Are Neural Language Models Good Plagiarists?
A Benchmark for Neural Paraphrase Detection

Jan Philip Wahle
University of Wuppertal
Wuppertal, Germany
wahle@uni-wuppertal.de

Terry Ruas
University of Wuppertal
Wuppertal, Germany
ruas@uni-wuppertal.de

Norman Meuschke
University of Wuppertal
Wuppertal, Germany
meuschke@uni-wuppertal.de

Bela Gipp
University of Wuppertal
Wuppertal, Germany
gipp@uni-wuppertal.de

Abstract—**Neural language models such as BERT allow for human-like text paraphrasing. This ability threatens academic integrity, as it aggravates identifying machine-obfuscated plagiarism. We make two contributions to foster the research on detecting these novel machine-paraphrases. First, we provide the first large-scale dataset of documents paraphrased using the Transformer-based models BERT, RoBERTa, and Longformer. The dataset includes paragraphs from scientific papers on arXiv, theses, and Wikipedia articles and their paraphrased counterparts (1.5M paragraphs in total). We show the paraphrased text maintains the semantics of the original source. Second, we benchmark how well neural classification models can distinguish the original and paraphrased text. The dataset and source code of our study are publicly available.**

Index Terms—**Paraphrase detection, BERT, transformers**

I. INTRODUCTION

Transformer-based language models [1] have reshaped natural language processing (NLP) and become the standard paradigm for most NLP downstream tasks [2], [3]. Now, these models are rapidly advancing to other domains such as computer vision [4]. We anticipate Transformer-based models will similarly influence plagiarism detection research in the near future [5]. Plagiarism is the use of ideas, concepts, words, or structures without proper source acknowledgment. Often plagiarists employ paraphrasing to conceal such practices [6]. Paraphrasing tools, such as *SpinBot*[1] and *SpinnerChief*[2], facilitate the obfuscation of plagiarised content and threaten the effectiveness of plagiarism detection systems (PDS).

We expect that paraphrasing tools will abandon deterministic machine-paraphrasing approaches in favor of neural language models, which can incorporate intrinsic features from human language effectively [3]. The ability of models such as GPT-3 [3] to produce human-like texts raises major concerns in the plagiarism detection community as statistical and traditional machine learning solutions cannot distinguish semantically similar texts reliably [7]. Using Transformer-based models for the classification seems to be intuitive to counteract this new form of plagiarism. However, Transformer-based solutions typically require sufficiently large sets of labeled training data to achieve high classification effectiveness. As the use of neural language models for paraphrasing is a recent trend, data for the training of PDS is lacking.

This paper contributes to the development of future detection methods for paraphrased text by providing, to our knowledge, the first large-scale dataset of text paraphrased using Transformer-based language models. We study how word-embeddings and three Transformer-based models used for paraphrasing (BERT [2], RoBERTa [8], and Longformer [9]) perform in classifying paraphrased text to underline the difficulty of the task and the dataset's ability to reflect it. The **dataset and source code** of our study are publicly available[3]. We grant access to the source code after accepting the terms and conditions designed to prevent misuse. Please see the repository for details.

II. RELATED WORK

Paraphrase identification is a well-researched NLP problem with numerous applications, e.g., in information retrieval and digital library research [6]. To identify paraphrases, many approaches combine lexical, syntactical, and semantic text analysis [7]. The Microsoft Research Paraphrase Corpus (MRPC) [10], a collection of human-annotated sentence pairs extracted from news articles, is among the most widely-used datasets for training and evaluating paraphrase identification methods. Another popular resource for paraphrase identification is the *Quora Question Pairs* (QQP) dataset included in GLUE [11]. The dataset consists of questions posted on Quora[4], a platform on which users can ask for answers to arbitrary questions. The task is to identify questions in the dataset that share the same intent. The datasets published as part of the PAN workshop series on plagiarism detection, authorship analysis, and other forensic text analysis tasks[5] are the most comprehensive and widely-used resource for evaluating plagiarism detection systems.

Neither the PAN nor the MRPC and QQP datasets include paraphrases created using state-of-the-art neural language models. The MRPC and QQP datasets consist of human-made content, which is unsuitable for training classifiers to recognize machine-paraphrased text. The PAN datasets contain cases that were obfuscated using basic automated heuristics that do not maintain the meaning of the text. Examples of such heuristics include randomly removing, inserting, or replacing words or

---

[1]https://spinbot.com
[2]https://spinnerchief.com/

[3]https://doi.org/10.5281/zenodo.4621403
[4]https://quora.com/about
[5]https://pan.webis.de/

TABLE I

AN ILLUSTRATIVE SAMPLE FOR EACH PARAPHRASING MODEL AND DATA SOURCE. RED BACKGROUND HIGHLIGHTS CHANGED TOKENS COMPARED TO THE ORIGINAL VERSION. THE ELLIPSIS "..." INDICATES THE REMAINDER OF THE PARAGRAPH.

Original Parapgraphs:

– A mathematically rigorous approach to quantum field theory based on operator algebras is called an algebraic quantum field theory...
– "Nuts" contains 5 instrumental compositions written and produced by Streisand, with the exception of "The Bar", including additional writing from Richard Baskin. All of the songs were recorded throughout 1987...
– Agriculture is the foundation for economic growth, development and poverty annihilation in developing countries. Ghana is endowed with a variety of mineral and agricultural product (Breisinger, 2008) Ghana is a country...

BERT Paraphrased	**Source**	**MLM Prob.**
The mathematically rigorous approach to quantum field theory based upon operator equations is called an algebraic Quantum field theory...	arXiv	0.15

RoBERTa Paraphrased		
"Nuts" contains five instrumental compositions written or produced by Streisand, with the exception of "Yourbars", which includes credited writing from Richard Baskin. All of these songs were recorded in 1987...	Wikipedia	0.15

Longformer Paraphrased		
Agriculture is the foundation builder for economic growth, development and poverty annihilation in developing countries. Ghana is endowed with a variety of biodiversity and agricultural product (Breisinger, 2008) Ghana became a country...	thesis	0.15

phrases and substituting words with their synonyms, antonyms, hyponyms, or hypernyms selected at random [12]. These cases are not representative of the sophisticated paraphrases produced by state-of-the-art Transformer-based models.

Currently, the HuggingFace API offers few neural language models capable of paraphrasing text excerpts. Most models are based on the same technique and trained to process short sentences. Plagiarists reuse paragraphs most frequently [6]. Hence, the ability to identify paragraph-sized paraphrases is most relevant for a PDS in practice. Prior to our study, no dataset of paragraphs paraphrased using Transformer-based models existed and could be used for training PDS.

Prior studies mitigated the lack of suitable datasets by paraphrasing documents using the paid services SpinBot and SpinnerChief [7], [13]. As the evaluations in these studies showed, text obfuscated by these tools already poses a significant challenge to current plagiarism detection systems. Nevertheless, the sophistication of the paraphrased text obtained from such tools to date is lower than that of paraphrases generated by Transformer-based models. Therefore, we extend the earlier studies [7] and [13] by using Transformer-based architectures [1] to generate paraphrases that reflect the strongest level of disguise technically feasible to date.

III. DATASET CREATION

Our neural machine-paraphrased dataset is derived from previous studies [7], [13]. The dataset of Foltynek et al. [7] consists of *featured Wikipedia articles*[6] in English. The dataset of Wahle et al. [13] comprises scientific papers randomly sampled from the *no problems* category of the arXMLiv[7] project, and randomly selected graduation theses by *English as a Second Language* (ESL) students at the Mendel University in Brno, Czech Republic.

[6]https://en.wikipedia.org/wiki/Wikipedia:Content_assessment
[7]https://kwarc.info/projects/arXMLiv/

TABLE II
OVERVIEW OF THE ORIGINAL PARAGRAPHS IN OUR DATASET.

Features	arXiv	Theses	Wiki	Wiki-Train
Paragraphs	20 966	5 226	39 241	98 282
# Words	3 194 695	747 545	5 993 461	17 390 048
Avg. Words	152.38	143.04	152.73	176.94

The earlier studies employed the paid online paraphrasing services SpinBot and SpinnerChief for text obfuscation. Since we investigate neural language models for paraphrasing, we only use the 163 715 original paragraphs from the earlier dataset. Table II shows the composition of these original paragraphs used for our dataset.

For paraphrasing, we used BERT [2], RoBERTa [8], and Longformer [9]. We chose BERT as a strong baseline for transformer-based language models; RoBERTa and Longformer improve BERT's architecture through more training volume and an efficient attention mechanism, respectively. More specifically, we used the masked language model (MLM) objective of all three Transformer-based models to create the paraphrases. The MLM hides a configurable portion of the words in the input, for which the model then has to infer the most probable word-choices. We excluded named entities and punctuation, e.g., brackets, digits, currency symbols, quotation marks from paraphrasing to avoid producing false information, or inconsistent punctuation compared to the original source. Then, we masked words and forwarded them to each model to obtain word candidates and their confidence scores. Lastly, we replaced each masked word in the original with the corresponding candidate word having the highest confidence score. Examples of original and paraphrased text using different models and data sources are illustrated in Table I. We also experimented with sampling uniformly from the top-k word predictions but neglected this method because of poor paraphrasing quality.
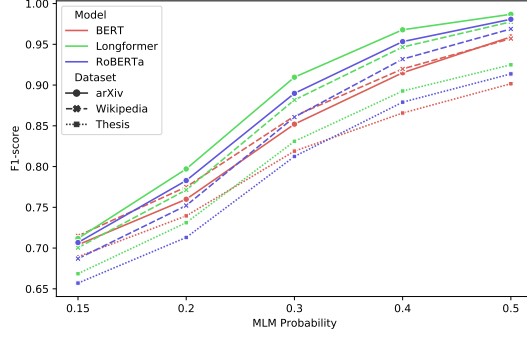
Fig. 1. Classification accuracy of fastText + SVM for neural-paraphrased test sets depending on masked language model probabilities.
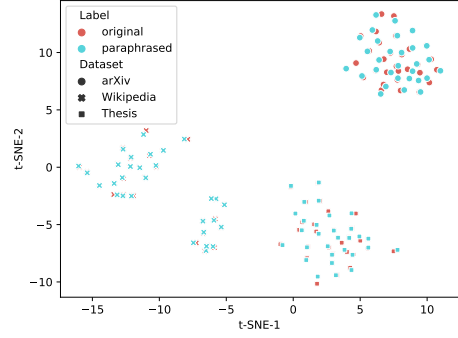

Fig. 2. Two-dimensional representation of BERT embeddings for 30 original and paraphrased paragraphs from each source. The overlap of the embeddings suggests semantic equivalence of the original and paraphrased content.

We ran an ablation study to understand how the masking probability of the MLM affects the difficulty of classifying documents as either paraphrased or original. For this purpose, we employed each neural language model with varying masking probabilities to paraphrase the arXiv, theses, and Wikipedia subsets. We encoded all original and paraphrased texts as features using the sentence embedding of fastText (subword)[8], which was trained on a 2017 dump of the full English Wikipedia, the UMBC WebBase corpus, and StatMT news dataset with 300 dimensions. Lastly, we applied the same SVM classifier to all fastText feature representations to distinguish between original and paraphrased content.

Fig. 1 shows the results of the ablation study. Higher masking probabilities consistently led to higher classification accuracy. In other terms, replacing more words reduced the difficulty of the classification task. This correlation has also been observed for non-neural paraphrasing tools [13]. Paragraphs from theses were most challenging for the classifier regardless of the paraphrasing model. We hypothesize that sub-optimal word choice and grammatical errors in the texts written by ESL students increase the difficulty of classifying these texts. The F1-scores for paragraphs from arXiv and Wikipedia were consistently higher than for theses. We attribute the high score on the Wikipedia test set to the documents' similarity with the training set which consists only of Wikipedia articles.

Masking 15% of the words posed the hardest challenge for the classifier. This ratio corresponds to the masking probability used for pre-training BERT [2], and falls into the percentage range of words that paid online paraphrasing tools replace on average (12.58% to 19.37%) [13]. Thus, we used a masking probability of 15% for creating all paraphrased data.

As a proxy for paraphrasing quality, we evaluated the semantic equivalence of original and paraphrased text. Specifically, we analyzed the BERT embeddings of 30 randomly selected original paragraphs from arXiv, theses, and Wikipedia and their paraphrased counterparts created using BERT, RoBERTa, and Longformer. Fig. 2 visualizes the results using a t-distributed Stochastic Neighbor Embedding (t-SNE) for

dimensionality reduction. The embeddings of original and paraphrased text overlap considerably despite changing approx. 15% of the words. This indicates the Transformer-based language models maintain the original text's semantics.

IV. CLASSIFICATION BENCHMARK

To check whether our dataset poses a realistic challenge for state-of-the-art classifiers and to establish a performance benchmark, we employed four models to label paragraphs as either original or paraphrased. A prior study showed that current plagiarism detection systems, which are essentially text-matching software, fail to identify machine-paraphrased text reliably while word embeddings, machine-learning classifiers, and particularly Transformer-based models performed considerably better [13]. Therefore, we evaluated the classification effectiveness of the three BERT-related models used for paraphrasing and the fastText + SVM classifier we applied and described for our ablation study (cf. Section III). We limited the number of input tokens for each model to 512 for a fair comparison of the models without losing relevant context information[9]. Unless specified differently, we used all hyperparameters in their default configuration.

We derived training data exclusively from Wikipedia as it is the largest of the three collections. We used arXiv papers and theses to obtain test sets that allow verifying a model's ability to generalize to data from sources unseen during training. We used BERT to generate the paraphrased training set (Wiki-Train) and BERT, RoBERTa, and Longfomer to create three paraphrased test sets. The classification models were exposed to mutually exclusive paragraphs to avoid memorizing the differences between aligned paragraphs. Evaluating each model using text paraphrased by the same model allows us to verify an assumption from related work, i.e., the best classifier is the language model used to generate the paraphrased text [14].

Table III shows the F1-Macro scores of each classification model for the paraphrased test sets consisting of arXiv, theses, and Wikipedia paragraphs. The baseline model (fastText +

---

[8] https://fasttext.cc/docs/en/english-vectors.html

[9] 99.35% of the datasets' text can be represented with less than 512 tokens.

TABLE III
CLASSIFICATION RESULTS (F1-MACRO SCORES). **BOLDFACE** SHOWS THE
BEST RESULT PER CLASSIFICATION MODEL.

Classification Model	Dataset	Paraphrase Model		
		BERT	**RoBERTa**	**Longformer**
fastText + SVM (baseline)	arXiv	70.40%	70.68%	71.17%
	Theses	68.94%	65.70%	66.85%
	Wikipedia	71.50%	68.70%	70.05%
	Average	70.28%	68.36%	69.36%
BERT	arXiv	80.83%	68.90%	68.49%
	Theses	74.74%	67.39%	66.04%
	Wikipedia	83.21%	68.85%	69.46%
	Average	**79.59%**	68.38%	68.00%
RoBERTa	arXiv	70.41%	85.40%	82.95%
	Theses	68.99%	79.13%	77.76%
	Wikipedia	72.18%	84.20%	82.15%
	Average	70.53%	**82.91%**	80.95%
Longformer	arXiv	65.18%	85.46%	89.93%
	Theses	65.72%	77.96%	81.31%
	Wikipedia	69.98%	81.76%	86.03%
	Average	66.96%	81.73%	**85.76%**

SVM) performed similarly for all paraphrasing models with scores ranging from F1=68.36% (RoBERTa) to F1=70.28% (BERT). With scores ranging from F1=79.59% (BERT) to F1=85.76% (Longformer), neural language models consistently identified text paraphrased using the same model best. This observation supports the findings of Zellers et al. [14].

Neural language models applied to paraphrases created by other models (e.g., BERT classifies text paraphrased by Longformer), typically achieved comparable scores to fast-Text+SVM. The average F1-scores for text paraphrased by unseen models range from F1=68.00% (BERT for Longformer paraphrases) to F1=81.73% (Longformer for RoBERTa paraphrases) with an average of 72.75%. These results are lower than the average scores for classifying paraphrases created for the same subsets using paid paraphrasing services (i.e., F1=99.65% to F1=99.87% for SpinBot) [13]. This finding shows Transformer-based neural language models produce hard-to-identify paraphrases, which make our new dataset a challenging benchmark task for state-of-the-art classifiers.

RoBERTa and Longformer achieved comparable results for all datasets, which we attribute to their overlapping pre-training datasets. BERT uses a subset of RoBERTa's and Longformer's training data and identifies the text paraphrased by the other two models with comparable F1-scores.Averaged over all paraphrasing techniques, RoBERTa achieved the best result (F1=78.15%), making it the most general model we tested for detecting neural machine-paraphrases.

All classification models performed best for Wikipedia articles, which is expected given their overlapping training corpus. The three neural language models identified arXiv articles similarly well which is in line with our ablation study (cf. Fig. 1). As in our ablation study, theses by ESL students were most challenging for our classification models, again corroborating our assumption that a higher ratio of gram-matical and linguistic errors causes the drop in classification effectiveness.

V. CONCLUSION AND FUTURE WORK

We presented a large-scale aligned dataset[3] of original and machine-paraphrased paragraphs to foster the research on plagiarism detection methods. The paragraphs originate from arXiv papers, theses, and Wikipedia articles and have been paraphrased using BERT, RoBERTa, and Longformer. We showed that the machine-paraphrased texts have a high semantic similarity to their original sources which reinforces our manual observation that neural language models produce hard-to distinguish, human-like paraphrases.

Furthermore, we showed Transformers are comparable in classifying original and paraphrased content to static word embeddings (i.e., fastText) and most effective for identifying text that was paraphrased using the same model. RoBERTa achieved the best overall result for detecting paraphrases.

In our future work, we will investigate other autoencoding models, and add autoregressive models to our study such as GPT-3 [3] for paraphrase generation and detection.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems 30*. Curran Associates, Inc., 2017, pp. 5998–6008.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv:1810.04805 [cs]*, May 2019, arXiv: 1810.04805.

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, and D. ... Amodei, "Language Models are Few-Shot Learners," *arXiv:2005.14165 [cs]*, Jun. 2020, tex.ids: BrownMRS20a arXiv: 2005.14165.

[4] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in Vision: A Survey," *arXiv:2101.01169 [cs]*, Feb. 2021, arXiv: 2101.01169.

[5] N. Dehouche, "Plagiarism in the age of massive generative pre-trained transformers (gpt-3)," *Ethics in Science and Environmental Politics*, vol. 21, pp. 17–23, 2021.

[6] T. Foltýnek, N. Meuschke, and B. Gipp, "Academic Plagiarism Detection: A Systematic Literature Review," *ACM Computing Surveys*, vol. 52, no. 6, pp. 112:1–112:42, 2019.

[7] T. Foltýnek, T. Ruas, P. Scharpf, N. Meuschke, M. Schubotz, W. Grosky, and B. Gipp, "Detecting Machine-obfuscated Plagiarism," in *Proceedings of the iConference 2020*, ser. LNCS. Springer, 2020.

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A Robustly Optimized BERT Pretraining Approach," *arXiv:1907.11692 [cs]*, Jul. 2019, arXiv: 1907.11692.

[9] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The Long-Document Transformer," *arXiv:2004.05150 [cs]*, Apr. 2020, arXiv: 2004.05150.

[10] B. Dolan and C. Brockett, "Automatically constructing a corpus of sentential paraphrases," in *Third International Workshop on Paraphrasing (IWP2005)*. Asia Fed. of Natural Language Processing, January 2005.

[11] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman, "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding," *arXiv:1804.07461 [cs]*, Feb. 2019.

[12] M. Potthast, B. Stein, A. Barrón-Cedeño, and P. Rosso, "An Evaluation Framework for Plagiarism Detection," in *Proceedings Int. Conf. on Computational Linguistics*, vol. 2, pp. 997–1005.

[13] J. P. Wahle, T. Ruas, T. Foltýnek, N. Meuschke, and B. Gipp, "Identifying Machine-Paraphrased Plagiarism," *arXiv:2103.11909 [cs]*, Jan. 2021.

[14] R. Zellers, A. Holtzman, H. Rashkin, Y. Bisk, A. Farhadi, F. Roesner, and Y. Choi, "Defending Against Neural Fake News," *arXiv:1905.12616 [cs]*, Oct. 2019, arXiv: 1905.12616.