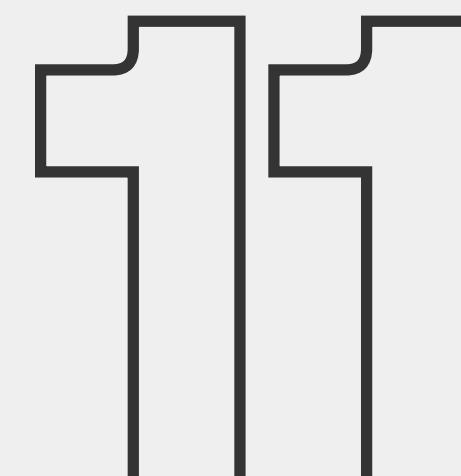
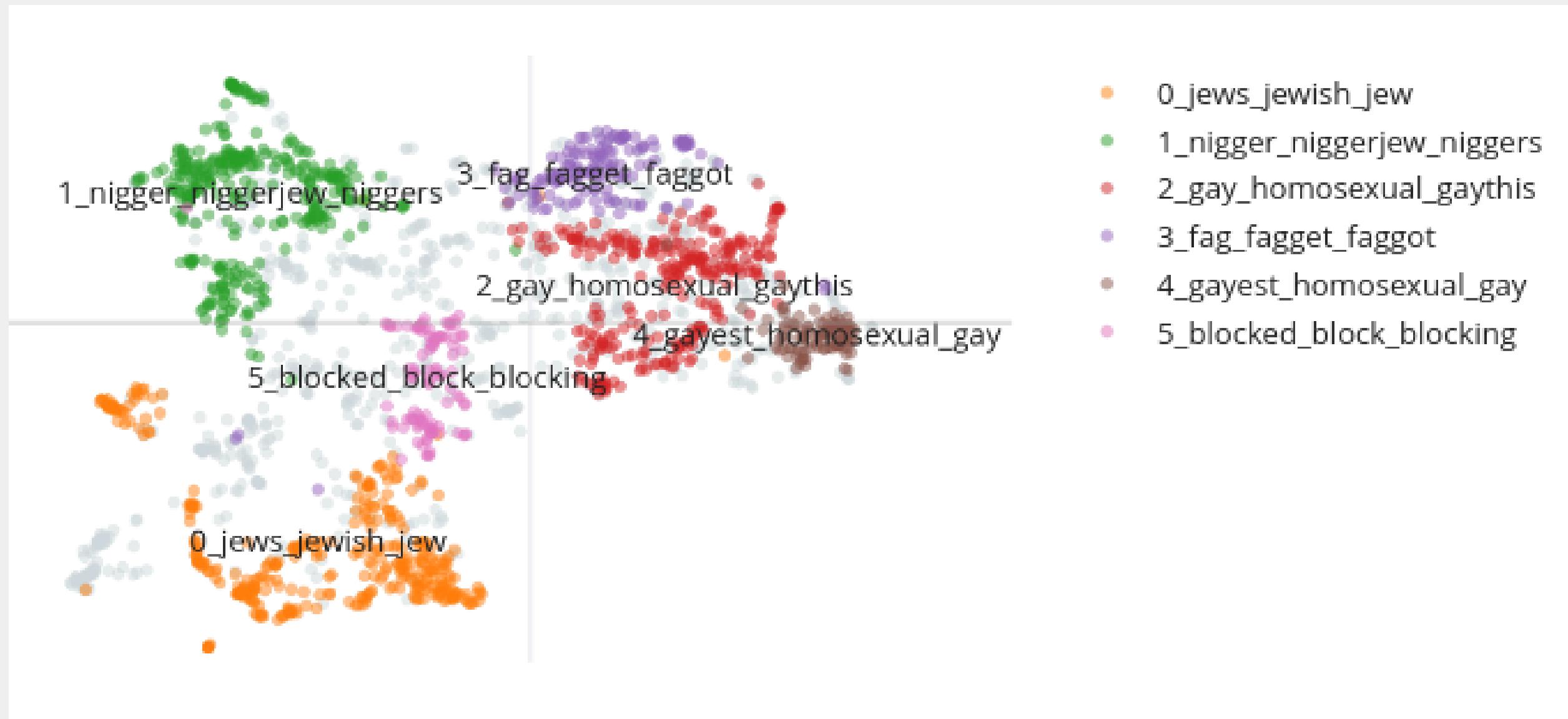


Topic 5



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



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