

BaselineAvengers at PAN 2024

Often-Forgotten Baselines for LLM-Generated Text Detection

Ludwig Lorenz , Funda Zeynep Aygüler , Ferdinand Schlatt and Nailia Mirza-khmedova

December 17, 2024

CLEF 2024 Conference and Labs of the Evaluation Forum, 9-12 September 2024, Grenoble - France

TABLE OF CONTENTS

1. Setting

2. Experiments

3. Results

4. Qualitative Analysis

Setting

- **Data:** 1359 pairs of real and LLM generated US news articles
- **Approach:** employ non-neural models
- Simple models are often overlooked in recent research, despite their proven effectiveness and their ability to serve as efficient baselines for comparison with more complex models [1]

- **Data:** 1359 pairs of real and LLM generated US news articles
- **Approach:** employ non-neural models
- Simple models are often overlooked in recent research, despite their proven effectiveness and their ability to serve as efficient baselines for comparison with more complex models [1]

- **Data:** 1359 pairs of real and LLM generated US news articles
- **Approach:** employ non-neural models
- Simple models are often overlooked in recent research, despite their proven effectiveness and their ability to serve as efficient baselines for comparison with more complex models [1]

Experiments

Table 1: Overview of the accuracy of the different classifiers (rows) and features (columns) evaluated on the validation set.

Classifier	tf-idf	Term Frequencies
Multinomial Naive Bayes	0.770	0.874
Logistic Regression	0.927	0.922
SVM	0.932	0.925

Results

Table 2: Overview of the performance of our approaches, baselines, and the summary statistics of the performance of all submissions in the competition over 10 variants of the test set. We report the minimum, median and maximum of the arithmetic mean of all metrics.

Approach	Minimum	Median	Max
naive-bayes	0.884	0.945	0.969
logistic-regression	0.837	0.957	0.989
svm	0.832	0.969	0.999
Baseline Binoculars	0.342	0.844	0.996
Baseline Fast-DetectGPT (Mistral)	0.095	0.842	0.958
Baseline PPMd	0.270	0.750	0.863
Baseline Unmasking	0.250	0.673	0.762
Baseline Fast-DetectGPT	0.159	0.677	0.982

Table 3: Overview of the performance of our approaches, baselines, and the summary statistics of the performance of all submissions in the competition. We report ROC-AUC, Brier, C@1, F_1 , $F_{0.5u}$ and their arithmetic mean.

Approach	ROC-AUC	Brier	C@1	F_1	$F_{0.5u}$	Mean
naive-bayes	0.998	0.859	0.975	0.975	0.974	0.956
logistic-regression	0.996	0.884	0.97	0.97	0.97	0.958
svm	0.994	0.923	0.976	0.976	0.975	0.969
Baseline Binoculars	0.972	0.957	0.966	0.964	0.965	0.965
Baseline Fast-DetectGPT (Mistral)	0.876	0.8	0.886	0.883	0.883	0.866
Baseline PPMd	0.795	0.798	0.754	0.753	0.749	0.77
Baseline Unmasking	0.697	0.774	0.691	0.658	0.666	0.697
Baseline Fast-DetectGPT	0.668	0.776	0.695	0.69	0.691	0.704

Qualitative Analysis

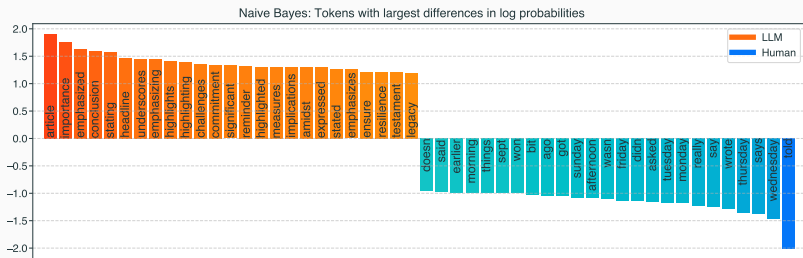


Figure 1: Top 50 tokens with the largest differences in log probabilities for multinomial Naive Bayes. Positive values indicate the probability is higher for LLM-generated texts, negative values indicate the probability is higher for human-written texts.

- LLM-generated texts frequently use specific terms such as “article”, “importance”, “emphasized”, “context”, and “despite”
- human-written texts show a higher probability of tokens related to everyday language and temporal expressions such as “told”, “says”, “asked”, “wrote”, and “really”
- The word “told” is particularly prominent in human-written texts, as it is frequently used in direct and indirect speech, which is also common in news articles

- LLM-generated texts frequently use specific terms such as “article”, “importance”, “emphasized”, “context”, and “despite”
- human-written texts show a higher probability of tokens related to everyday language and temporal expressions such as “told”, “says”, “asked”, “wrote”, and “really”
- The word “told” is particularly prominent in human-written texts, as it is frequently used in direct and indirect speech, which is also common in news articles

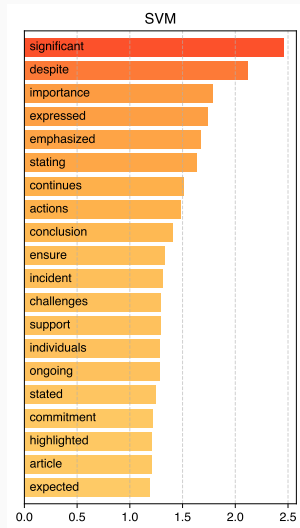
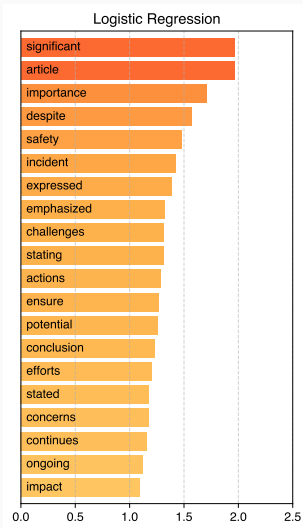


Figure 2: Top 20 tokens for identifying LLM-generated texts using Logistic Regression (left) and SVM (right). The importance of each token is based on the size of the coefficients assigned to them by the trained models.

- “significant”, “article”, “importance”, “despite”, “stating” and “conclusion” are the most notable tokens in both models, logistic regression and SVM
- This suggests that LLM-generated texts often contain terms that convey formality, which might be less prevalent in human-written texts
- frequent appearance of the word “significant” in LLM-generated texts can be attributed to the tendency of language models to produce polished and systematic content

Hypothesis

LLMs use a distinctive vocabulary that can be captured using lexical features.

- “significant”, “article”, “importance”, “despite”, “stating” and “conclusion” are the most notable tokens in both models, logistic regression and SVM
- This suggests that LLM-generated texts often contain terms that convey formality, which might be less prevalent in human-written texts
- frequent appearance of the word “significant” in LLM-generated texts can be attributed to the tendency of language models to produce polished and systematic content

Hypothesis

LLMs use a distinctive vocabulary that can be captured using lexical features.

- “significant”, “article”, “importance”, “despite”, “stating” and “conclusion” are the most notable tokens in both models, logistic regression and SVM
- This suggests that LLM-generated texts often contain terms that convey formality, which might be less prevalent in human-written texts
- frequent appearance of the word “significant” in LLM-generated texts can be attributed to the tendency of language models to produce polished and systematic content

Hypothesis

LLMs use a distinctive vocabulary that can be captured using lexical features.

CONCLUSION

LLM-generated

formal and structured
discourse

Human-written

more narrative and less
formal vocabulary

- assumption that LLMs use a particular vocabulary, which can be captured using lexical features
- logistic regression and SVM models using tf-idf feature vectors are highly effective
- submissions outperform all official baselines, demonstrating that simple and interpretable models can be more effective than complex and computationally expensive methods

CONCLUSION

LLM-generated

formal and structured
discourse

Human-written

more narrative and less
formal vocabulary

- assumption that LLMs use a particular vocabulary, which can be captured using lexical features
- logistic regression and SVM models using tf-idf feature vectors are highly effective
- submissions outperform all official baselines, demonstrating that simple and interpretable models can be more effective than complex and computationally expensive methods

Questions?

BACKUP SLIDE: EVALUATION SCORES

- **ROC-AUC:** The area under the ROC (Receiver Operating Characteristic) curve
- **Brier:** The complement of the Brier score (mean squared loss)
- **C@1:** A modified accuracy score that assigns non-answers (score = 0.5) the average accuracy of the remaining cases
- **F1:** The harmonic mean of precision and recall
- **F0.5u:** A modified F0.5 measure (precision-weighted F measure) that treats non-answers (score = 0.5) as false negatives
- The arithmetic mean of all the metrics above.

References

- [1] Y.-C. Lin, S.-A. Chen, J.-J. Liu, and C.-J. Lin. Linear classifier: An often-forgotten baseline for text classification. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1876–1888, Toronto, Canada, July 2023. Association for Computational Linguistics.