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


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], [7]. 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 [8]. 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[3], 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 the three Transformer-based models used for paraphrasing (BERT [2], RoBERTa [9], and Longformer [10]) 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[4]. 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 [8]. The Microsoft Research Paraphrase Corpus (MRPC) [11], 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 the General Language Understanding Evaluation (GLUE) resource collection [12]. The dataset consists of questions posted on Quora[5], 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[6] 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

1https://spinbot.com
2http://www.spinnerchief.com/
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 phrases and substituting words with their synonyms, antonyms, hyponyms, or hypernyms selected at random [13]. These cases are not representative of the sophisticated paraphrases produced by state-of-the-art Transformer-based models.

Currently, the HuggingFace API\(^3\) 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], [14]. 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 [8], [15]. 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 is date lower than that of paraphrases generated by Transformer-based models. Therefore, we extend the earlier studies [8] and [15] 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 the datasets of the studies [8] and [15]. The dataset of Foltynek et al. [8] consists of featured Wikipedia articles\(^7\) in English. The dataset of Wahle et al. [15] comprises scientific papers randomly sampled from the no problems category of the arXMLiv\(^8\) project, and randomly selected graduation theses by English as a Second Language (ESL) students at the Mendel University in Brno, Czech Republic.

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 I shows the composition of these original paragraphs used for our dataset.

For paraphrasing, we used BERT [2], RoBERTa [9], and Longformer [10]. The well-established model BERT offers a strong baseline; 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. Table II shows the proportion of ignored tokens per paragraph. 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. We also experimented with sampling uniformly from the top-k word predictions but neglected this method because of poor paraphrasing quality.

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\(^9\), 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\(^10\) 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 [15]. 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.


\(^{8}\)https://kwarc.info/projects/arXMLiv/

\(^{9}\)https://fasttext.cc/docs/en/english-vectors.html

\(^{10}\)https://scikit-learn.org/stable/modules/svm.html
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%) [15]. 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\(^\text{11}\) (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 [15]. 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\(^\text{12}\). 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 (WikiTrain) 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 [16].

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 + 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. [16].

Neural language models applied for classifying paraphrases created by other models (e.g., BERT classifies text paraphrased by Longformer), typically outperformed static word embeddings and achieved comparable scores otherwise. 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

\(^{11}\)https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html

\(^{12}\)99.35% of the datasets’ text can be represented with less than 512 tokens.
been paraphrased using BERT, RoBERTa, and Longformer. from arXiv papers, theses, and Wikipedia articles and have machine-paraphrased paragraphs to foster the research causes the drop in classification effectiveness. tion that a higher ratio of grammatical and linguistic errors for our classification models, again corroborating our assump-

neural language models identified arXiv articles similarly well articles, which is an outcome we expected given our training 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 pretraining 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, further supporting this hypothesis. 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 an outcome we expected given our training corpus exclusively consists of Wikipedia content. 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 grammatical and linguistic errors causes the drop in classification effectiveness.

V. CONCLUSION AND FUTURE WORK

We presented a large-scale aligned dataset 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. To the best of our knowledge, this is the first time that the three Transformer-based language models have been applied for this use case. 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 more effective in classifying original and paraphrased content than 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 plan to investigate the performance of autoregressive models such as T5 [7] and GPT-3 [3] for paraphrase generation and detection.

REFERENCES


TABLE III

<table>
<thead>
<tr>
<th>Classification Model</th>
<th>Dataset</th>
<th>Paraphrase Model</th>
<th>F1-score</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>fastText + SVM (baseline)</td>
<td>arXiv</td>
<td>BERT</td>
<td>70.40%</td>
<td>70.68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RoBERTa</td>
<td>70.40%</td>
<td>70.68%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longformer</td>
<td>70.40%</td>
<td>70.68%</td>
</tr>
<tr>
<td>BERT</td>
<td>arXiv</td>
<td>BERT</td>
<td>80.83%</td>
<td>68.90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RoBERTa</td>
<td>80.83%</td>
<td>68.90%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longformer</td>
<td>80.83%</td>
<td>68.90%</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>arXiv</td>
<td>BERT</td>
<td>74.74%</td>
<td>67.39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RoBERTa</td>
<td>74.74%</td>
<td>67.39%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longformer</td>
<td>74.74%</td>
<td>67.39%</td>
</tr>
<tr>
<td>Longformer</td>
<td>arXiv</td>
<td>BERT</td>
<td>65.18%</td>
<td>85.46%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>RoBERTa</td>
<td>65.18%</td>
<td>85.46%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Longformer</td>
<td>65.18%</td>
<td>85.46%</td>
</tr>
</tbody>
</table>

BibTeX:

```bibtex
@InProceedings{Wahle2021,
    title = {Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection},
    booktitle = {Proceedings of the {ACM}/{IEEE} {Joint} {Conference} on {Digital} {Libraries} ({JCDL})},
    author = {Wahle, Jan Philip and Ruas, Terry and Meuschke, Norman and Gipp, Bela},
    month = {Sep.},
    topic = {pd},
    year = {2021}
}
```

RIS:

```
TY - CONF
TI - Are Neural Language Models Good Plagiarists? A Benchmark for Neural Paraphrase Detection
AU - Wahle, Jan Philip
AU - Ruas, Terry
AU - Meuschke, Norman
AU - Gipp, Bela
ID - Wahle2021
C3 - Proceedings of the ACM/IEEE Joint Conference on Digital Libraries (JCDL)
DA - 2021/04/
PY - 2021
DO - 
ER -
```

Related Publications: [www.gipp.com/pub](http://www.gipp.com/pub)