

1. **Automated Hate Speech Detection and the Problem of Offensive Language\***

*Thomas Davidson,1 Dana Warmusley,2 Michael Macy,1,3 Ingmar Weber4*

Research Paper: <https://arxiv.org/pdf/1703.04009.pdf>

**Models Implemented:** logistic regression, naïve Bayes, decision trees, random forests, and linear SVMs. One-versus-rest framework where a separate classifier is trained for each class and the class label with the highest predicted probability across all classifiers is assigned to each tweet. All modeling was performed using scikit-learn

**Pros and Cons:** Logistic Regression and Linear SVM tended to perform significantly better than other models. Logistic regression with L2 regularization for the final model as it more readily allows us to examine the predicted probabilities of class membership

2. **A Literature Review of Textual Hate Speech Detection Methods and Datasets**

*Fatimah Alkomah 1,2,\* and Xiaogang Ma 1*

Literature Review: <https://www.mdpi.com/2078-2489/13/6/273/pdf>

**Models Implemented:** TFIDF Methods, Lexicon-Based Methods, CNNs use filters and a set of pooling layers for hate speech detection tasks, LSTMs, GRUs, BiLSTM models, generative pretrained transformer models, BERT, Hybrid Models.

**Pros and Cons:** Best method was one-vs -one SVM, followed by LR, Followed by LSTM. LSTMs and other methods do not work well with data that has a lot of pre required context. Needs a lot of spoonfeeding and data background data that might not be available always

**MODEL TO BE USED: ONE-VS-ONE LINEAR SVM**