Term Project Report for HS Empirical Methods for NLP and Data Science: Inferential Reproducibility Examination on a Retrained RoBERTa model for Hate Speech Detection

Jinghua Xu
Heidelberg University, Germany
jinghua.xu@stud.uni-heidelberg.de

Abstract

This report summarises the major analysis and key findings of an Inferential Reproducibility examination on a State-of-the-Art (SOTA) model proposed in a previous research for the task of hate speech detection. The analysis consists of two major steps: Significance Testing using Linear Mixed Effects Models (LMEMs) and Reliability Analysis through Variance Component Analysis (VCA). The reported statistics indicate statistically significant improvement of the SOTA model performance from a strong Baseline. Meanwhile, the Reliability Analysis reveals poor reliability of the SOTA model and the contributions of a set of model meta-parameters to the total variance.

1 Introduction

A research procedure is reliable when it responds to the same phenomena in the same way regardless of the circumstances of its implementation (Krippendorff, 2018). Various other terms such as "agreement", This analysis intends to examine the inferential reproducibility (Riezler and Hagmann, 2021) of the State-of-the-Art (SOTA) model proposed in Barbieri et al. (2020) for hate speech detection, a RoBERTa (Liu et al., 2019) model retrained on over 60M tweets and fine-tuned on the hate speech dataset proposed in Basile et al. (2019). The analysis consists of two major steps. It first conducts Significance Testing using Linear Mixed Effect Models (Pinheiro and Bates, 2000; Riezler and Hagmann, 2021) to examine the statistical significance of the SOTA model performance improvement from a strong Baseline. This is followed by Reliability Analysis through Variance Component Analysis (Koo and Li, 2016; Riezler and Hagmann, 2021) to untangle sources of variability in measurement and assess general robustness of model by ratio of substantial variance out of total variance. Additionally, further analysis examines the interaction with various data properties including data length, average word frequency, and readability. Based on the test statistics and visualizations, the analysis finds the performance improvement of the SOTA model from Baseline statistically significant, and the reliability of the SOTA model to be poor. The following sections of this report first introduces the task, dataset, and models involved with the analysis, and then discusses the analysis, findings and final remarks. This report summarizes the major analysis and the key findings by the author, while extended analyses with more details (further statistical tests, visualizations, etc.) can be found in the original Jupyter Notebooks released at github.com/JINHXu/Inferencial-Reproducibility-RoB-RT-hate.

2 Task & Dataset

The task is binary text classification to detect hate speech on Twitter (Basile et al., 2019).[1] The dataset (Basile et al., 2019) consists of 13000 English tweets labelled as hateful/non-hateful against immigrants and women, with 10000 for training and 2970 for testing.[2]

3 Models

3.1 Baseline Model: SVM

The Baseline model is based on SVM (Cortes and Vapnik, 1995) with both word and character n-gram features. It serves as a strong baseline since great success using this approach had been seen in previous work on tweet classification such as Çöltekin and Rama (2018) and Mohammad et al. (2018). In Barbieri et al. (2020) the reported F-1 score of the SVM Baseline model is 35.7.

[1]The original shared-task consists of two subtasks including binary classification and multi-class classification. This analysis is only concerned with the binary classification subtask.

[2]The original dataset consists of both English and Spanish tweets, while the SOTA model studied in this analysis is only concerned with English hate speech detection, the Spanish data is thus disregarded in this analysis.

3.2 SOTA Model: Retrained RoBERTa

The SOTA model is a RoBERTa model retrained on 60M English tweets (584 million tokens), and fine-tuned on the training data created in Basile et al. (2019) for hate speech detection. In Barbieri et al. (2020), the reported F-1 score of the SOTA model is 55.5.

4 Analysis

4.1 Analysis Setup

4.1.1 Data Measurements

This analysis considers the following data characteristics to measure data difficulty:

- Length: The number of tokens in each tweet
- Readability: Flesch-Kincaid readability (Kincaid et al., 1975) of each tweet
- Frequency:[3] The average word frequency of tokens in each tweet. (Zhang et al., 2018). Frequency is multiplied by 10000.

The analysis groups data into levels according to the above measurements. Data samples shorter than 15 tokens are classified as "short", with 15-55 as "typical", and longer than 55 as "very long". According to Kincaid et al. (1975), data samples with a readability score lower than 50 are considered "difficult", within 50-80 "fair", while above 80 "easy". Lastly, to define frequency into three levels, this analysis considers three different setups for classification: with threshold set at [30, 50], [40, 80], and [50, 100] to groups data into "low-frequency", "regular-frequency" and "high-frequency".[4]

4.1.2 Meta-parameters

For the Reliability Analysis of the SOTA model, the following meta-parameters have been considered in this analysis:

- Training Batch Size: 8, 32, 64
- Random Seed: 1, 2, 3
- Learning Rate: 1e-3, 1e-4, 1e-5

Number of training epochs is set to 5 to be consistent with the setup in Barbieri et al. (2020).

4.2 Data Overview

	length	readability	frequency
Min.	1.00	-1191.10	0.00
1st Qu.	14.00	50.53	26.72
Median	20.00	75.71	40.06
Mean	21.41	66.45	40.49
3rd Qu.	27.00	91.11	54.24
Max.	90.00	119.19	124.65

Table 1: Data overview.

It can be seen from Table 1, the data samples are overall short (mean length at 21.41), and fairly easy to read (mean readability at 66.45).

4.3 Significance Testing[5]

4.3.1 System Comparison: SOTA vs. Baseline

In order to test weather the performance improvement of the SOTA model from Baseline is statistically significant, the analysis first applies generalized likelihood ratio test (GLRT) to a data model and a restricted model.

chi_square	df	p
190.3342578991069	1	0.0

Table 2: System Comparison by applying GLRT to LMEMs: Baseline vs. SoTA

It can be seen from Table 2, that the p-value (0.0) is lower than the typical alpha level of 0.05, the null hypothesis can thus be rejected. Therefore, the SOTA model does not perform equally well with the Baseline model.

	Coef.	Std.Err.	z	P>|z|
system[Baseline]	0.472	0.009	51.763	0.000
system[SoTA]	0.577	0.009	63.290	0.000

Table 3: Model summary.

Furthermore, it can be seen from model summary in Table 3 that the estimated mean performance for Baseline is 0.472, and 0.577 for SOTA.

[3] While better ways to represent word rarity to measure data difficulty had been proposed in Platanios et al. (2019). It proposes to use the negative multiplication of log frequency of words in each sentence to reflect the overall word rarity of a sentence. Due to the nature of this dataset being composed of tweets, which contain various misspelled words, Internet memes and combined hashtags such as "#netflix#tv#hulu" with frequency zero as illegal input to log, the approach cannot be used.

[4] This report only presents analysis with the [30, 50] setup for frequency level definition, more detailed analysis including the other two setups can be found in the original Jupyter Notebooks.

[5] The full analysis (Colab notebook) is released here.

SOTA model therefore significantly outperforms the Baseline.

4.3.2 System Comparison: Condition on Data Properties

In addition to examining the overall model performance improvement, the following analysis conducts system comparison conditioning on the three data properties.

Length

In order to examine the interaction between data length and the performance gap between the two models, the analysis first applies GLRT to an LMEM with the interaction term and one without the interaction term.

chi_square	df	p
2.3953713894152315	2	0.3018920744929181

Table 4: Interaction with data length.

As shown in Table 4, p-value is higher than the typical alpha level of 0.05, the interaction term is thus statistically insignificant.

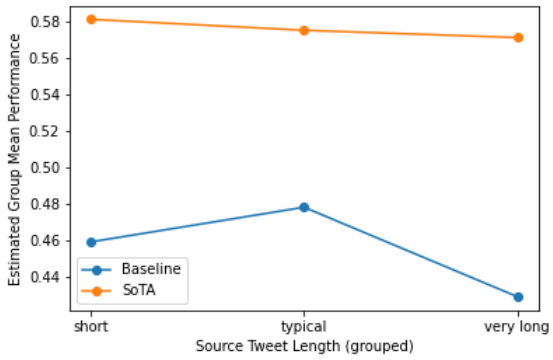

Figure 1: Estimated mean performance in each length group: Baseline vs. SOTA.

As can be seen from Figure 1, SOTA model outperforms Baseline in each length group, with the performance gap the smallest for data in the "typical" length group.

	Coef.	Std.Err.	z	P>|z|
c0	-0.1216	0.013	-9.201	0.000
c1	-0.0971	0.009	-10.659	0.000
c2	-0.1429	0.154	-0.926	0.354

Table 5: Pairwise comparison: performance gap in each data group (length).

In order to see if the performance gap between the two systems is statistically significant, it is important to conduct conditional post-hoc analysis. As shown in Table 5, p-value is lower than 0.05 for short and typical data, the performance gap is thus statistically significant in these two data groups. Whereas p-value for the very long data class is higher than 0.05, the performance gap is therefore not significant in this data group.

Readability

As presented in Table 6, p-value is smaller than the typical alpha level of 0.05, the null hypothesis is thus rejected. The interaction term is statistically significant.

chi_square	df	p
50.741522282434744	2	9.585554572311139e-12

Table 6: Interaction with readability.

As can be seen from Figure 2, the Baseline system works the best with “easy" tweets, while worse with tweets fairly readable. The SoTA system works better with tweets with better readability. The SOTA system constantly outperforms the Baseline system, with the performance gap the smallest when it comes to “difficult" tweets, the largest for “easy" tweets.

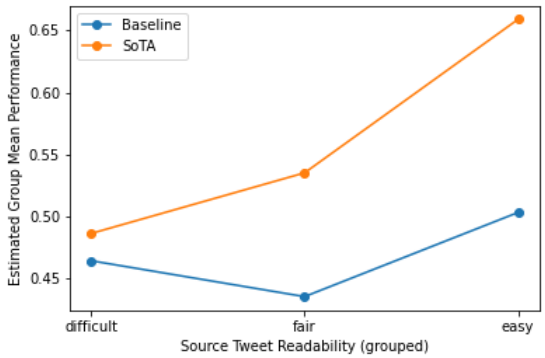

Figure 2: Estimated mean performance in each readability group: Baseline vs. SOTA.

As shown in Table 7, the performance improvement for SOTA over Baseline is statistically significant (p-value < 0.05) for tweets with “fair" and “easy" readability, yet insignificant (p-value > 0.05) for “difficult" tweets.

Frequency

As shown in Table 8, p-value is higher than the typical alpha level of 0.05, the interaction term with frequency is thus not statistically significant.

It can be seen from Figure 3, SOTA still outperforms Baseline in each data group, both systems

	Coef.	Std.Err.	z	P>|z|
c0	-0.0220	0.015	-1.465	0.143
c1	-0.0997	0.013	-7.601	0.000
c2	-0.1559	0.011	-13.828	0.000

Table 7: Pairwise comparison: performance gap in each data group (readability).

chi_square	df	p
3.824801916387514	2	0.14772527824866688

Table 8: Interaction: frequency

work the best with tweets in the regular-frequency group, yet the contrast between both systems does not vary notably across frequency groups. In the meantime, it is interesting to see that both models work the worst with tweets in the high-frequency group, which is anti-intuitive.

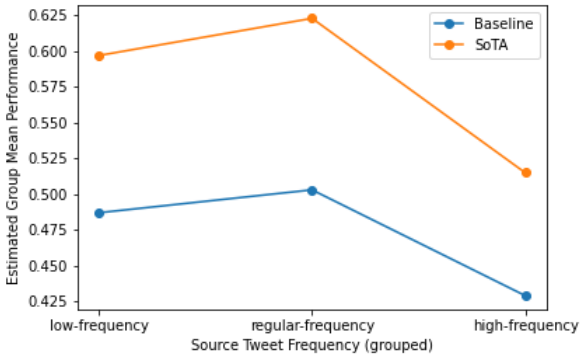

Figure 3: Estimated mean performance in each frequency group: Baseline vs. SOTA.

Table 9 reports the statistical test results of the conditional post-hoc analysis. The contrast in terms of system performance between Baseline and SOTA is statistically significant for tweets in each "frequency" group, according to the p-values (= 0.000, < 0.05).

4.4 Reliability Analysis[6]

4.4.1 Variance Component Analysis

As mentioned in a previous section, the test dataset consists of 2970 tweets, which are the objects of interest in this analysis. The meta-parameters examined in the analysis make up a grid of 27 models (3 random seeds × 3 learning rates × 3 batch sizes). Table 10 shows that the substantial variance for these experimental data amounts to only 8.2% of the total variance, the reliability coefficient is

[6]The full analysis (Colab Notebook) is released here.

	Coef.	Std.Err.	z	P>|z|
c0	-0.1105	0.014	-7.854	0.000
c1	-0.1198	0.013	-9.559	0.000
c2	-0.0861	0.012	-6.899	0.000

Table 9: Pairwise comparison: performance gap in each data group (readability).

poor (lower than 50%) according to the guidelines of Koo and Li (2016).

Variance component v	Variance σ_v^2	Percent
Residual	0.2262137	91.8
tweet_id	0.0202055	8.2

Table 10: Variance components for substantial variance check.

An interpretation of each random effect as a variance component is shown in Table 11. It can be seen that all meta-parameters introduce negligible variance of < 1%. Amongst the meta-parameters, learning rate (lr) induces the largest variance. It contributes 0.11% to the total variance. While contribution of variance due to replications under random seeds is only 0.04%, and there is essentially no contribution of variance due to different training batch sizes. The variance contributed by objects of interest is only 8.2% of total variance. According to the 80% threshold theory of Jiang (2018), this coefficient value tells that predictions of the SOTA model are inconsistent across the meta-parameter configurations of the grid search. According to the guidelines of Koo and Li (2016), this value of can be interpreted as poor reliability.

Variance component v	Variance σ_v^2	Percent
Residual	2.259413e-01	91.64
tweet_id	2.021556e-02	8.20
lr	2.753851e-04	0.11
seed	1.083961e-04	0.04
batch	9.759875e-06	0.00

Table 11: Variance components in meta-parameter grid search for SOTA model fine-tuning.

4.4.2 Interaction of Meta-parameters with Data Properties

In addition to using LMEMs to assess meta-parameter "importance" through variance component analysis, this analysis is also interested in examining the interaction of meta-parameters and

test data characteristics. The following interaction examination mainly focuses on learning rate amongst all the meta-parameters, and data properties including: length, average word frequency, and readability.

Interaction: Learning rate & Length

As can be seen from Figure 4, the choice of learning rate causes a performance difference of the SOTA model for tweets of typical length. For very long and short tweets, a higher learning rate (1e-3) results in different performance, whereas the system have the same performance with learning rates of 1e-4 and 1e-5.

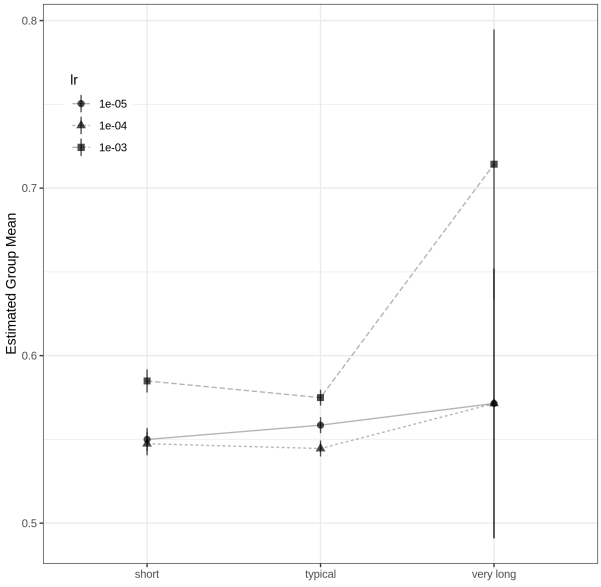

Figure 4: Estimated group mean performance for the SOTA model trained with different learning rates (per each length data group).

As can be seen in Table 12, p-value is higher than the typical alpha level of 0.05, the interaction between learning rate and length is statistically insignificant.

	F value	Pr(>F)
src_length_class	0.452434	0.63612190
lr	3.787069	0.02266614
src_length_class:lr	1.743087	0.13736869

Table 12: Interaction: Learning rate & Length

Interaction: Learning rate & Readability

As shown in Figure 5, the choice of learning rate causes a performance difference of the SOTA model for data in all readability groups. As shown in Table 13, p-value is lower than the typical alpha level of 0.05, the interaction between learning rate and readability is statistically significant.

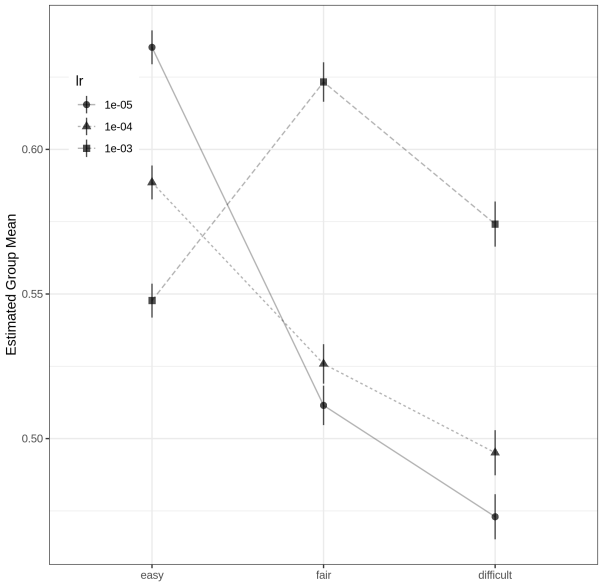

Figure 5: Estimated group mean performance for the SOTA model trained with different learning rates (per each readability data group).

	F value	Pr(>F)
src_readability_class	50.149	3.8051e-22
lr	71.572	8.8116e-32
src_readability_class:lr	144.816	1.3927e-123

Table 13: Interaction: Learning rate & Readability

Interaction: Learning rate & Frequency

It can be seen from Figure 6, the choice of learning rate causes a performance difference of the SOTA model for data in all frequency groups. As shown in Table 14, p-value is lower than the typical alpha level of 0.05, the interaction between learning rate and frequency is statistically significant.

	F value	Pr(>F)
src_frequency_class	15.686	1.6724e-07
lr	42.565	3.3444e-19
src_frequency_class:lr	48.668	5.9123e-41

Table 14: Interaction: Learning rate & Frequency

Further Analysis of the Interaction

The above visualizations provide a rough idea on the performance gap between models trained with different learning rates in each data group. However, in order to inspect the statistical significance of the performance gap between models trained

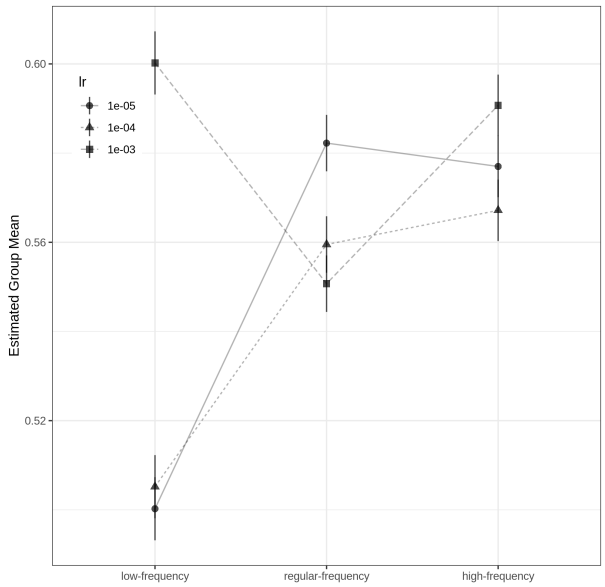

Figure 6: Estimated group mean performance for the SOTA model trained with different learning rates (per each frequency data group).

with each pair of learning rates in each data group, it is not sufficient to observe from the visualization, instead, it is important to run pairwise comparisons to obtain the p-values to interpret the significance. These further analysis are not included in this report, but can be found in the original Jupyter Notebook for Reliability Analysis.

5 Conclusion

This analysis conducted Significance Testing using LMEMs and Reliability Analysis through Variance Component Analysis using a Twitter dataset composed of data samples averagely short and fairly easy to read. The analysis found the performance improvement of the SOTA model from the Baseline statistically significant overall. The performance improvement is, however, statistically insignificant for data difficult to read and data that are very long. The conditional analysis found the interaction between system performance improvement and data readability statistically significant, while the interaction is insignificant with the other two data properties. Through VCA, it was found out that the SOTA model has poor reliability, and that the meta-parameters introduce negligible variance (< 1%), with learning rate contributes the largest variance (0.11%) amongst all meta-parameters. Future work can consider extending the meta-parameter search space, i.e. a larger search grid for reliability analysis.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. 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*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Çağrı Çöltekin and Taraka Rama. 2018. Tübingen-Oslo at SemEval-2018 task 2: SVMs perform better than RNNs in emoji prediction. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 34–38, New Orleans, Louisiana. Association for Computational Linguistics.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Zhehan Jiang. 2018. Using the linear mixed-effect model framework to estimate generalizability variance components in r: A lme4 package application. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, 14(3):133.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Terry K Koo and Mae Y Li. 2016. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

José C Pinheiro and Douglas M Bates. 2000. Linear mixed-effects models: basic concepts and examples. *Mixed-effects models in S and S-Plus*, pages 3–56.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom M Mitchell. 2019. Competence-based curriculum learning for neural machine translation. *arXiv preprint arXiv:1903.09848*.

Stefan Riezler and Michael Hagmann. 2021. Validity, reliability, and significance: Empirical methods for nlp and data science. *Synthesis Lectures on Human Language Technologies*, 14(6):1–165.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *arXiv preprint arXiv:1811.00739*.