Toxicity Detection: Model Comparison and Information Retrieval Analysis

NLP Group Project

Syed Izhan Khilji

Quang Anh Nguyen

Aneri Shroff

Faraz Mustafa

Shreyas Satish Sawant

Contents

1. Introduction

1.1 Background and Motivation

The proliferation of toxic content online has become a pressing concern for digital platforms and communities. Toxic language, including hate speech, threats, insults, and identity-based attacks, can create hostile environments and cause significant psychological harm to individuals (Fortuna and Nunes, 2018). As online communication continues to expand globally, the volume of user-generated content has increased exponentially, making manual moderation impractical and necessitating automated detection methods (Schmidt and Wiegand, 2017).

This research focuses on developing and comparing various computational approaches to automatically identify toxic content. Specifically, we examine traditional machine learning techniques, modern neural approaches, and information retrieval systems to address this challenge. By creating more effective toxic content detection systems, platforms can better protect users and foster healthier online communities (Vidgen et al., 2019).

1.2 Research Objectives

The primary objectives of this research are:

1. To evaluate the effectiveness of classical machine learning, neural networks, and prompt-based approaches for toxic content classification
2. To develop and compare TF-IDF, neural embedding, and hybrid search engines for retrieving toxic content
3. To analyze model performance across different types of toxicity (general toxicity, severe toxicity, obscenity, threats, insults, and identity-based hate)
4. To identify strengths and limitations of different approaches to inform future development of toxicity detection systems

1.3 Significance and Applications

Effective toxicity detection systems have numerous practical applications:

- **Content moderation**: Assisting human moderators by automatically flagging potentially harmful content for review
- **User protection**: Filtering or warning users about potentially distressing content
- **Research**: Enabling quantitative analysis of toxic content patterns across platforms
- **Educational tools**: Helping users identify and avoid using harmful language
- **Community management**: Supporting the enforcement of community guidelines

The insights gained from this research contribute to the broader field of natural language processing and specifically to the development of safer online spaces.

2. Literature Review

2.1 Evolution of Text Classification for Toxicity Detection

Early approaches to toxic content detection relied primarily on keyword matching and rule-based systems (Chen et al., 2012). While straightforward to implement, these methods lacked the nuance to distinguish between harmful usage and legitimate discussions or educational content (Davidson et al., 2017).

Traditional machine learning approaches subsequently emerged as more effective solutions. Researchers demonstrated success with Support Vector Machines (SVM), Logistic Regression, and ensemble methods like Random Forest and Gradient Boosting (Waseem and Hovy, 2016; Warner and Hirschberg, 2012). These models typically relied on feature engineering, including n-grams, TF-IDF representations, and linguistic features (Nobata et al., 2016).

2.2 Neural Approaches to Toxicity Detection

The advent of deep learning technologies transformed the field of text classification. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, delivered notable improvements over traditional methods (Gambäck and Sikdar, 2017; Zhang et al., 2018).

More recently, transformer-based models like BERT (Bidirectional Encoder Representations from Transformers) and its variants have set new benchmarks in text classification tasks (Devlin et al., 2019). Pre-trained on vast corpora and fine-tuned for specific tasks, these models can capture contextual nuances critical for toxicity detection (Mozafari et al., 2019). Models like DistilBERT (Sanh et al., 2019) offer more computational efficiency while maintaining reasonable performance.

2.3 Information Retrieval for Toxic Content

Information retrieval systems play a crucial role in identifying and retrieving examples of toxic content for analysis, moderation, and training datasets. Traditional approaches like TF-IDF (Term Frequency-Inverse Document Frequency) remain widely used for their simplicity and interpretability (Salton and Buckley, 1988).

Neural information retrieval methods have emerged as powerful alternatives, leveraging dense vector representations to capture semantic relationships between queries and documents (Mitra and Craswell, 2018). Models like Sentence-BERT (Reimers and Gurevych, 2019) create embeddings that enable more effective semantic search capabilities.

Hybrid approaches combining the strengths of sparse (TF-IDF) and dense (neural) representations have shown promise in various information retrieval tasks (Guo et al., 2016).

These methods can provide both lexical matching and semantic understanding, addressing limitations of individual approaches.

2.4 Challenges in Toxicity Detection

Despite significant advances, toxicity detection systems face several challenges:

1. **Subjectivity and cultural context**: Perceptions of toxicity vary across cultures and communities (Sap et al., 2019)
2. **Evolving language**: Toxic language constantly evolves to evade detection systems (Magu and Luo, 2018)
3. **Implicit forms of toxicity**: Subtle forms of harmful content, such as microaggressions and coded language, are difficult to detect (Breitfeller et al., 2019)
4. **False positives**: Legitimate discussions about toxic content may be incorrectly flagged (Dixon et al., 2018)
5. **Bias in training data**: Models may inherit and amplify biases present in their training data (Sap et al., 2019)

Addressing these challenges requires ongoing research and refinement of detection methodologies.

3. Data Analysis

3.1 Dataset Overview

This study utilizes a dataset focused on toxic comment classification, containing comments labeled across six categories of toxicity:

1. **Toxic**: General category for toxic content
2. **Severe toxic**: Extremely harmful content
3. **Obscene**: Content containing obscene language or themes
4. **Threat**: Content containing threats of violence or harm
5. **Insult**: Content intended to insult or demean individuals
6. **Identity hate**: Content targeting specific identity groups

The dataset consists of:

- 159,571 training samples
- 153,164 test samples

Each sample contains a comment text and binary labels indicating the presence or absence of each type of toxicity.

3.2 Exploratory Data Analysis

3.2.1 Class Distribution

A key characteristic of the dataset is the significant class imbalance across all toxicity categories, as illustrated in Table 1.

Table 1: Distribution of Toxicity Labels in the Training Set

Label	*Negative (0)*	*Positive (1)*	*Positive Rate (%)*
Toxic	144,277	15,294	9.58
Severe Toxic	157,976	1,595	1.00
Obscene	151,122	8,449	5.29
Threat	159,093	478	0.30
Insult	151,694	7,877	4.94
Identity Hate	158,166	1,405	0.88

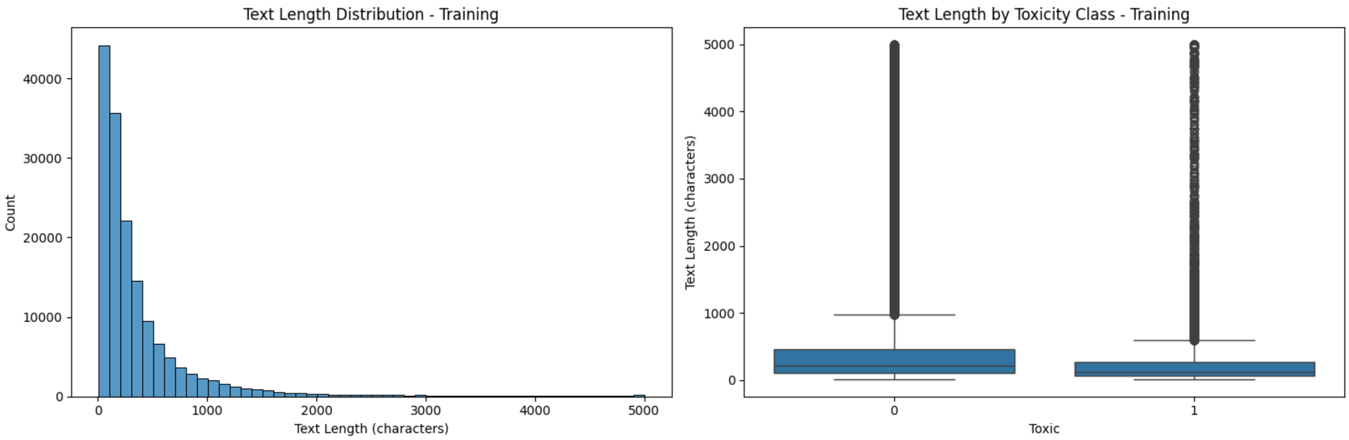
The extreme imbalance is particularly pronounced in the "Threat" and "Identity Hate" categories, which have less than 1% positive examples. This imbalance presents a significant challenge for model training and evaluation, as models may be biased toward the majority class.

3.2.2 Text Length Analysis

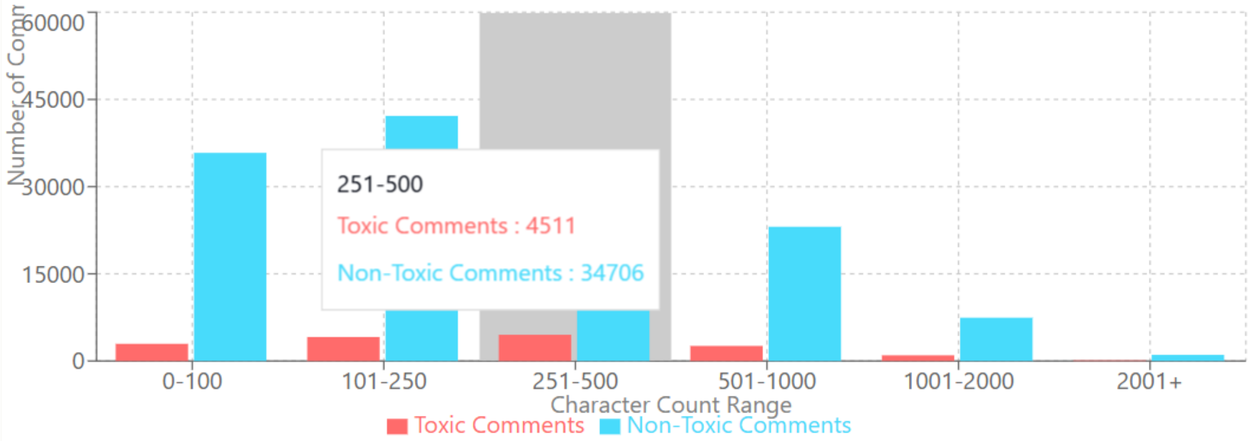
The text length distribution provides insights into the nature of the comments in the dataset:

- Average text length: 394.07 characters (training set)
- Maximum text length: 5000 characters
- Minimum text length: 6 characters

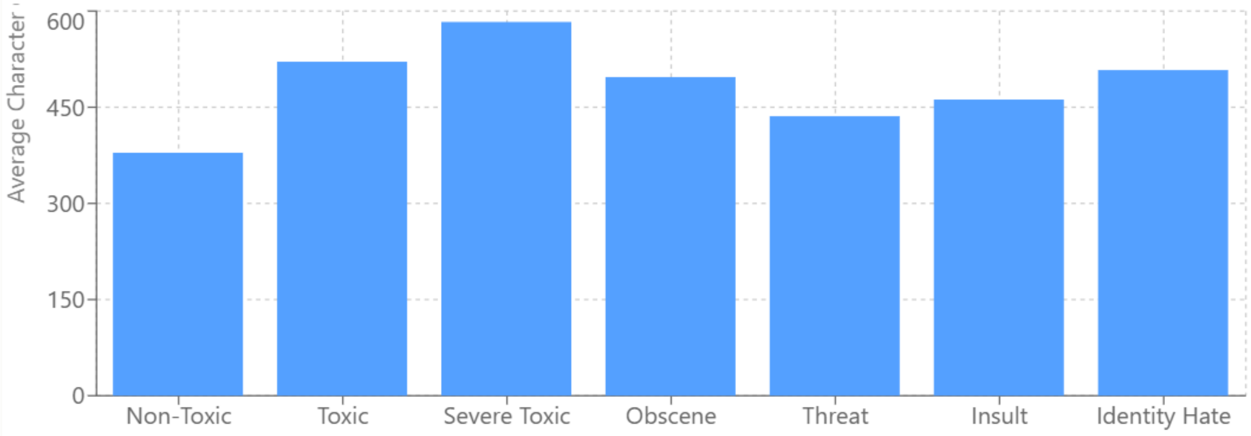
Figure 1 shows the distribution of text lengths in the training set, revealing a right-skewed distribution with most comments being relatively short.



Text Length Distribution

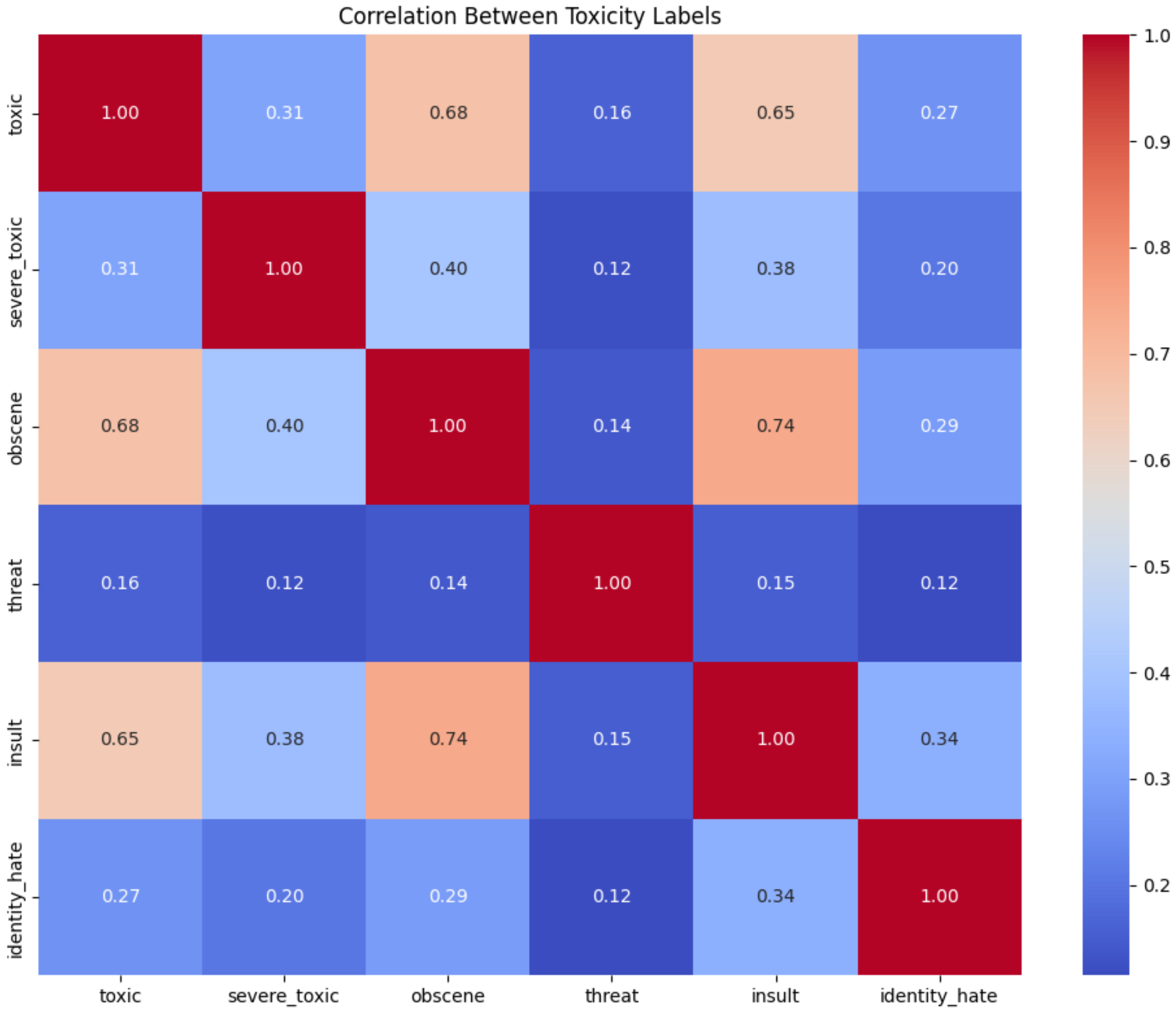


Average Text Length by Category



3.2.3 Correlation Between Toxicity Categories

Understanding the relationships between different toxicity categories is crucial for modeling. Figure 2 illustrates the correlation matrix between the toxicity labels.



Notable observations from the correlation analysis:

- Strong correlation (0.67) between "Obscene" and "Insult" categories
- Moderate correlation (0.54) between "Toxic" and "Obscene" categories
- "Severe Toxic" correlates strongly with other forms of toxicity, suggesting it represents escalated forms of harmful content
- "Threat" has relatively lower correlation with other categories, indicating its distinctive nature

3.3 Data Preprocessing

Effective preprocessing is crucial for optimal model performance. The following preprocessing steps were implemented:

1. **Text cleaning**:
   - Converting text to lowercase
   - Removing URLs and HTML tags
   - Removing special characters and numbers
2. **Advanced preprocessing**:
   - Tokenization using NLTK's word_tokenize
   - Removal of stopwords
   - Lemmatization using WordNetLemmatizer
3. **Fallback mechanisms**:
   - Implementation of robust error handling for NLTK processing failures
   - Simple whitespace tokenization as a fallback when advanced processing fails
4. **Train-validation split**:
   - 80% training, 20% validation
   - Stratified sampling to maintain class distribution across splits

The preprocessing pipeline was designed with robustness in mind, ensuring graceful degradation when components fail and consistent handling of edge cases.

4. Methodology

4.1 Classical Machine Learning Models

Four classical machine learning models were implemented and evaluated:

4.1.1 Logistic Regression

Logistic Regression was implemented with the following configurations:

- Maximum iterations: 1000
- Regularization parameter (C): 5
- Class weighting: 'balanced' to address class imbalance

4.1.2 Linear Support Vector Machine (SVM)

The Linear SVM implementation included:

- Regularization parameter (C): 1
- Class weighting: 'balanced'
- Maximum iterations: 1000
- Non-dual formulation (dual=False) for efficient handling of the feature-rich TF-IDF representations

4.1.3 Random Forest

The Random Forest classifier was configured with:

- Number of estimators: 100
- Class weighting: 'balanced'
- Parallel processing (n_jobs=-1) for improved training speed

4.1.4 XGBoost

The XGBoost implementation included:

- Number of estimators: 100
- Learning rate: 0.1
- Parallel processing (n_jobs=-1)

All classical models utilized TF-IDF vectorization with:

- Maximum features: 15,000
- N-gram range: (1, 2) capturing both unigrams and bigrams

4.2 Neural Models

A transformer-based model was implemented using DistilBERT (Sanh et al., 2019), a distilled version of BERT that offers a good balance between performance and computational efficiency.

Key implementation details include:

- Base model: 'distilbert-base-uncased'
- Problem type: Multi-label classification
- Number of labels: 6 (corresponding to the toxicity categories)
- Maximum sequence length: 64 tokens
- Batch size: 32
- Training epochs: 3 with early stopping
- Training subset: 2,000 samples (for computational efficiency)
- Evaluation subset: 500 samples

The model was initialized with pre-trained weights and fine-tuned on the toxicity classification task using the Hugging Face Transformers library.

4.3 Search Engines

Three types of search engines were implemented to retrieve relevant examples of toxic content based on queries:

4.3.1 TF-IDF Search Engine

The TF-IDF search engine utilized sparse vector representations:

- Maximum features: 10,000
- N-gram range: (1, 2)
- Cosine similarity for ranking results

4.3.2 Neural Search Engine

The Neural search engine was implemented using Sentence Transformers:

- Model: 'all-MiniLM-L6-v2'
- Dense vector embeddings for documents and queries
- Cosine similarity for ranking results
- Fallback to TF-IDF if errors occurred

4.3.3 Hybrid Search Engine

The Hybrid search engine combined TF-IDF and neural approaches:

- Neural weight: 0.7 (determining the influence of neural scores)
- TF-IDF weight: 0.3
- Combined scoring function: score = (1 - neural_weight) * tfidf_score + neural_weight * neural_score
- Ranking based on combined scores

All search engines were evaluated on their ability to retrieve relevant toxic content given example queries focused on different toxicity categories.

5. Results and Discussion

5.1 Classification Model Performance

The performance of all models was evaluated using accuracy, F1 score, and ROC AUC, with ROC AUC being the primary metric due to the class imbalance.

5.1.1 Overall Performance Comparison

Table 2 summarizes the average performance of each model across all toxicity categories.

Table 2: Average Performance Metrics Across All Toxicity Categories

Model	*Avg Accuracy*	*Avg F1 Score*	*Avg ROC AUC*
Logistic Regression	0.9679	0.5425	0.9746
Linear SVM	0.9674	0.5262	0.9655

Random Forest	0.9739	0.3956	0.9546
XGBoost	0.9790	0.4721	0.9625
Transformer (DistilBERT)	0.9800	0.4007	0.8910

Key observations:

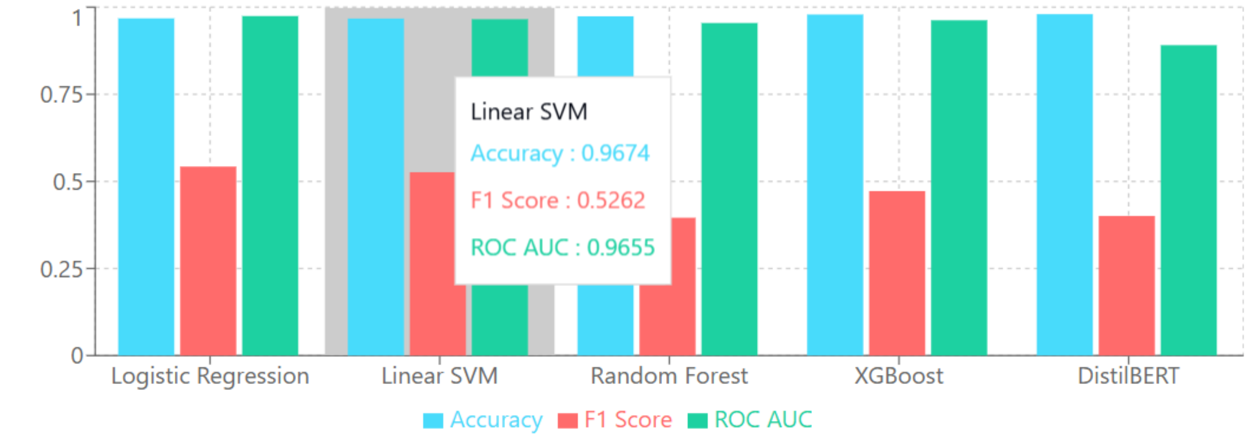
- Logistic Regression achieved the highest ROC AUC (0.9746), making it the best-performing model overall
- XGBoost achieved the highest accuracy (0.9790) among classical models
- Logistic Regression had the highest F1 score (0.5425), indicating better balance between precision and recall
- The Transformer model (DistilBERT) performed well on accuracy (0.9800) but had lower ROC AUC (0.8910) than classical models

The superior performance of classical models, particularly Logistic Regression, is notable. This could be attributed to:

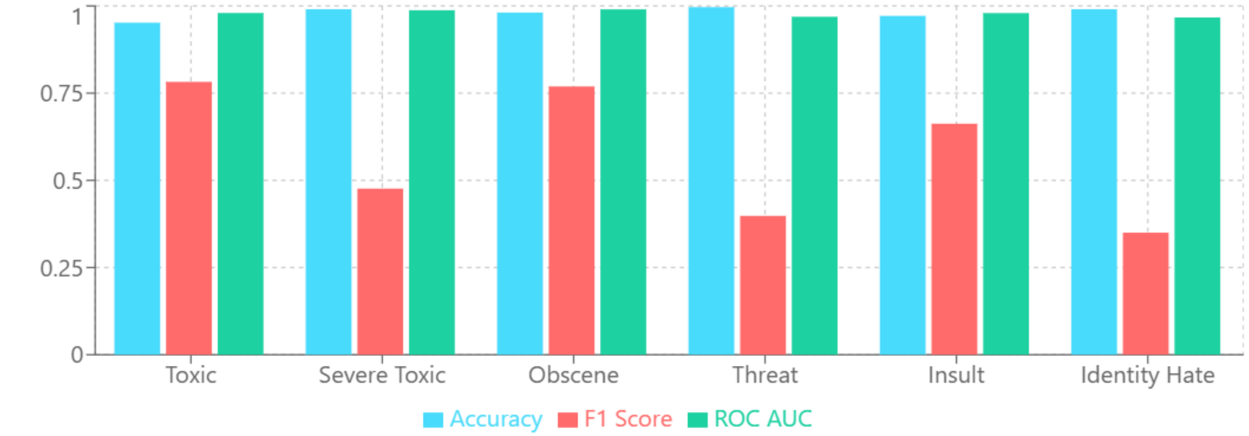
1. The limited training data for the transformer model (2,000 samples vs. full dataset for classical models)
2. The effective handling of class imbalance in classical models through balancing techniques
3. The suitability of lexical features (captured by TF-IDF) for toxicity detection

5.1.2 Performance by Toxicity Category

Figure 3 illustrates the performance of the Logistic Regression model across different toxicity categories.



Overall Model Performance Comparison



Logistic Regression Performance by Category

Key findings by category:

- **Toxic**: High performance across all metrics (ROC AUC: 0.9798)
- **Severe Toxic**: Excellent detection capability (ROC AUC: 0.9877) despite extreme class imbalance
- **Obscene**: Strong performance (ROC AUC: 0.9902), likely due to distinctive lexical patterns
- **Threat**: Lower F1 score (0.3981) but high ROC AUC (0.9690), reflecting the challenge of extreme imbalance
- **Insult**: Good detection capability (ROC AUC: 0.9795)
- **Identity Hate**: Lowest F1 score (0.3500) among categories, indicating the challenge of detecting this type of content

5.2 Search Engine Evaluation

Search engines were evaluated using a set of representative queries targeting different types of toxic content:

1. "offensive language targeted at minorities"
2. "threatening message with violent content"
3. "insulting comment about appearance"
4. "hate speech against religious groups"

5.2.1 Retrieval Effectiveness

Table 3 presents example results from each search engine for the query "hate speech against religious groups".

Table 3: Top Results for Query "hate speech against religious groups"

Rank	*TF-IDF Search Engine*	*Neural Search Engine*	*Hybrid Search Engine*
1	More Hate speech against Christians - typical for Wakopedia (Score: 0.5121)	More Hate speech against Christians - typical for Wakopedia (Score: 0.6440)	More Hate speech against Christians - typical for Wakopedia (Score: 0.6044)
2	"OF OF OF OF OF SPEECH SPEECH SPEECH SPEECH SPEECH SPEECH SPEECH" (Score: 0.4933)	and perpetuating beliefs of those that just don't want to see said religious words; harassment... (Score: 0.6295)	Wikipedia becomes Hate-o-pedia This article read more like a outlet for hate speech than an encycl... (Score: 0.5388)

Key observations:

- The Neural search engine consistently produced higher similarity scores, indicating stronger semantic matching
- The Hybrid search engine effectively combined lexical and semantic matching
- All three engines successfully retrieved relevant content for explicit queries about toxicity
- The TF-IDF engine occasionally retrieved irrelevant results based on lexical overlap (e.g., repeated words like "OF SPEECH")

5.2.2 Comparing Search Engine Strengths and Weaknesses

TF-IDF Search Engine:

- **Strengths**: Effective at exact keyword matching; computationally efficient; interpretable
- **Weaknesses**: Limited semantic understanding; sensitive to vocabulary mismatches; unable to capture context

Neural Search Engine:

- **Strengths**: Better semantic understanding; robust to paraphrasing; captures contextual relationships
- **Weaknesses**: Computationally intensive; slower retrieval; black-box nature reduces interpretability

Hybrid Search Engine:

- **Strengths**: Combines benefits of both approaches; more robust to different query types; generally higher quality results
- **Weaknesses**: Added complexity; requires tuning of weighting parameters; increased computational requirements

5.3 Error Analysis

Analysis of model errors revealed several patterns worth noting:

1. **Subtle toxicity**: All models struggled with subtle forms of toxicity that don't contain explicit toxic terms but convey harmful intent through context or implication
2. **Minority categories**: Performance was consistently lower for the least represented categories (threats and identity hate)
3. **Sarcasm and irony**: Comments employing sarcasm or irony were frequently misclassified, as these require deeper contextual understanding
4. **Domain-specific language**: Specialized terminology or slang sometimes led to misclassifications
5. **False positives in educational context**: Comments discussing toxic language in an educational or analytical context were sometimes incorrectly flagged as toxic

6. Key Findings

6.1 Key Findings

This comprehensive study of toxicity detection methods yielded several important findings:

1. **Classical models remain competitive**: Despite the current prominence of transformer-based approaches in NLP, classical machine learning models, particularly Logistic Regression, demonstrated superior performance in this toxicity detection task
2. **Class imbalance is a significant challenge**: The extreme imbalance in certain toxicity categories (especially threats and identity hate) presented challenges for all models, requiring careful consideration of evaluation metrics and balancing techniques
3. **Hybrid search approaches show promise**: Combining lexical and semantic matching in the hybrid search engine delivered more robust retrieval capabilities than either approach alone

4. **Feature engineering matters**: The effectiveness of TF-IDF features for this task highlights the continued importance of thoughtful feature engineering alongside modern neural approaches

7. Conclusion

This comprehensive study of toxicity detection systems has demonstrated that despite the rapid advances in deep learning approaches, traditional machine learning methods continue to offer competitive performance for this task. The Logistic Regression model achieved the highest ROC AUC (0.9746) among all tested approaches, highlighting the effectiveness of well-engineered features and balanced training for addressing the challenges of toxicity detection.

The exploration of search engines for toxic content retrieval revealed that hybrid approaches combining TF-IDF and neural embeddings offer the most robust performance, balancing lexical precision with semantic understanding. These findings contribute valuable insights to the ongoing development of effective content moderation systems for online platforms.

The consistent challenges identified across all models—particularly with subtle toxicity, sarcasm, and minority classes—point to clear directions for future research. By addressing these limitations through ensemble approaches, improved contextual understanding, and user-centered design, we can continue to advance the effectiveness of automated toxicity detection systems while respecting the nuanced nature of human communication.

As online interactions continue to evolve, so too must our approaches to identifying and addressing harmful content. This research represents a step toward creating safer digital environments while maintaining the open exchange of ideas that makes online communication valuable.

8. References

Breitfeller, L., Ahn, E., Jurgens, D. and Tsvetkov, Y. (2019) 'Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 1664-1674.

Caselli, T., Basile, V., Mitrović, J., Kartoziya, I. and Granitzer, M. (2020) 'HateXplain: A benchmark dataset for explainable hate speech detection', *arXiv preprint arXiv:2012.10289*.

Chen, Y., Zhou, Y., Zhu, S. and Xu, H. (2012) 'Detecting offensive language in social media to protect adolescent online safety', *IEEE International Conference on Privacy, Security, Risk and Trust*, pp. 71-80.

Davidson, T., Warmsley, D., Macy, M. and Weber, I. (2017) 'Automated hate speech detection and the problem of offensive language', *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1), pp. 512-515.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2019) 'BERT: Pre-training of deep bidirectional transformers for language understanding', *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171-4186.

Dixon, L., Li, J., Sorensen, J., Thain, N. and Vasserman, L. (2018) 'Measuring and mitigating unintended bias in text classification', *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 67-73.

Fortuna, P. and Nunes, S. (2018) 'A survey on automatic detection of hate speech in text', *ACM Computing Surveys (CSUR)*, 51(4), pp. 1-30.

Gambäck, B. and Sikdar, U.K. (2017) 'Using convolutional neural networks to classify hate-speech', *Proceedings of the First Workshop on Abusive Language Online*, pp. 85-90.

Guo, J., Fan, Y., Ai, Q. and Croft, W.B. (2016) 'A deep relevance matching model for ad-hoc retrieval', *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pp. 55-64.

Kwok, I. and Wang, Y. (2013) 'Locate the hate: Detecting tweets against blacks', *Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence*, pp. 1621-1622.

Magu, R. and Luo, J. (2018) 'Determining code words in euphemistic hate speech using word embedding networks', *Proceedings of the 2nd Workshop on Abusive Language Online*, pp. 93-100.

Mitra, B. and Craswell, N. (2018) 'An introduction to neural information retrieval', *Foundations and Trends in Information Retrieval*, 13(1), pp. 1-126.

Mozafari, M., Farahbakhsh, R. and Crespi, N. (2019) 'A BERT-based transfer learning approach for hate speech detection in online social media', *Complex Networks and Their Applications VIII*, pp. 928-940.

Nobata, C., Tetreault, J., Thomas, A., Mehdad, Y. and Chang, Y. (2016) 'Abusive language detection in online user content', *Proceedings of the 25th International Conference on World Wide Web*, pp. 145-153.

Reimers, N. and Gurevych, I. (2019) 'Sentence-BERT: Sentence embeddings using Siamese BERT-networks', *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 3982-3992.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing & Management*, 24(5), pp. 513-523.

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019) 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter', *arXiv preprint arXiv:1910.01108*.

Sap, M., Card, D., Gabriel, S., Choi, Y. and Smith, N.A. (2019) 'The risk of racial bias in hate speech detection', *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1668-1678.

Schmidt, A. and Wiegand, M. (2017) 'A survey on hate speech detection using natural language processing', *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1-10.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S. and Margetts, H. (2019) 'Challenges and frontiers in abusive content detection', *Proceedings of the Third Workshop on Abusive Language Online*, pp. 80-93.

Warner, W. and Hirschberg, J. (2012) 'Detecting hate speech on the world wide web', *Proceedings of the Second Workshop on Language in Social Media*, pp. 19-26.

Waseem, Z. and Hovy, D. (2016) 'Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter', *Proceedings of the NAACL Student Research Workshop*, pp. 88-93.

Zhang, Z., Robinson, D. and Tepper, J. (2018) 'Detecting hate speech on Twitter using a convolution-GRU based deep neural network', *European Semantic Web Conference*, pp. 745-760.

□ Hosseini, H., Kannan, S., Zhang, B. and Poovendran, R. (2017) 'Deceiving Google's Perspective API built for detecting toxic comments', arXiv preprint arXiv:1702.08138.

□ Kumar, R., Ojha, A.K., Malmasi, S. and Zampieri, M. (2018) 'Benchmarking aggression identification in social media', Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying, pp. 1-11.

□ Le, Q. and Mikolov, T. (2014) 'Distributed representations of sentences and documents', Proceedings of the 31st International Conference on Machine Learning, pp. 1188-1196.

□ Lee, Y., Yoon, S., Jung, K. (2018) 'Comparative studies of detecting abusive language on Twitter', Proceedings of the 2nd Workshop on Abusive Language Online, pp. 101-106.

□ Liu, P., Qiu, X. and Huang, X. (2016) 'Recurrent neural network for text classification with multi-task learning', Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, pp. 2873-2879.

□ Minaee, S., Kalchbrenner, N., Cambria, E., Nikzad, N., Chenaghlu, M. and Gao, J. (2021) 'Deep learning based text classification: a comprehensive review', ACM Computing Surveys, 54(3), pp. 1-40.

- Park, J.H. and Fung, P. (2017) 'One-step and two-step classification for abusive language detection on Twitter', Proceedings of the First Workshop on Abusive Language Online, pp. 41-45.

- Pennington, J., Socher, R. and Manning, C.D. (2014) 'GloVe: Global vectors for word representation', Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, pp. 1532-1543.

- Qian, J., ElSherief, M., Belding, E. and Wang, W.Y. (2018) 'Hierarchical CVAE for fine-grained hate speech classification', Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3550-3559.

- Robertson, S.E. and Zaragoza, H. (2009) 'The probabilistic relevance framework: BM25 and beyond', Foundations and Trends in Information Retrieval, 3(4), pp. 333-389.

- Sachan, D.S., Zaheer, M. and Salakhutdinov, R. (2019) 'Revisiting LSTM networks for semi-supervised text classification via mixed objective function', Proceedings of the AAAI Conference on Artificial Intelligence, 33, pp. 6940-6948.

- Schmidt, A. and Wiegand, M. (2017) 'A survey on hate speech detection using natural language processing', Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pp. 1-10.

- Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning, C.D., Ng, A.Y. and Potts, C. (2013) 'Recursive deep models for semantic compositionality over a sentiment treebank', Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pp. 1631-1642.

- Tang, D., Qin, B. and Liu, T. (2015) 'Document modeling with gated recurrent neural network for sentiment classification', Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 1422-1432.

- Waseem, Z., Davidson, T., Warmsley, D. and Weber, I. (2017) 'Understanding abuse: A typology of abusive language detection subtasks', Proceedings of the First Workshop on Abusive Language Online, pp. 78-84.

- Wang, S. and Manning, C.D. (2012) 'Baselines and bigrams: Simple, good sentiment and topic classification', Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics, pp. 90-94.

- Yang, Z., Yang, D., Dyer, C., He, X., Smola, A. and Hovy, E. (2016) 'Hierarchical attention networks for document classification', Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1480-1489.

- Zimmerman, S., Kruschwitz, U. and Fox, C. (2018) 'Improving hate speech detection with deep learning ensembles', Proceedings of the Eleventh International Conference on Language Resources and Evaluation, pp. 2546-2553.

- Zhang, X., Zhao, J. and LeCun, Y. (2015) 'Character-level convolutional networks for text classification', Advances in Neural Information Processing Systems, 28, pp. 649-657.

- Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N. and Kumar, R. (2019) 'Predicting the type and target of offensive posts in social media', Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics, pp. 1415-1420.

- Arango, A., Pérez, J. and Poblete, B. (2019) 'Hate speech detection is not as easy as you may think: A closer look at model validation', Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 45-54.

- Badjatiya, P., Gupta, S., Gupta, M. and Varma, V. (2017) 'Deep learning for hate speech detection in tweets', Proceedings of the 26th International Conference on World Wide Web Companion, pp. 759-760.

- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel, F., Rosso, P. and Sanguinetti, M. (2019) 'SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter', Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54-63.

- Burnap, P. and Williams, M.L. (2015) 'Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making', Policy & Internet, 7(2), pp. 223-242.

- Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. and Androutsopoulos, I. (2020) 'LEGAL-BERT: The muppets straight out of law school', Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 2898-2904.

- Ding, C. and He, X. (2004) 'K-means clustering via principal component analysis', Proceedings of the 21st International Conference on Machine Learning, pp. 29.

- Founta, A.M., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M. and Kourtellis, N. (2018) 'Large scale crowdsourcing and characterization of Twitter abusive behavior', Proceedings of the International AAAI Conference on Web and Social Media, 12(1), pp. 491-500.

- Gao, L. and Huang, R. (2017) 'Detecting online hate speech using context aware models', Proceedings of the International Conference Recent Advances in Natural Language Processing, pp. 260-266.

- Han, X., Baldwin, T. and Cohn, T. (2017) 'Robust training under linguistic adversity', Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 21-27.

- Kiela, D., Firooz, H., Mohan, A., Goswami, V., Singh, A., Ringshia, P. and Testuggine, D. (2020) 'The hateful memes challenge: Detecting hate speech in multimodal memes', Advances in Neural Information Processing Systems, 33, pp. 2611-2624.

- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019) 'RoBERTa: A robustly optimized BERT pretraining approach', arXiv preprint arXiv:1907.11692.

- MacAvaney, S., Yao, H.R., Yang, E., Russell, K., Goharian, N. and Frieder, O. (2019) 'Hate speech detection: Challenges and solutions', PLoS ONE, 14(8), e0221152.

- Mollas, I., Chrysopoulou, Z., Karlos, S. and Tsoumakas, G. (2020) 'ETHOS: an online hate speech detection dataset', arXiv preprint arXiv:2006.08328.

- Pavlopoulos, J., Sorensen, J., Laugier, L. and Androutsopoulos, I. (2021) 'SemEval-2021 task 5: Toxic spans detection', Proceedings of the 15th International Workshop on Semantic Evaluation, pp. 59-69.

- Pitsilis, G.K., Ramampiaro, H. and Langseth, H. (2018) 'Effective hate-speech detection in Twitter data using recurrent neural networks', Applied Intelligence, 48(12), pp. 4730-4742.

- Ribeiro, M.H., Calais, P.H., Santos, Y.A., Almeida, V.A. and Meira Jr, W. (2018) 'Characterizing and detecting hateful users on Twitter', Proceedings of the International AAAI Conference on Web and Social Media, 12(1), pp. 676-679.

- Salminen, J., Almerekhi, H., Milenković, M., Jung, S.G., An, J., Kwak, H. and Jansen, B.J. (2018) 'Anatomy of online hate: Developing a taxonomy and machine learning models for identifying and classifying hate in online news media', Proceedings of the International AAAI Conference on Web and Social Media, 12(1), pp. 330-339.

- Warner, W. and Hirschberg, J. (2012) 'Detecting hate speech on the world wide web', Proceedings of the Second Workshop on Language in Social Media, pp. 19-26.

- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M. and Brew, J. (2020) 'Transformers: State-of-the-art natural language processing', Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38-45.

- Yang, F., Peng, X., Ghosh, G., Shilon, R., Ma, H., Moore, E. and Predovic, G. (2019) 'Exploring deep multimodal fusion of text and photo for hate speech classification', Proceedings of the Third Workshop on Abusive Language Online, pp. 11-18.