Toxic Comments Classification

Nikita Pudov

May 2025

Abstract

The article describes the definition of toxic comments on the web. The paper compares classical machine learning approaches with Bert-based neural network approaches. The neural network approach shows the best results according to the F_1 score metric (0.7912).

Project link: `https://github.com/Npudov/OdsToxicCommentClassification`.

1 Introduction

A huge number of comments are written on modern social networks. Manual processing of these comments requires a lot of time and effort. There is a need for automated review of comments before they are published for compliance with the rules. Therefore, there is a need to identify toxic comments. This problem is solved by NLP methods.

1.1 Team

Nikita Pudov prepared this document and wrote code.

2 Related Work

Early research in toxic comment classification predominantly focused on traditional machine learning techniques combined with feature engineering. [Nobata et al., 2016] pioneered this domain by developing a hate speech detection system using linear support vector machines (SVMs) trained on lexical, syntactic, and semantic features extracted from online comments. Their work demonstrated the viability of machine learning for abuse detection but highlighted challenges in generalizing across diverse linguistic patterns. Subsequent studies expanded on these foundations, with [Khieu and Narwal,] achieving 92.7 % label accuracy using logistic regression and random forest models on TF-IDF vectorized text features. These approaches relied heavily on manual feature engineering, particularly through text preprocessing steps like lemmatization and stop-word removal. A systematic review [Andročec, 2020] analyzed 31 studies and found

that traditional methods consistently struggled with class imbalance and contextual understanding. For instance, models frequently misclassified identity-related terms (e.g., "Muslim") as inherently toxic, revealing biases in training data. This limitation spurred interest in more sophisticated feature representation techniques, including word2vec embeddings, though these still required careful curation of lexical resources [Andročec, 2020].

The field witnessed a paradigm shift with the adoption of deep neural networks capable of learning hierarchical text representations. [Georgakopoulos et al., 2018] achieved 91.2 % mean accuracy using convolutional neural networks (CNNs) to detect local semantic patterns in toxic comments. Recurrent architectures like LSTMs further improved performance by modeling sequential dependencies, with [Manav Kohli and Palowitch,] reaching 97.78 % accuracy through custom word embeddings. These models demonstrated superior handling of contextual nuances compared to bag-of-words approaches but remained computationally intensive.

Transformer-based models marked the next evolutionary stage, though their application is less documented in the reviewed literature. The survey [Andročec, 2020] noted emerging work on BERT variants fine-tuned for toxicity detection, which leverage self-attention mechanisms to capture long-range dependencies.

3 Model Description

From the beginning, classical machine learning methods TF-IDF with CatBoost and TF-IDF with logistic regression were used to identify toxic comments. Next, I used an approach based on the Bert neural network.

4 Dataset

I used Youtube Toxic Comment Dataset from Kaggle[1].

It has 1000 text comments from youtube and label IsToxic. On the Tab. 1 you can see class distribution for the mentioned dataset.

	Toxic	Non Toxic
Class	462	538

Table 1: Class distribution.

The distribution of comment lengths in the dataset is presented on Fig. 1.

I done preprocessing english text comments (lower case, tokenization + lemmatization) with spacy library for classic ML approaches (TF-IDF + Catboost, TF-IDF + Logistic Regression). The length of the comment text has also been trimmed to 400 characters for all approaches in this paper.

[1] Youtube Toxic Comment Dataset.

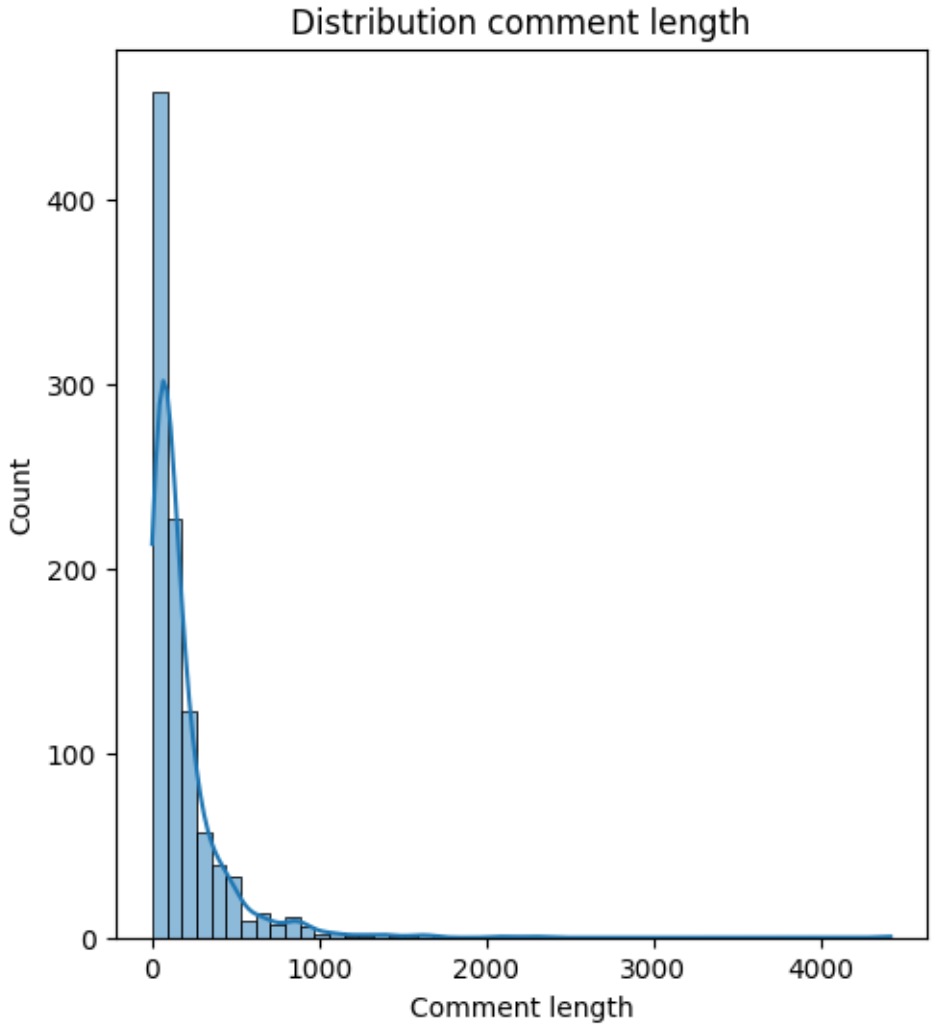

Figure 1: Comment Length Distribution.

On the Tab. 2 you can see the quantitative statistics train/test for the mentioned dataset.

	Train	Test
Comments quantitative	900	100

Table 2: Statistics of the Youtube Toxic Dataset.

5 Experiments

5.1 Metrics

Since the classification problem was solved, classical metrics (precision, recall, F_1 score) were used.

5.2 Experiment Setup

Firstly, i used TF-IDF with Catboost[2] pipeline. TF-IDF Vectorizer had hyperparameters $ngram\_range = (1, 2)$, $min\_df = 3$, $max\_df = 0.9$. Catboost had hyperparameters $iterations = 2000$, $depth = 6$.

Secondly, i used TF-IDF with Logistic Regression pipeline. TF-IDF Vectorizer had hyperparameters $ngram\_range = (1, 2)$, $min\_df = 3$, $max\_df = 0.9$. Logistic regression had hyperparameters $C = 0.9$,$penalty = "l2"$, $solver = "liblinear"$, $max\_iter = 1000$.

At the end, i used Bert model (bert-base-uncased)[3] with tokenizer ($padding = max\_length$, $truncation = True$, $max\_length = 128$). I trained Bert 10 epochs with $learning\_rate = 0.00002$, $weight\_decay = 0.01$.

5.3 Baselines

TF-IDF with Catboost and TF-IDF with Logistic Regression can be considered as a baseline.

6 Results

The results of our experiments to determine the toxicity of comments are presented on the Tab. 3

Method	Precision	Recall	F_1 score
TF-IDF + Catboost	0.7625	0.7536	0.7552
TF-IDF + Logistic Regression	0.7397	0.7351	0.7362
BERT	0.7912	0.8000	0.7826

Table 3: Results Toxic Comment Classification

The results show that the neural network approach based on the Bert demonstrates the best accuracy in terms of the F_1 score metric. In addition, it is easier to adjust the hyperparameters compared to Catboost. Logistic regression showed the worst result.

[2] Catboost.
[3] Bert model.

7 Conclusion

The paper reviewed classical ML approaches for determining the toxicity of comments and the neural network approach of Bert. The Bert neural network approach showed better results in classifying comments compared to the classical ones.

References

[Andročec, 2020] Andročec, D. (2020). Machine learning methods for toxic comment classification: a systematic review. *Acta Universitatis Sapientiae, Informatica*, 12(2):205–216.

[Georgakopoulos et al., 2018] Georgakopoulos, S. V., Tasoulis, S. K., Vrahatis, A. G., and Plagianakos, V. P. (2018). Convolutional neural networks for toxic comment classification. In *Proceedings of the 10th hellenic conference on artificial intelligence*, pages 1–6.

[Khieu and Narwal,] Khieu, K. and Narwal, N. *CS224N: Detecting and Classifying Toxic Comments*.

[Manav Kohli and Palowitch,] Manav Kohli, E. K. and Palowitch, J. *Paying attention to toxic comments online*.

[Nobata et al., 2016] Nobata, C., Tetreault, J. R., Thomas, A. O., Mehdad, Y., and Chang, Y. (2016). Abusive language detection in online user content. *Proceedings of the 25th International Conference on World Wide Web*.