Toxic Content Classification on Social Media Using ML

Kaylie Gupta

Analysis completed: January 2024
Repository uploaded: September 2025

Abstract

Social media is ever-present, especially among young users, and exposure to toxic content has been linked to negative mental health outcomes. This project explores whether machine learning (ML) can help filter toxic content on social platforms. Using publicly available NLP datasets labeled for toxicity, I evaluated five ML algorithms and achieved accuracy in the 73–88% range. These results indicate that ML models can effectively support personalized content filtering. Future work will focus on fine-grained labeling to enable more precise personalization and improved model performance

Introduction

Excessive social media use has been associated with a range of mental health issues, particularly among children and adolescents. My research aims to mitigate these effects by developing an ML-based content filter.

Specifically, I propose an NLP algorithm that classifies text from captions and comments into categories such as toxic, abusive, obscene, pro-eating disorder, or discriminatory. Users can specify sensitive topics or dislikes, allowing the model to filter content and promote safer online experiences

Background

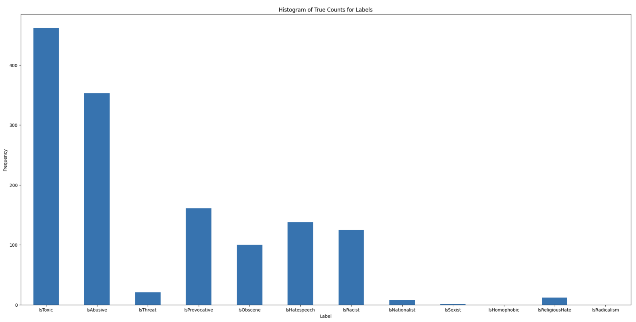
A Stanford CS paper has outlined advanced machine learning techniques to tackle the problem of toxic comments. The purpose was to create a model that could detect toxic comments online. They concluded that they could attain good accuracy on public data sets.

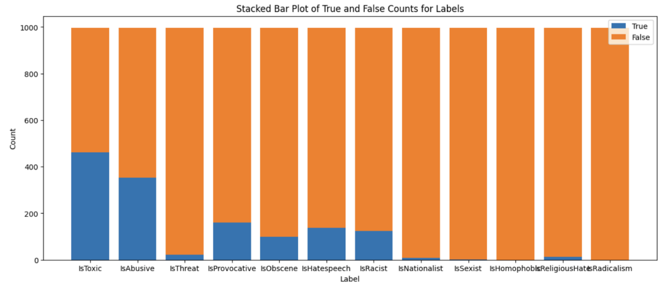
This research paper concluded that LogisticRegression and AdaBoost were best classifiers for toxicity classification. The investigation highlights the importance of choosing the right combination of ML classifier and feature set for effective toxic comment analysis.

Dataset

1. YouTube Toxic Comments Dataset

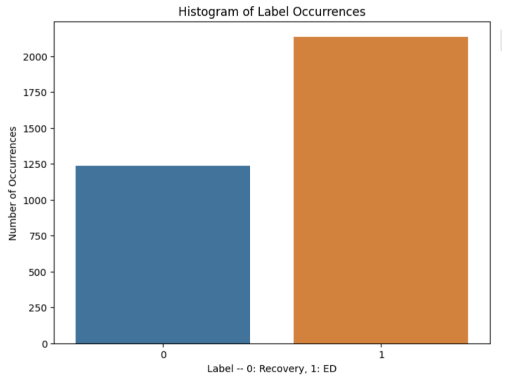
- Contains 1,000 comments labeled across multiple categories: isToxic, isAbuse, isThreat, isObscene, isHatespeech, isRacist, etc.
- Labels are highly imbalanced; most true values fall under isToxic.
- Data split: 80% training, 20% testing.





2. Pro-Ana vs Pro-Recovery Dataset

- NLP dataset classifying text as promoting recovery (0) or an eating disorder (1).
- Contains 3,369 inputs, with more pro-eating disorder samples.
- Data split: 80% training, 20% testing.

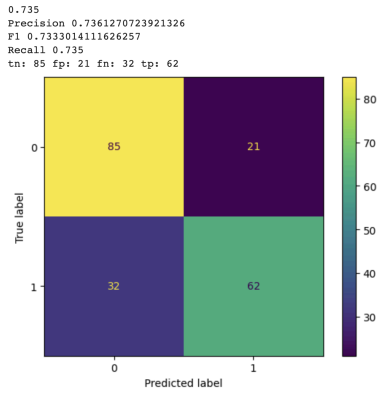


Methodology/Models

I evaluated five ML models: Logistic Regression, RidgeClassifier, RandomForestClassifier, DecisionTreeClassifier, and SVC.

YouTube Dataset:

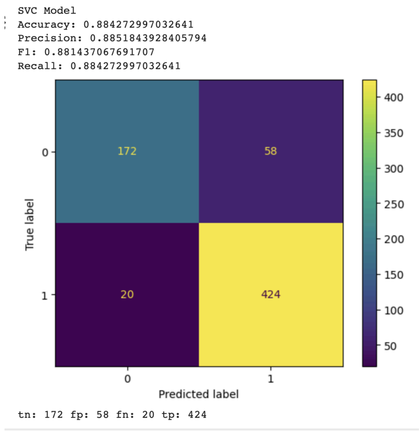
- Logistic Regression achieved the highest accuracy (73%).
- RandomForest and SVC were similar (≈73%).
- Parameters such as max_iter and max_depth were tuned to improve performance.



```
0.735
Precision 0.7361270723921326
F1 0.7333014111626257
Recall 0.735
tn: 85 fp: 21 fn: 32 tp: 62
```

ED Dataset

- SVC achieved the highest accuracy (88%).
- Confusion matrices were used to assess true positives, false positives, true negatives, and false negatives.

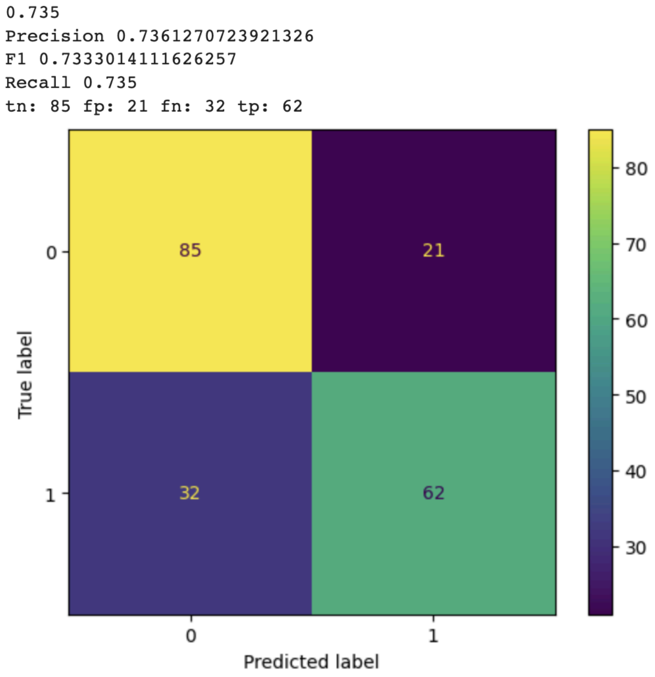


```
SVC Model
Accuracy: 0.884272997032641
Precision: 0.8851843928405794
F1: 0.881437067691707
Recall: 0.884272997032641
```

```
tn: 172 fp: 58 fn: 20 tp: 424
```

Results and Discussion

- ML models consistently achieved high 70–80% accuracy, demonstrating their potential for toxic content classification.
- The main challenge lies in fine-grained labeling. Most public datasets rely on general categories (e.g., “isToxic”), which limits personalization.
- Future work will focus on creating balanced, fine-grained datasets that distinguish toxicity across specific groups and content types.



```
0.735
Precision 0.7361270723921326
F1 0.7333014111626257
Recall 0.735
tn: 85 fp: 21 fn: 32 tp: 62
```

Conclusion

Machine learning can effectively classify toxic content with moderate to high accuracy. To deploy such models in real-world social media platforms, the largest effort will be collecting and maintaining fine-grained labeled datasets for better personalization and content filtering.

Acknowledgements

Thanks to Varsha Sandadi for guidance, and to my parents for idea brainstorming and support

References

1. Toxic comment detection and classification
   https://cs229.stanford.edu/proj2019spr/report/71.pdf
2. How to Machine Learning Algorithms classify toxicity in comments? An empirical study
   https://cs229.stanford.edu/proj2019spr/report/71.pdf
3. Youtube toxic comments data set from Kaggle
   https://www.kaggle.com/datasets/reihanenamdari/youtube-toxicity-data