# Twitter Hate Speech Identification

**Vadym Gryshchuk**

## 1. Introduction

Approaches that are based on learned contextual embeddings of words (Pennington et al., 2014; Liu et al., 2019) have been successfully applied in different natural language processing (NLP) tasks (Basile et al., 2019). Thus, two models are considered: a CNN-based model using GloVe embeddings (Pennington et al., 2014) and a RoBERTa-based model (Barbieri et al., 2020).

**CNN model** The CNN model uses three convolutional layers followed by batch normalization, $ReLU$ activation function and $max$ pooling. Filter outputs are concatenated after the last pooling layer. Dropout is applied to the concatenated filter outputs. A concatenated output is passed then through a linear layer to make a prediction. An input to this model is a tweet consisting of pre-trained embeddings from the GloVe model (Pennington et al., 2014). An output of the model is a real number. Thus, a sigmoid function is applied to the output to produce values between 0 and 1. A threshold for binary classification is set to 0.5. A *binary cross-entropy with logits* function is used to train a model[1].

**RoBERTa-based model** The RoBERTa-based model utilizes a pre-trained RoBERTa model (Liu et al., 2019) on 58M tweets (Barbieri et al., 2020). A RoBERTa model is used as a feature extractor and thus is not trained. The recurrent neural network component processes the extracted embeddings and provides the last hidden state representation to a linear layer. The same loss function is used as in the CNN model.

**Dataset** Twitter Hate Speech dataset[2] contains tweets categorized into hateful and non-hateful. The dataset is unbalanced. Non-hateful tweets are more prevalent.

## 2. Results and Evaluation

The dataset (only train set is downloaded) is divided into train, validation and test sets. The test set is always the same, but the distribution of tweets between train and validation sets vary. Stratified sampling is used to ensure that enough samples from each tweet category are sampled, thus preserving the ratio of hateful and non-hateful tweets as in the original data.

*Table 1.* Recall, precision, $F_1$ and $F_2$ scores with standard errors of two trained models evaluated on a test set with three random initializations of classifiers.

| METRIC | CNN MODEL | ROBERTA-BASED MODEL |
|---|---|---|
| RECALL | 0.37± 0.04 | 0.93± 0.02 |
| PRECISION | 0.40± 0.01 | 0.58± 0.05 |
| $F_1$ SCORE | 0.38± 0.02 | 0.71± 0.04 |
| $F_2$ SCORE | 0.38± 0.03 | 0.83± 0.01 |

**Performance Measure** Recall, precision and $F_\beta$ score are used to evaluate the models trained on unbalanced data. Recall measures the fraction of the actual hateful tweets (True Positives) that were retrieved (True Positives + False Negatives). Precision tells how many actual tweets (True Positives) were correctly classified as hateful among all the tweets that a model classified as hateful (True Positives + False Positives). Thus, *Recall* is defined as $TP/(TP+FN)$ and *Precision* equals to $TP/(TP+FP)$, where $TP$ are True Positives, $FP$ are False Positives, and $FN$ are False Negatives. $F_\beta$ score is the weighted harmonic mean of precision and recall and is defined as

$$F_\beta = (1+\beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}, \qquad (1)$$

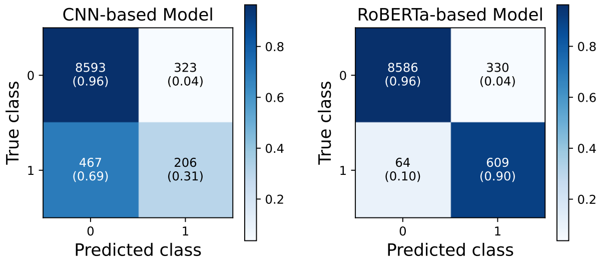
where $\beta$ is a positive real factor. Larger values for $\beta$ favour recall more than precision, which can be more meaningful for scenarios when the identification of hateful tweets is very important. The optimal value of $F_\beta$ is at 1. Table 1 shows the performance of two models on a test set.

**Error Analysis** Fig. 1 shows a confusion matrix for the CNN and RoBERTa-based models. The identification of hateful tweets is a challenging task. The RoBERTa-based model identifies more hateful tweets correctly. Fig. 2 shows the density function of predictions for actual labels.

**Generalization** Generalization is one of the main challenges of machine learning algorithms that is often expressed as a bias-variance tradeoff. A model that exhibits high bias does not learn representations of train data properly, which is known as underfitting. On the contrary, a

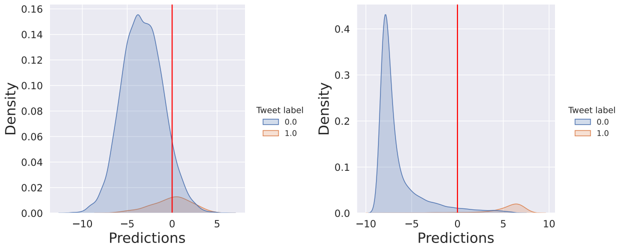
[1] Binary cross entropy with logits function in PyTorch: https://bit.ly/3oZjsON

[2] Twitter hate speech dataset: https://bit.ly/3HUgNPo



(a) (b)

*Figure 1.* Confusion matrix on test data: (a) CNN model (b) RoBERTa-based model.



(a) (b)

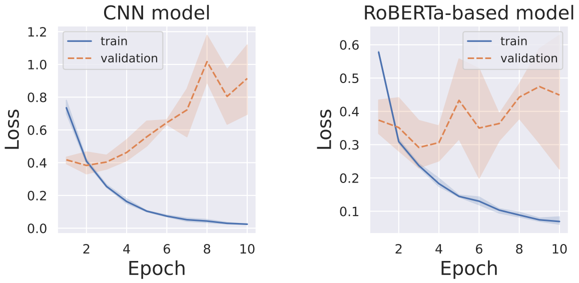
*Figure 2.* Density estimation of predictions: (a) CNN model (b) RoBERTa-based model (red line denotes a threshold).

model that fits perfectly to train data may generalize poorly and thus show high variance. Fig. 3 visualizes train and validation binary cross-entropy loss over epoch. Both models start to overfit already after a few first epochs. However, the RoBERTa-based model shows not so strong overfitting.

**McNemar's Test** McNemar's test (McNemar, 1947) is used to compare two methods on the same data. We have seen that the RoBERTa-based model provides better identification of hateful tweets (higher sensitivity) than a CNN model. However, specificity (True Negative Rate) is similar for both models. McNemar's test compares the sensitivity and specificity of two models. After creating a contingency table and performing a test the calculated statistic is 273.78 and the p-value is 0.00. Thus, the null hypothesis of marginal homogeneity at the significance level of 0.05 is rejected. Different proportions of errors are observed for two classifiers.

## 3. Conclusion

We have shown that the use of pre-trained embeddings or pre-trained architectures provides a solid basis for achieving high results on the identification of hatefull tweets. Furthermore, McNEmar's test showed that the evaluated models have a different proportion of errors on the the test set. The RoBERTa-based model achieves better results than the CNN model. The increase in performance is due to the pre-trained embeddings provided by a model trained on tweets. Thus, the RoBERTa-based model should generalizes better also to the data of similar domain. Classifiers that can learn or utilize powerful representations are the candidates for better generalization.



(a) (b)

*Figure 3.* Train and validation loss values over three lerning trials: (a) CNN model (b) RoBERTa-based model (shaded areas indicate 95% confidence interval using standard error).

## References

Barbieri, F., Camacho-Collados, J., Espinosa Anke, L., and Neves, L. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1644–1650, November 2020.

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

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019. URL `http://arxiv.org/abs/1907.11692`.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Pennington, J., Socher, R., and Manning, C. D. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543, 2014.