Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4692–4700
Marseille, 20-25 June 2022

A Benchmark Corpus for the Detection of Automatically Generated Text in Academic Publications

Vijini Liyanage, Davide Buscaldi, Adeline Nazarenko
Université Sorbonne Paris Nord, LIPN, CNRS, UMR 7030
F-93430, Villetaneuse, France
{liyanage, davide.buscaldi, adeline.nazarenko}@lipn.univ-paris13.fr

Abstract

Automatic text generation based on neural language models has achieved performance levels that make the generated text almost indistinguishable from those written by humans. Despite the value that text generation can have in various applications, it can also be employed for malicious tasks. The diffusion of such practices represent a threat to the quality of academic publishing. To address these problems, we propose in this paper two datasets comprised of artificially generated research content: a completely synthetic dataset and a partial text substitution dataset. In the first case, the content is completely generated by the GPT-2 model after a short prompt extracted from original papers. The partial or hybrid dataset is created by replacing several sentences of abstracts with sentences that are generated by the Arxiv-NLP model. We evaluate the quality of the datasets comparing the generated texts to aligned original texts using fluency metrics such as BLEU and ROUGE. The more natural the artificial texts seem, the more difficult they are to detect and the better is the benchmark. We also evaluate the difficulty of the task of distinguishing original from generated text by using state-of-the-art classification models.

Keywords: Automatic Text Generation, GPT-2, Arxiv-NLP, Detection, Dataset, Classification

1. Introduction

The Transformer model (Vaswani et al., 2017) can be considered as a major milestone in the domain of deep learning and Natural Language Processing (NLP). Since then, various forms of state-of-the-art transformer models such as the Generative Pre-training model (GPT) (Radford et al., 2018), BERT (Devlin et al., 2018) and Transformer-XL (Dai et al., 2019) have been introduced and utilized for a diverse amount of NLP tasks. To cite some, Natural Language Generation (NLG) (Radford et al., 2019), text classification (Yang et al., 2019), machine translation (Lample and Conneau, 2019) and text summarization (Lewis et al., 2019). The aforementioned research show that the transformer models are capable of producing outstanding results.

The GPT model has been recently updated to significantly improve the quality of the generated text. For example, the GPT-2 model (Radford et al., 2019) is competent enough to produce natural text that looks like as if written by a human (Solaiman et al., 2019). Regardless of the valuable contribution provided by these models for the betterment of NLP in general, some concerns have been raised about the a potential risks associated to such models. These models can be misused for malicious tasks such as fake news generation (Zellers et al., 2019), (Vosoughi et al., 2018), fake review generation (Adelani et al., 2020) and viral story generation (Faris et al., 2017), (Wardle and Derakhshan, 2017).

Recently, some of these models have been applied for the computer-assisted writing of research papers (Wang et al., 2019), reviews (Wang et al., 2020) or theses (Abd-Elaal et al., 2019). Despite the advantages in alleviating researchers' workload, a risk for misuse of these technologies exists. A recent work (Cabanac and Labbé, 2021) shows that old textual generation models based on context-free grammars, such as SciGen[1], are being actively used in academic publishing, although their detection is relatively easy since they tend to produce nonsense or "tortured phrases" – weirdly paraphrased versions of scientific terms. Therefore, immediate research on detection of academic texts that are artificially generated is imperative. In this way, researchers will also have a tool to determine whether these powerful models were used in a responsible way or not. To develop such detection methods, there is a vital requirement of a corpus composed of automatically generated academic content. This is the main purpose of the present research.

2. Related Work

2.1. Human Detection of Machine Generated Text

Several research works have been conducted regarding human-detection of machine generated text. (Bakhtin et al., 2019) considers human-detection as a ranking task. Based on a thorough analysis of how human detection is affected by factors such as sampling method and the length of the text excerpt, the authors of (Ippolito et al., 2019b) consider that human detectors perform well in finding semantic errors in machine generated text. (Gehrmann et al., 2019) claims that the accuracy of human detection of artificially generated text without any tool is only 54%. (Ippolito et al., 2019a) has identified that humans focus mostly on semantic er-

[1] `https://github.com/strib/scigen.git`

rors, while the discriminative models such as fine-tuned BERT focus more on statistical artifacts. One of the latest tools, RoFT (Real or Fake Text) tool (Dugan et al., 2020), is used to evaluate human detection showing that the text generation models are capable to fool humans by one or two sentences. A recent research (Clark et al., 2021) shows that training humans on evaluation task for GPT-3 generated text only improves the overall accuracy upto 50%.Despite the interest in measuring the ability of humans to detect automatically generated text, not much research has been conducted to develop automatic tools to distinguish machine generated text from human written text.

2.2. Automatic Detection of Machine Generated Text

The introduction of the statistical model GLTR (Giant Language model Test Room) (Gehrmann et al., 2019) can be considered as a major milestone for the detection of automatically generated text. The authors consider the stylometric details of texts by incorporating three types of tests: the probability of the word, the absolute rank of a word and the entropy of the predicted distribution. Afterwards, they compute per-token likelihoods and visualize histograms over them to support humans in detection of automatically generated content. A recent research (Al-Kadhimi and Löwenström, 2020) has extended the work done in GLTR by feeding the output of GLTR to a Convolutional Neural Network, which automatically classifies whether the input reviews are human written or machine generated.

Another turning point is the establishment of the GROVER (Zellers et al., 2019), in which it's architecture is a combination of a generation model and detection model. News are generated using a transformer based model which has an architecture similar to GPT-2 (Radford et al., 2019). But (Radford et al., 2019) have used conditional generation (on article meta data) and nucleus sampling. Afterwards, a zero shot detection is performed using a simple linear classifier on top of the pre-trained GROVER model. The authors have experimented detection with existing other models as well(fastText (Bojanowski et al., 2017) and BERT(Devlin et al., 2018)) reporting the highest accuracy for their own GROVER model and claiming that the best models in forming fake content are also the best models in detection. Contrary to that, Uchendu shows, however, that GROVER cannot correctly detect texts generated by language models other than GROVER itself.

Lots of research have leveraged the RoBERTa (Liu et al., 2019), a masked and non-generative language model to detect automatically generated text. (Solaiman et al., 2019) has proved that the discriminative model of RoBERTa outperforms generative models such as GPT-2 in detection task. Such findings contradict with GROVER authors' claim that the generative model is better in detecting text generated by itself. (Fagni et al., 2021) reveals that the RoBERTa can defeat traditional machine learning models, such as bag of words, and neural network models, such as RNNs and CNNs, regarding the detection of automatically generated tweets. Moreover, (Uchendu et al., 2020) shows that RoBERTa outperforms existing detectors in detecting automatically generated news articles and product reviews which are generated by state of the art models like GPT-2. Despite the success of RoBERTa, recent research (Jawahar et al., 2020) shows that its dependence on large amounts of data, limits limits its use for detection. (Wolff and Wolff, 2020) challenges the RoBERTa model by exposing it with homoglyph and misspelling attacks and their results show a drastic drop in recall. This shows that the future detection models should be robust against such attacks. To test and evaluate the detection models, it is of key importance to have a dataset that incorporates such attacks or traps, if possible independent of any specific detection method.

Many of the research on detection assumes that the generator is known ((Gehrmann et al., 2019), (Zellers et al., 2019)). Therefore their approaches are incapable in generalizing the settings so that it works well when the generator architectures and corpora are different in training and testing stages. However, in real world settings, a detection model faces indeterminate and evolving data. This issue has been addressed to a certain level by (Bakhtin et al., 2019), which provides a thorough experimental setup and quantitative results. (Ippolito et al., 2019a) fine-tunes the BERT model for detection and analyzes how factors like sampling strategies and text excerpt length impact the detection task. Another research (Varshney et al., 2020) produces a formal hypothesis testing framework and sets error exponents limits (in terms of perplexity and cross-entropy) for large scale models such as GPT-2 and CTRL in order to find limits in detecting text generated by them. One of the latest research (Maronikolakis et al., 2020) leverages Transformers to detect headlines generated by GPT-2 model. In this approach, the models learn from the datasets and classifies text as machine generated text or human-written text. It makes use of 4 types of classifiers: Baselines (Logistic Regression, Elastic Net), Deep learning (CNN, Bi-LSTM, Bi-LSTM with Attention), Transfer learning via ULMFit and Transformers (BERT, DistilBERT) for the detection task. The results show that BERT scores an overall accuracy of 85.7%.

In addition to transformer based models many research works conducted for the detection make use of other types of deep learning models as well as statistical models. (Vijayaraghavan et al., 2020) experiments numeric representation such as Tf-idf and word2vec, as well as neural networks such as ANNs and LSTMs on detection of fake news. A latest research (Harada et al., 2021) suggests two approaches – CPCO (Consistency of anti preceding sentence using Cosine words Overlapping), and CICO (Consistency of opposing Input sentences using Cosine words Overlapping) –, which utilize sentence coherence for the detection task. Also (Jawahar,) leverages different discourse models for detection. By exposing style-based classifiers to syntactic and semantic permutations, (Bhat and Parthasarathy, 2020) shows the limitations of style-based classifiers which highly rely on provenance to detect fake text. Furthermore, (Pérez-Rosas et al., 2017) highlights the importance of linguistic features such as semantic, syntactic and lexical features in distinguishing the machine generated news from human written news.

Although there are lots of research conducted for the detection of automatically generated content such as news paper articles, reviews, tweets, headlines and so on, only a limited number of detection research is dedicated to academia. These works (Xiong and Huang, 2009; Lavoie and Krishnamoorthy, 2010) have mainly focused on the authenticity of the references. (Amancio, 2015) has examined topological properties in natural and generated papers as a source for detection. (Williams and Giles, 2015) has proposed various measures in detecting generated academic content. SciDetect (Nguyen and Labbé, 2016) is another research which considers inter-textual distance using all the words and nearest neighbour classification for detection. The authors have analyzed the existing methods for detecting automatically generated paper and relied on Probabilistic Context Free Grammar (PCFG) to detect fake academia. However, their approach relies on the fact that word distributions are identical to that of human written content, an assumption that is not always verified. A recent research (Cabanac and Labbé, 2021) proposes a rule based detection mechanism which leverage third party search engines to distinguish automatically generated papers based on improbable word sequences found in them, but their approach can only detect grammar based computer generated papers. However, among all the works about the detection of automatically generated text, there is no research on texts in the academic or scholarly domain, despite the availability of such data and the potential danger in the misuse of generative models in this domain.

In order to leverage the deep learning models to detect automatically generated research content from human written content, a corpus of artificially generated academic content is required. Many latest research (Cabanac and Labbé, 2021), (Xiong and Huang, 2009) have leveraged SCIgen to generate a dataset for their research. But SCIgen generates nonsense (because it focuses on amusement rather than coherence) using context free grammar. To test methods and encourage research in this area, we propose a benchmark of academic papers that are generated using state-of-the-art models like GPT-2 an that look as if they were written by humans. This paper presents the dataset that composes this benchmark, explain how it has been built and evaluate its usefulness for the detection of automatically generated academic texts.

3. Benchmark datasets

To the best of our knowledge, there are no publicly available corpora which can be used as a benchmark dataset to experiment the detection of automatically generated academic content. However, as explained above, such a benchmark is a prerequisite for any research in the field of fake text detection, which represents an important issue for the academic world. This paper presents datasets for this purpose.

We built two corpora: one containing automatically generated papers and a hybrid dataset which contains original (human written) abstracts in which some sentences are substituted with machine generated sentences. We designed them as a mixture of natural (written by humans) and artificial (generated) content that should be not easily detected by a machine and we used different generation strategies. In order to achieve this, we had to tune the models to generate content in a window of "credibility", or in other words that the generated content could appear as original to an uninformed reader. The automatically generated paper dataset and the hybrid dataset contain average word counts of 1243 and 177, respectively.

The first corpus is composed of artificially generated research papers, aligned to the original abstract from which the writing prompt was chosen. It corresponds to the situation in which a malicious author would create a full text to submit, in a similar way to Sci-Gen. Their length is variable as they can be composed from the abstract alone to more sections such as introduction, related work and conclusion. We eliminated sections such as Methodology and Results and Evaluation, because they mostly contain diagrams and tables which can reveal too easily their (synthetic) origin. Few examples showing content from the original paper and the respective generated paper are available in Table 1. Here the original papers correspond to the papers which were used as prompts for the models to perform generation.

The second corpus is a collection of abstracts which are composed of a combination of original content and machine generated sentences. It corresponds to a situation in which an author would recur to language generation to fill in certain parts of hers paper. To compose this corpus, we ignored papers containing the name of the proposed model, product or project in the proposal statement of the abstract, as the Arxiv-NLP model might suggest a name that would not be consistent with the rest of the abstract (the sentences extracted from the original abstract), thus making the generated abstract too easily distinguishable from human written abstracts. Table 2 shows some examples of the original abstracts and their corresponding generated abstracts.

The dataset is available at the following URL: `https://github.com/vijini/GeneratedTextDetection.git`.

Original Paper	Generated Paper
Finally, in this task, we used comparative evaluations to **determine** the quality of the sets of generated **KPs. Obviously, this results with rankingbased evaluation. Establishing an absolute evaluation metric in this context merits further investigation. We thank all participating teams for taking the time to participate in this challenging new shared task. We believe that these contributions, along with the data collected and shared in this report, will be valuable to further advance the research around KPA and related tasks.**	Finally, in this task, we used comparative evaluations to **estimate** the quality of the sets of generated **texts. We expect the task and the findings reported in this paper to be relevant for researchers working on text summarization and argument mining. Introduction Key Point Analysis (KPA) was introduced in BarHaim et al. (2020a,b) as a challenging NLP task with tight relations to Computational Argumentation, Opinion Analysis, and Summarization, and with many practical applications (Bar-Haim et al., 2021).**
We show that the mixed-domain transfer learning approach **does not always work well. We found** the **combination** of the **non-graph DSM (for capturing local relationships) and CNN (for extracting global relationships) to be a promising approach towards creating a gener- alizable solution for detecting COVID-19 online misinformation. In the future, we plan to investigate the gener- alizability of the DSM models created using deep learning architectures such as BERT.**	We show that the mixed-domain transfer learning approach **is more effective across most of** the **dimensions** of the **COVID-19 data. Specifically, we show that the shallow mixed-domain transfer learning approach is more effective across the following dimensions of the COVID-19 data: granular (grained), shallow (skewed), and deep (deeper). We show that the shallow mixed-domain transfer learning approach is more effective across the following dimensions of the COVID-19 data: temporal dimension (the context in the dataset evolves), length dimension.**

Table 1: Some examples of original *vs.* generated papers in the "fully generated" corpus.

Original Abstract	Generated Abstract
Our experiments suggest **that models possess belief-like qualities to only a limited extent, but update methods can both fix incorrect model beliefs and greatly improve their consistency.** Although off-the-shelf optimizers are surprisingly strong belief- updating baselines, our learned optimizers can outperform them in more difficult settings than have been considered in past work.	Our experiments suggest **the importance of model beliefs in learning models, and we show that the approach outperforms automatic model updating systems using word representations.** Although off-the- shelf optimizers are surprisingly strong belief- updating baselines, our learned optimizers can outperform them in more difficult settings than have been considered in past work.
Simultaneously evolving morphologies (bodies) and controllers (brains) of robots can cause a mismatch between the inherited body and brain in the offspring. To mitigate this problem, the addition of an infant learning period by the so-called Triangle of Life framework has been proposed relatively long ago. However, an empirical assessment is still lacking to-date. In this paper we **investigate the effects of such a learning mechanism from different perspectives.**	Simultaneously evolving morphologies (bodies) and controllers (brains) of robots can cause a mismatch between the inherited body and brain in the offspring. To mitigate this problem, the addition of an infant learning period by the so-called Triangle of Life framework has been proposed relatively long ago. However, an empirical assessment is still lacking to-date. In this paper , **we present a method to evaluate the effect of an algorithm based on the development of a hybrid human/bot learning framework, which combines the development of both a hybrid robot and a human model on the same domain.**

Table 2: Some examples of original *vs.* generated abstracts from the "hybrid" corpus.

4. Benchmark Design Methodology

For the first corpus ("fully generated"), we used the GPT-2 model for the generation of research papers and the model was fine-tuned by feeding it with papers extracted from ArXiv[2] . We specifically selected the Computation and Language domain and the 100 latest papers were chosen to make sure that the papers we considered were not used to train the original GPT-2

[2] `https://arxiv.org`

model. Moreover, papers with IEEE citation style were not selected to train the model, because this style represent citations only as numbers, which does not allow a reader to verify whether the citations are appropriate (we do not consider the "References" section for model training). Each paper was separately used to fine-tune the GPT-2 model. By choosing the first 50 words of each original paper as the seed text, a new paper which is of same length as the original paper was generated using the fine-tuned model. Likewise, 100 new papers were generated. The temperature parameter was set to 0.7 to make sure that the generated content are neither too much random nor too much alike the original paper. For clarity purposes, we ignored the sections such as methodology, results and evaluation and discussion which contain diagrams, tables, equations.

The second ("hybrid") corpus was generated by combining the authentic content and the machine generated content. For this task we utilized some of the latest abstracts from Artificial Intelligence domain in ArXiv. Each hybrid abstract is made of 4 parts. The initial content is extracted from an original abstract up to the point where it reveals about the proposal (*e.g.* "In this paper,", "We propose", "Here we"). The next sentence is generated until the first full-stop, using the Arxiv-NLP provided by the Huggingface team (Wolf et al., 2020). Then, the rest of the original abstract is copied until the point that corresponds to the conclusion (*e.g.* "We conclude", "Our results show that"). Again, using the Arxiv-NLP, the rest of the abstract is generated. Likewise 100 new abstracts were composed. The temperature parameter was set to 1 and top-p was 0.9. This generation is done with human intervention, so that it is biased towards the objective strategy of making the generated content difficult to detect.

5. Evaluation

We present a double approach to evaluate the utility of the produced corpora for the task of classifying artificially generated and human written academic texts in a context where neural-based generation models have become common. We first evaluate the intrinsic quality of the generated texts, assuming that the more natural they seem the more difficult and the more misleading they are for the detection methods. We also perform an application based evaluation using a panel of state-of-art detection models to assess the difficulty of the classification task and to check that our benchmark is not biased towards a specific detection approach.

5.1. Evaluation of the quality of the generated texts

To assess the instrinsic quality of our benchmark, we leveraged two metrics: BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and ROUGE (Recall-Oriented Understudy for Gisting Evaluation) (Lin, 2004) to compare the generated contents with their respective original contents. BLEU and ROUGE are two traditional metrics used to compare the candidate (generated) text with the reference (original) text. They are traditionally considered as fluency metrics that indicates how natural an artificial text is compared to a natural original one. BLEU is a precision based score while ROUGE is a recall based score. We calculated the BLEU score at uni-gram and sentence levels and ROUGE at uni-gram (1), bi-gram (2) and Longest Common Sub-sequence (L) levels. All the results with respect to BLEU and ROUGE scores are provided in Table 3 and Table 4 for the fully generated and hybrid datasets, respectively. For all the evaluations, 80/20 training/test split was done on the datasets.

In general, all the average figures are quite high (more than 0.8), which indicates that the artificial texts are quite similar to natural ones. Of course come of the generated texts could be identified as "copies" of the original texts but they have nevertheless been artificially generated and it is interesting to evaluate the detection models against these texts that "look" natural. The tables represent both the minimum and average scores, so that a reader can get an idea about the quality of the overall dataset by the average score, while minimum score provides an idea of the quality of individual fully generated or hybrid research articles. Minimum *vs.* average scores comparison shows that the methodology ensures a wide diversity of generated data: some artificial texts are quite different from the original ones and presumably easier to detect.

Rouge-1 and Rouge-L scores are almost similar when they are rounded off to three decimal places. But when we consider more decimal places, it can be seen that the Rouge-1 scores are slightly higher than the Rouge-L scores. If we consider overall Unigram *vs.* sentence BLEU, as the unigram BLEU metric is more tolerant than the sentence one, it is normal that the sentence BLEU scores, although quite high, are lower than the unigram BLEU scores.

When we look at the results on Table 3, it can be seen that the uni-gram BLEU scores are higher than uni-gram ROUGE scores. On the other hand, when we measure the scores considering a longer sequence (either sentence level in BLEU or Longest Common Sub-sequence in ROUGE), the BLEU scores are lower than ROUGE ones.

When we look at the Table 4, it can be seen that the average BLEU scores are always less than the respective ROUGE scores. When we refer the two tables 3 and 4, it can be seen that the average BLEU scores are always higher in the fully generated dataset compared to hybrid dataset, which means the fully generated dataset has a better precision (that is, a good portion of the generated n-grams are also in the original text). On the other hand, it can be seen that the average ROUGE scores of the fully generated dataset is lower than the hybrid dataset. This means that the recall of the hybrid dataset is at a better level compared to the other dataset (this is expected as parts of the original text are included in the generated one). The minimum scores are relatively high for both BLEU and Rouge, indicating that there is a good degree of similarity between the texts even when they are most dissimilar.

5.2. Evaluation of the classification difficulty

Since our datasets are to be used for benchmarking classification models, we also evaluate the difficulty of the classification task.

To verify that the proposed benchmark is not biased towards a specific model type, we consider a variety of models, *i.e.* statistical models such as Bag of Words (Harris, 1954) as well as deep learning based models. As Bag of Words models, we consider Multinomial Naive Bayes, Passive Aggressive Classifier Multinomial Classifier with Hyper-parameter (alpha) algorithms and Support Vector Machine (Cortes and Vapnik, 1995). For the vocabulary, we considered not only single words but n-grams of size 1 - 3 words. As deep learning based models, we consider basic LSTM, Bi-LSTM (with the same configuration used by (Maronikolakis et al., 2020)), BERT and DistilBERT (Sanh et al., 2019).

The classification results are presented in Table 5 and Table 6 for fully generated and hybrid datasets, respectively. As per the results, the highest classification score was obtained by the DistilBERT model regarding both the datasets: the scores are 62.5% and 70.2% for the fully generated dataset and the hybrid dataset, respectively. These results show that the generated corpora are competent enough to be used as a baseline datasets to experiment detection.

As expected, the scores differs from one model to the other, the deep learning based models having higher accuracy scores than the statistical models. The higher scores are obtained by DistilBERT model on both the datasets: 62.5% and 70.2% for the fully generated dataset and the hybrid dataset, respectively. Interestingly, if most models perform slightly better on the Hybrid Dataset, it is not the case of LSTM model which achieves a much better score on the Fully Generated Dataset and of BERT and passive aggressive classifier at lesser degrees. This is an argument for including both datasets and different types of generated data in our benchmark. Despite these differences, we observe globally that the accuracy scores are not very high, even for DistilBERT. This is the most important point to assess the quality of our benchmark in terms of classification difficulty.

We did a comparison of our results to the latest research works (Maronikolakis et al., 2020) and the results are depicted in Table 7. The overall accuracies regarding our datasets are lower when compared with the aforementioned research. This may be due to the fact that Maronikolakis *et al. focus on the generation of short content (headlines) but it shows that our datasets are more difficult to classify than theirs, which makes it a better benchmark proposal.*

6. Conclusion and Future Work

This paper presents a benchmark proposal for detecting automatically generated research content and describes how it has been produced. We have considered the detection as a binary classification task and tested various classification algorithms. The results show that the existing state of the art models for classification could provide a maximum accuracy of 70.2% on our dataset. This result shows that this problem is open and there is room for further improvement in term of accuracy.

Faced with the proliferation of automatically generated content and the risks of scientific misconduct that this represents for the academic world, it is essential to develop efficient detection methods to as to disqualify fake research articles or artificially reviews. The production of datasets of automatically generated academic texts at the level of quality of the best automatic content generation methods (such as deep learning based models) is an indispensable prerequisite. Our proposed benchmark fits into this framework, the next step being the design of classification methods capable of detecting artificially generated academic content.

As a future work, we plan to develop original methods oriented to the task of detecting automatically generated texts and fragments, possibly leveraging knowledge to detect contradictions and out-of-context sentences. Moreover, we hope to increase the size of the dataset by adding more papers to the corpus and thereby re-generating the results.

7. Acknowledgements

This work is funded by Université Sorbonne Paris Nord and is partially supported by a public grant overseen by the French National Research Agency (ANR) as part of the program “Investissements d’Avenir” (reference: ANR-10-LABX-0083).

8. Bibliographical References

Abd-Elaal, E., Gamage, S., Mills, J., et al. (2019). Artificial intelligence is a tool for cheating academic integrity. In *30th Annual Conference for the Australasian Association for Engineering Education (AAEE 2019): Educators Becoming Agents of Change: Innovate, Integrate, Motivate*, page 397. Engineers Australia.

Adelani, D., Mai, H., Fang, F., Nguyen, H., Yamagishi, J., and Echizen, I. (2020). Generating sentiment-preserving fake online reviews using neural language models and their human-and machine-based detection. In *International Conference on Advanced Information Networking and Applications*, pages 1341–1354. Springer.

Al-Kadhimi, S. and Löwenström, P. (2020). Identification of machine-generated reviews: 1d cnn applied on the gpt-2 neural language model.

Amancio, D. R. (2015). Comparing the topological properties of real and artificially generated scientific manuscripts. *Scientometrics*, 105(3):1763–1779.

Score	unigram level BLEU	sentence BLEU	Rouge-1	Rouge -2	Rouge-L
Min. Score	0.659	0.583	0.591	0.559	0.591
Avg. Score	0.867	0.809	0.853	0.810	0.853

Table 3: Minimum and Average BLEU and ROUGE Scores (Recall) for the Fully Generated Dataset.

Score	ngram level BLEU	sentence BLEU	Rouge-1	Rouge -2	Rouge-L
Min. Score	0.629	0.495	0.598	0.473	0.598
Avg. Score	0.824	0.792	0.882	0.840	0.881

Table 4: Minimum and Average BLEU and ROUGE Scores (Recall) for the Hybrid Dataset.

Bakhtin, A., Gross, S., Ott, M., Deng, Y., Ranzato, M., and Szlam, A. (2019). Real or fake? learning to discriminate machine from human generated text. *arXiv preprint arXiv:1906.03351*.

Bhat, M. M. and Parthasarathy, S. (2020). How effectively can machines defend against machine-generated fake news? an empirical study. In *Proceedings of the First Workshop on Insights from Negative Results in NLP*, pages 48–53.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Cabanac, G. and Labbé, C. (2021). Prevalence of nonsensical algorithmically generated papers in the scientific literature. *Journal of the Association for Information Science and Technology*.

Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., and Smith, N. A. (2021). All that's' human'is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*.

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

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Dugan, L., Ippolito, D., Kirubarajan, A., and Callison-Burch, C. (2020). Roft: A tool for evaluating human detection of machine-generated text. *arXiv preprint arXiv:2010.03070*.

Fagni, T., Falchi, F., Gambini, M., Martella, A., and Tesconi, M. (2021). Tweepfake: About detecting deepfake tweets. *Plos one*, 16(5):e0251415.

Faris, R., Roberts, H., Etling, B., Bourassa, N., Zuckerman, E., and Benkler, Y. (2017). Partisanship, propaganda, and disinformation: Online media and the 2016 us presidential election. *Berkman Klein Center Research Publication*, 6.

Gehrmann, S., Strobelt, H., and Rush, A. (2019). Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

Harada, A., Bollegala, D., and Chandrasiri, N. P. (2021). Discrimination of human-written and human and machine written sentences using text consistency. In *2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS)*, pages 41–47. IEEE.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2-3):146–162.

Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2019a). Automatic detection of generated text is easiest when humans are fooled. *arXiv preprint arXiv:1911.00650*.

Ippolito, D., Duckworth, D., Callison-Burch, C., and Eck, D. (2019b). Human and automatic detection of generated text.

Jawahar, G. (). Detecting human written text from machine generated text by modeling discourse coherence.

Jawahar, G., Abdul-Mageed, M., and Lakshmanan, L. V. (2020). Automatic detection of machine generated text: A critical survey. *arXiv preprint arXiv:2011.01314*.

Lample, G. and Conneau, A. (2019). Cross-lingual language model pretraining.

Lavoie, A. and Krishnamoorthy, M. (2010). Algorithmic detection of computer generated text. *arXiv preprint arXiv:1008.0706*.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2019). Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Model used	Accuracy
Bag of ngrams(1-3), Multinomial Naive Bayes Algorithm	19.7
Bag of ngrams(1-3), Passive Aggressive Classifier Algorithm	31.8
Bag of ngrams(1-3), Multinomial Classifier with Hyperparameter (alpha)	19.7
Bag of ngrams(1-3), SVM	37.9
LSTM model	59.1
Bi-LSTM (Latest Paper)	40.9
BERT	52.5
DistilBERT	**62.5**

Table 5: Classification Results for Fully Generated Dataset

Model used	Accuracy
Bag of ngrams(1-3), Multinomial Naive Bayes Algorithm	24.2
Bag of ngrams(1-3), Passive Aggressive Classifier Algorithm	30.3
Bag of ngrams(1-3), Multinomial Classifier with Hyperparameter (alpha)	22.7
Bag of ngrams(1-3), SVM	37.9
LSTM model	50.0
Bi-LSTM (Latest Paper)	47.0
BERT	50.0
DistilBERT	**70.2**

Table 6: Classification Results for Hybrid Dataset

Metric	Full-text Dataset	Hybrid Dataset	Maroniko-lakis et al., 2020
Bi-LSTM	40.9	47.0	82.8
BERT	52.5	50.0	**85.7**
DistilBERT	**62.5**	**70.2**	85.5

Table 7: Classification Result Comparison for Bi-LSTM, BERT and DistilBERT models

Maronikolakis, A., Schutze, H., and Stevenson, M. (2020). Identifying automatically generated headlines using transformers. *arXiv preprint arXiv:2009.13375.*

Nguyen, M. T. and Labbé, C. (2016). Engineering a tool to detect automatically generated papers. In *BIR 2016 Bibliometric-enhanced Information Retrieval.*

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Pérez-Rosas, V., Kleinberg, B., Lefevre, A., and Mihalcea, R. (2017). Automatic detection of fake news. *arXiv preprint arXiv:1708.07104.*

Radford, A., Narasimhan, K., Salimans, T., and Sutskever, I. (2018). Improving language understanding by generative pre-training.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. volume 1, page 9.

Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108.*

Solaiman, I., Brundage, M., Clark, J., Askell, A., Herbert-Voss, A., Wu, J., Radford, A., Krueger, G., Kim, J., Kreps, S., et al. (2019). Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203.*

Uchendu, A., Le, T., Shu, K., and Lee, D. (2020). Authorship attribution for neural text generation. In *Conf. on Empirical Methods in Natural Language Processing (EMNLP).*

Varshney, L. R., Keskar, N. S., and Socher, R. (2020). Limits of detecting text generated by large-scale language models. In *2020 Information Theory and Applications Workshop (ITA)*, pages 1–5. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Vijayaraghavan, S., Wang, Y., Guo, Z., Voong, J., Xu, W., Nasseri, A., Cai, J., Li, L., Vuong, K., and Wadhwa, E. (2020). Fake news detection with different models. *arXiv preprint arXiv:2003.04978.*

Vosoughi, S., Roy, D., and Aral, S. (2018). The spread of true and false news online. *Science*, 359(6380):1146–1151.

Wang, Q., Huang, L., Jiang, Z., Knight, K., Ji, H., Bansal, M., and Luan, Y. (2019). Paperrobot: Incremental draft generation of scientific ideas. *arXiv preprint arXiv:1905.07870*.

Wang, Q., Zeng, Q., Huang, L., Knight, K., Ji, H., and Rajani, N. (2020). Reviewrobot: Explainable paper review generation based on knowledge synthesis. *arXiv preprint arXiv:2010.06119*.

Wardle, C. and Derakhshan, H. (2017). Information disorder: Toward an interdisciplinary framework for research and policy making. *Council of Europe*, 27.

Williams, K. and Giles, C. L. (2015). On the use of similarity search to detect fake scientific papers. In *International Conference on Similarity Search and Applications*, pages 332–338. Springer.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October. Association for Computational Linguistics.

Wolff, M. and Wolff, S. (2020). Attacking neural text detectors. *arXiv preprint arXiv:2002.11768*.

Xiong, J. and Huang, T. (2009). An effective method to identify machine automatically generated paper. In *2009 Pacific-Asia Conference on Knowledge Engineering and Software Engineering*, pages 101–102. IEEE.

Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., and Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. volume 32.

Zellers, R., Holtzman, A., Rashkin, H., Bisk, Y., Farhadi, A., Roesner, F., and Choi, Y. (2019). Defending against neural fake news. *arXiv preprint arXiv:1905.12616*.