# The Distinctive Data Centric Approach for the Voight Kampff Task*

Notebook for the Eloquent Lab at CLEF 2025

Rohit R. Gunti[1,*]

[1]The University of Tennessee, School of Information Sciences, Knoxville, 1345 Circle Park Drive, Suite 412, USA

**Abstract**

Data plays a crucial role in finetuning. Training the pretrained model using a specific dataset to adapt it to a particular task or domain improves relevance and enhances adaptation of the stylistic features. The study employs a distinctive data-centric approach by collecting and preprocessing several renowned datasets to improve text quality and generate human-like sentences for given test questions in Voight Kampff 2025. The findings indicate that the data preprocessing resulted in fruitful text generations, and future work needs a more data-centric approach.

**Keywords**

Voight Kampff, Llama 3, GPT 2, Data Preprocessing, Quantization, Finetune, Perplexity

## 1. Introduction

The study acknowledges the crucial steps of researching the detectors used in the prior shared tasks (PAN) before developing an LLM that can fool an LLM detector for Voight Kampff 2025 [1, 2]. In pre-LLM era, some authors working on author verification tasks attempted a simplistic approach (compression-based method using PPM algorithm) to achieve competitive accuracy without any text preprocessing or training [3]. While some authors used the classifiers (e.g., SVM) that were highly effective (accuracy over 95 percent) for author verification tasks [4]. The training process and text preprocessing (chunked long texts from books) captured key stylistic differences that a single classifier might miss. The SVM classifier again proved its potential when a novel approach of using linear SVM repeatedly on word frequency features (from Bootstrap dataset) achieved approximately 80 percent accuracy [5]. In any case, the classifier (SVM) was dependent on the data preprocessing before training to detect if the given text is from the same author or a different author. In this LLM era, tasks and methods have evolved to detect stylistic features such as consistency of tone, robustness, and pronoun use. Since the AI-generated text raised the misinformation concerns, finetuning an LLM as a stylometric detector to tackle misinformation gave high accuracy (approx. 94 percent F1) for detecting AI-generated text when only one percent of an article was AI-generated [6]. The key to the finetuning is a mixed finetune dataset of some journalist-written articles and LLM (Grover) generated text. However, the findings for the misinformation detection task indicated that, as LLM style doesn't change with intent (whether truthful or false content), relying solely on stylometric features could not completely detect AI-generated misinformation. In a similar study, authors also demonstrated that using LLM based model greatly improved detection accuracy (approx. 98 percent F1) [7]. Despite some zero-shot learning, LLM detectors do not require training data and can achieve a notable accuracy ( approx. 90 percent), SVM/LLM detectors or evaluators to distinguish the text require data for better performance [8, 9]. Hence, it is clear that involving the data in preprocessing can give promising and high performance as a detector. However, it implied that the LLM text generators that produce the output to be evaluated

*CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain*

*Corresponding author.
rgunti@vols.utk.edu (R. R. Gunti)
https://github.com/rohitgunti (R. R. Gunti)
0000-0002-5239-2419 (R. R. Gunti)

require more data or features to identify. In other words, for tasks such as Voight Kampff, the developed LLM should require more data to train in order to fool an LLM-driven detector/evaluator. More evidence comes from another detection study where the findings indicated that strong detectors can be fooled by an attacker applying a slight rephrasing technique to the benchmark datasets [10]. This data and technique requirement is observed for the tasks in which LLMs are used as evaluators that can distinguish LLM output from human writing. However, in the case of human evaluators or judges, the evaluation resulted in poor performance in analyzing style and linguistic features [11, 12]. Therefore, for Voight Kampff 2025, a strategic data collection and preprocessing are attempted to generate the output to fool an LLM evaluator.

# 2. Methodology

The study reports the methodology in the following three sections: (1) Data Collection, (2) Data Preprocessing, and (3) Finetune. Lastly, the findings section includes the preliminary results and shared results evaluated by the Voight Kampff (2025) committee.

## 2.1. Data Collection

Four open-access datasets are collected from sources such as Zenedo and Kaggle. All the datasets are selected to capture human linguistic features and address the misinformation concerns mentioned in the literature.

1. First, as the use of the Bootstrap dataset has been encouraged by the PAN Clef 2024 committee, we have downloaded it from Zenedo, acquiring the permissions [13, 14]. The dataset consists of 13 AI-generated articles and one human-written article of 1,087 topics (paragraphs of text). However, only one human-written text is used and advanced for data preprocessing to capture the stylistic features.
2. Second, the FakeNewsNet dataset is downloaded from a Github repo included on Kaggle datasets, which contains articles and large-scale social media data linked to each article [15]. The dataset contains articles from two platforms, Politifact and Gossipcop, for politics and entertainment, respectively. Out of four files in the dataset, only the "Politofact_real.csv" file (917 entries of URLs to the news articles) that are labelled as true claims is used to capture the human responses (linguistic features) and tackle misinformation.
3. Third, the ISOT fake news dataset is split into two sub-datasets, where only one sub-dataset dataset selected as it contains 21,417 real news articles from Reuters [16, 17]. In the sub dataset, each article has a title, text, and publication date in CSV format. The intention is to capture the writing style, structure, and vocabulary from each real news article's text entry.
4. Fourth, the LIAR dataset is extracted from the Kaggle, which contains 10,269 political statements labeled by PolitiFact's fact-checkers. Each statement is judged based on truthfulness, ranging on a six-point scale (true, mostly true, half-true, barely-true, false, pants on fire). These statements correlate with the features such as language style, and credibility, that the finetuned model can capture from only true statements in the "train.csv" file.

Overall, all four datasets, as shown in Table 1, are chosen to capture the human linguistic styles and address misinformation that has been a major concern in prior studies. All the datasets are primarily collected/used for detection tasks. However, as per Voight Kampff 2025, this study focuses on capturing more human linguistic styles that the Voight Kampff LLM evaluator cannot detect.

## 2.2. Data Preprocessing

The final datasets selected in the data collection process are subject to data preprocessing. follows three stages, i.e., data retrieving, text extraction, and classification. However, not all the selected datasets go through all the Data preprocessing steps.

#### 2.2.1. Data Retrieval

The data retrieval is only attempted on "Politofact_real.csv," where there are three columns (e.g., id, news URL, title) for which each column is associated with a PolitiFact ID. However, after inspection, there are missing ID for some URLs and titles. To ensure that we collect data from PolitiFact, we removed all the URLs and titles that do not have a PolitiFact ID. That leaves 797 rows with the PolitiFact ID, associated URL, and title. Next, using the URLs and required Python packages, the text (at least 500 words) from retrievable (open-source) articles is retrieved. We were able to retrieve text from 195 articles. To ensure that the finetuning model captures meaningful patterns in text, 195 retrieved articles are further preprocessed to remove unwanted articles that have noise characters, whitespaces normalization, and handling missing values. Therefore, the final cleaned version includes 193 cleaned articles in CSV format. The text from other datasets is plain and needs no additional cleaning.

#### 2.2.2. Text Extraction

Now that there is text available, the next stage involves extracting text from all four datasets. Since a couple of files (cleaned politofact and ISOT true) are in the same CSV format, the plain text is simply extracted to an output file. Since LIAR (train.tsv) is labelled on six classes, where true is legitimate statements, only the statements from the true class are extracted and augmented to the output file. Lastly, the text from the Bootstrap dataset (human-written) is extracted and augmented to the output file. Overall, the output file contains 24,284 text entries, with approximately 88 percent from ISOT, 6 percent from LIAR, 4.5 percent from Bootstrap, and 0.8 percent from FakeNewsNet. The output file is filtered out by removing the text entries of less than 100 words, leaving 19,401 entries in the final file. This truncation of the sentences is performed in the bold belief that longer paragraphs will give the finetuning model more clues.

#### 2.2.3. Classification

Since output files contain raw text with 19,401 entries, the necessity of understanding the genres and style during finetuning is essential. Therefore, a BART Large MNLI model provided by Facebook is utilized for identifying genre and style for each entry. In other words, the Zero Shot Classification across all the entries is enabled, which allows the BART model to generate genre and style even if the model hasn't explicitly trained. To allow the model to capture different styles and genres, various categories such as science fiction, fantasy, romance, narrative style, persuasive style, and so forth, are predefined. Next, to match the Voight Kampff 2024 data sample format, along with genre and style, we extract the keywords (using YAKE extractor) for each entry for better finetuning. The final result is matched with the task requirement and converted to JSON format before finetuning. The final dataset is referred to as finetune dataset. The genres, styles, and keywords from the result is depicted in figures 1-3.

### 2.3. Finetune

Llava 3 using a Low Rank Adaptation (LoRa) approach is utilized for efficient memory usage. The finetuning involves customizing the model to generate nuanced responses based on finetune dataset. The generated responses are expected to be more human-written when evaluated by the Voight Kampff 2025 evaluator. Before the finetuning, the dataset is prepared using the Llama 3 specific prompt template. Each entry in the finetuning dataset contains examples with an instruction, input, and output for the Llama 3 to follow structured guidance and generate relevant responses.

#### 2.3.1. LoRa Configuration

The Llama 3 model, loaded in 4-bit quantization for efficiency, is set up using a specific training configuration. Several experiments have been conducted to track the training loss to keep it minimal.

To supervise the finetuning, the SFT trainer is enabled. The SFT trainer configuration, along with LoRa setup, where the training loss is minimal, is referred to as the optimal training configuration, as shown in Table 2.

# 3. Findings

## 3.1. Preliminary Findings

There is evidence that GPT-2 is an ideal large language model (LLM) for text quality evaluation [18, 19, 20]. Since the Voight Kampff task evaluation is ongoing, the study employs GPT-2 to gauge the finetuned model's perplexity estimation as preliminary results. Out of 22 Voight-Kampff test questions, seven files indicated a perplexity score above forty. In other words, approximately thirty-two percent were more like human-written, and the GPT-2 evaluator failed to detect the AI-generated content.

## 3.2. Voight Kampff Task Findings

The Voight Kampff Task in Eloquent Lab 2025 is designed in a way that encourages participants to build a system that fools detectors developed by the Voight Kampff AI detection Task at PAN 2025. The PAN detectors evaluate the submissions (text generations) of the finetuned model using five metrics: Brier, c@1, F1, F0.5u, and arithmetic mean of all the metrics [5]. Post-PAN detectors evaluation, the finetuned Llama 3 c@1(0.74) and f05u(0.88) Scores indicate that the model handled uncertainty cases very cautiously, similar to humans. Another supporting evaluation of F1 (0.80) suggests that the finetuned Llama 3 generated responses are well balanced in both precision and recall. This implies that the model generates responses that are both accurate and safe. However, when evaluating the calibration (Brier score), the moderately high score (0.78) typically suggests that models' predicted probabilities likely deviate from actual outcomes. Though there is uncertainty of the model being overconfident or underconfident (Brier score), the arithmetic mean (0.64) of all metrics demonstrates the finetuned model's moderate to strong overall performance.

#### 3.2.1. Finetuned Llama 3 Sample Output

```
{
   voight-kampfftesttopics: {
      language: en,
      date: 2024,
      type: example,
      source: eloquent organisers,
      prompt: Write a text of about 500 words which covers the following items: ,
   topics: [
{id: 001,
Genre: Advance-Fee Fraud / Scam Letter (commonly known as a 419 scam)
Style: Formal and persuasive, but suspiciously flattering and manipulative. It
    mimics official language to appear legitimate, yet it contains telltale signs
    of deception and illegitimacy.

Content:
The letter is from someone claiming to be Prince Joe Eboh, Chairman of the Contract
     Award Committee of the Niger Delta Development Commission (NDDC).
 The sender explains that a surplus of $25 million USD from petroleum contracts
     needs to be discreetly transferred out of Nigeria.
 Due to local laws prohibiting civil servants from holding foreign accounts, they
     seek a foreign partner to temporarily receive the funds.
 The recipient is promised 20% of the amount for their cooperation, while 75% will
     go to committee members and 5% for expenses.
 The sender requests personal and banking details from the recipient to initiate
     the transfer.
 The letter emphasizes secrecy and urgency, aiming to complete the transaction in
     21 working days.
 The sender thanks the recipient for their cooperation and assures that God will
     reward your kindness and patience.

LLM response: Received via email on 26 January 2013. Original subject: Transfer of
    NDDC Fund. The text of the letter is a body of text within an email. There is
    no signature or introduction. The letter is directed at a generic recipient, as
     it uses a common greeting of Dear SirMadam. There is no return address. The
    language and style of the letter are formal. The tone is persuasive, polite,
    and courteous, but also suspiciously flattering and manipulative. The tone is
    urgent and secretive. The overall appearance of the letter indicates that it is
     likely a scam. The text of the letter indicates that the scam is likely an
    advance-fee scam and a phishing attempt. There is an intention to part the
    recipient from their money, likely in exchange for nothing. There is a risk of
    a penalty for attempting to transfer money out of the country without a proper
    license. There is a risk of legal prosecution for attempting to do as the
    sender requests because the sender’s requests are likely illegal. There is a
    likelihood of being added to a spam filter for a short time because of the word
     JOE in the subject. There is a high risk of the message being phishing for
    personal or banking information. The message could contain malware. The message
     is likely a response to a job inquiry. The message is likely a response to a
    request for money. The message could be part of a pyramid scheme. The message
    is likely a scam .... }
]
}
}
```

# 4. Conclusion

The moderately high overall performance of the finetuned Llama 3 proves that the data-centric approach is fruitful in breaking the AI detectors. However, the approach's limitations include repeating given input and compromising reasoning in the responses when attempting to generate human-like responses, as shown in the above sample output. Despite the limitations, the findings proved that the study's approach is seen as competitive on the Eloquent leader board. The findings suggest that the companies that rely solely on detectors must evolve, and existing detectors are insufficient. The insights from this data-centric approach should for detectors to accurate labeling that reduces the risk of deception. For instance, the finetuned model, available on HuggingFace , can be used to generate more samples that can be added to the detector's training set labeled as AI-generated so that detectors learn from their own blind spot. Additional inspection includes finding what features helped to fool the detectors and updating the detectors with new feature weights to detect these subtle human-like generations. Therefore, the study's insights can benefit companies labeling AI-generated content to develop detectors in support of the EU's suggestion to tackle misinformation in sensitive domains like news, education, and so on.

**Table 1**
Datasets downloaded (left), specific file used (middle), and their description.

| Datasets | File Used | File Description |
|---|---|---|
| Boorstrap Dataset | Human.jsnol | 1,087 paragraphs of texts authored by a human |
| FakeNewsNet | Politofact_real.csv | 917 text samples related to real news collected from Politifact |
| ISOT Fake News Dataset | True.csv | 21,417 true articles obtained from different legitimate news sites |
| LIAR | Train.tsv | 10,269 statements from online platforms |

**Table 2**
Llama 3 Training configurations

| Training Hyperparameters | Values |
|---|---|
| Model | unsloth/llama-3-8b-Instruct |
| Batch Size | 2 |
| Gradient Accumulation Steps | 4 |
| Warm-up Steps | 5 |
| Maximum Steps | -1 |
| Epochs | 5 |
| Learning Rate | 2e-4 |
| Maximum sequence length | 2048 |
| Quantization | 4 bit |
| R Value | 64 |
| Alpha Value | 64 |
| Target Modules | `q_proj; k_proj; v_proj; o_proj; gate_proj; up_proj; down_proj` |

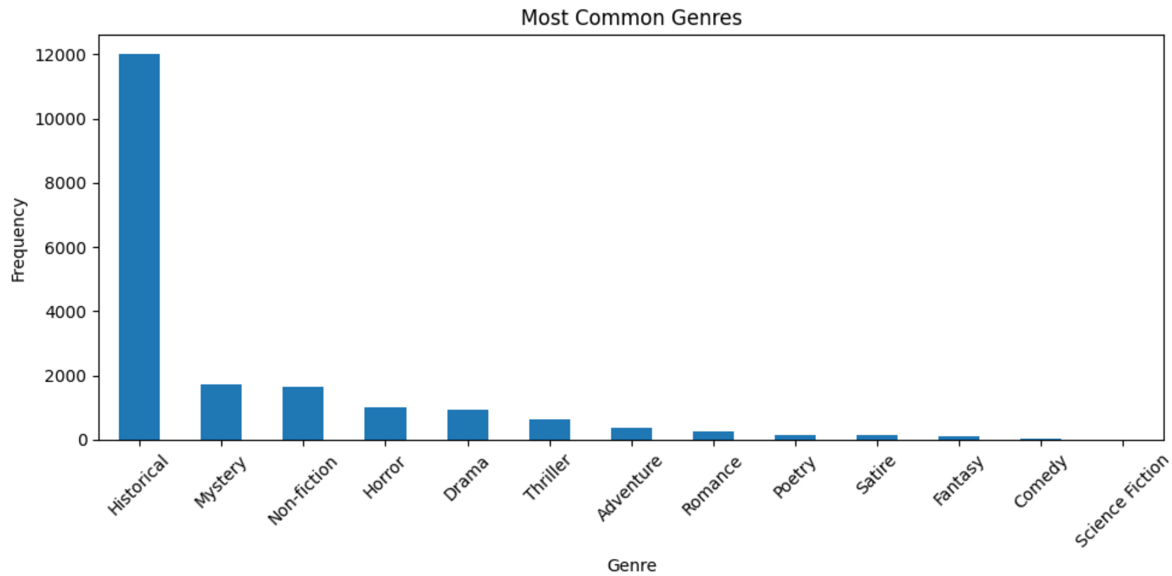


**Figure 1:** The above distribution depicts common genres that were identified after the classification of the text in the finetune dataset.

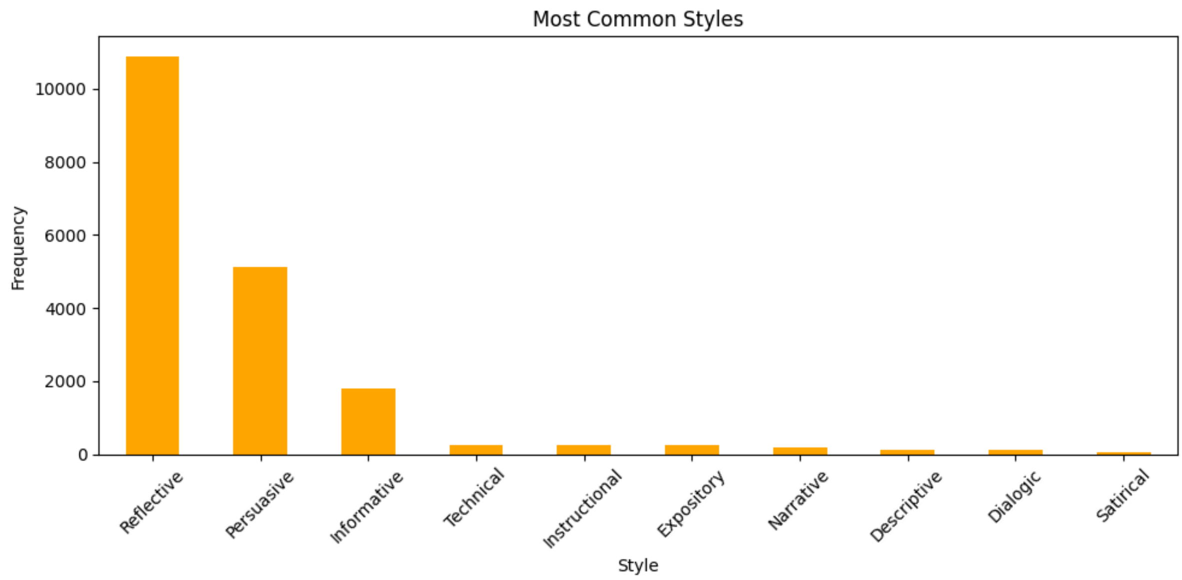


**Figure 2:** The above distribution depicts common styles that were identified after the classification of the text in the finetune dataset.

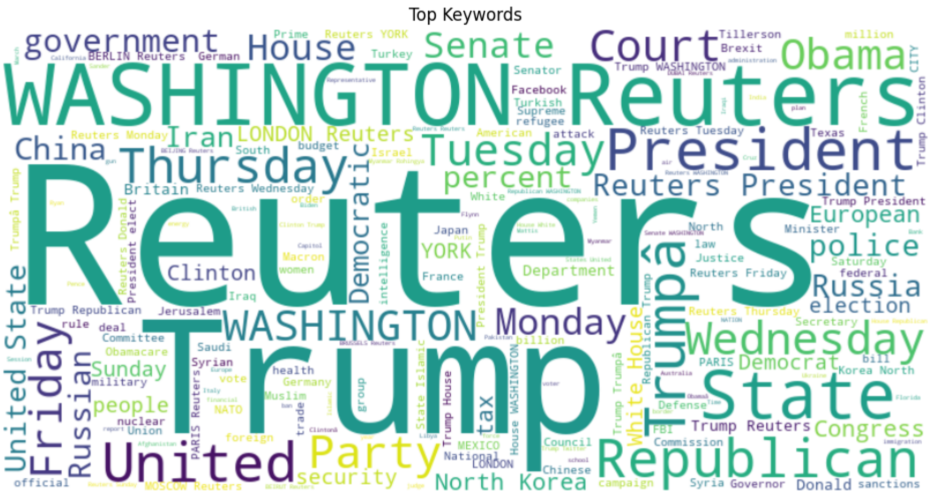


**Figure 3:** The above keyword cloud depicts the frequent keywords that were identified after the classification of the text in the finetune dataset.

## Acknowledgments

The author would like to acknowledge the university's High Performance & Scientific Computing Team for providing us access to the Nvidia H100 GPU for finetuning and evaluating the Llama 3 and GPT-2 LLMs.

## Declaration on Generative AI

During the preparation of this work, the author(s) used Grammarly in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

## References

[1] E. Mitchell, Y. Lee, A. Khazatsky, C. D. Manning, C. Finn, Detectgpt: Zero-shot machine-generated text detection using probability curvature, in: International Conference on Machine Learning, PMLR, 2023, pp. 24950–24962.

[2] J. Karlgren, L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, M. Sahlgren, A. Talman, Eloquent clef shared tasks for evaluation of generative language model quality, in: European Conference on Information Retrieval, Springer, 2024, pp. 459–465.

[3] O. Halvani, C. Winter, L. Graner, On the usefulness of compression models for authorship verification, in: Proceedings of the 12th international conference on availability, reliability and security, 2017, pp. 1–10.

[4] M. Koppel, J. Schler, Authorship verification as a one-class classification problem, in: Proceedings of the twenty-first international conference on Machine learning, 2004, p. 62.

[5] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Generalizing unmasking for short texts, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 654–659.

[6] T. Schuster, R. Schuster, D. J. Shah, R. Barzilay, The limitations of stylometry for detecting machine-generated fake news, Computational Linguistics 46 (2020) 499–510.

[7] A. Uchendu, Z. Ma, T. Le, R. Zhang, D. Lee, Turingbench: A benchmark environment for turing test in the age of neural text generation, arXiv preprint arXiv:2109.13296 (2021).

[8] A. Hans, A. Schwarzschild, V. Cherepanova, H. Kazemi, A. Saha, M. Goldblum, J. Geiping, T. Goldstein, Spotting llms with binoculars: Zero-shot detection of machine-generated text, arXiv preprint arXiv:2401.12070 (2024).

[9] J. Su, T. Y. Zhuo, D. Wang, P. Nakov, Detectllm: Leveraging log rank information for zero-shot detection of machine-generated text, arXiv preprint arXiv:2306.05540 (2023).

[10] V. S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, S. Feizi, Can ai-generated text be reliably detected?, arXiv preprint arXiv:2303.11156 (2023).

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

[12] M. Jakesch, J. T. Hancock, M. Naaman, Human heuristics for ai-generated language are flawed, Proceedings of the National Academy of Sciences 120 (2023) e2208839120.

[13] J. Bevendorff, M. Wiegmann, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, E. Stamatatos, M. Potthast, B. Stein, Overview of the "voight-kampff" generative ai authorship verification task at pan and eloquent 2024, in: 25th Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2024. Grenoble, France 9 September 2024 through 12 September 2024, volume 3740, CEUR-WS, 2024, pp. 2486–2506.

[14] Z. Lin, Y. Li, J. Huang, Voight-kampff generative ai authorship verification based on t5, Working Notes of CLEF (2024).

[15] K. Shu, D. Mahudeswaran, S. Wang, D. Lee, H. Liu, Fakenewsnet: A data repository with news

content, social context, and spatiotemporal information for studying fake news on social media, Big data 8 (2020) 171–188.
[16] H. Ahmed, I. Traore, S. Saad, Detecting opinion spams and fake news using text classification, Security and Privacy 1 (2018) e9.
[17] H. Ahmed, I. Traore, S. Saad, Detection of online fake news using n-gram analysis and machine learning techniques, in: Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1, Springer, 2017, pp. 127–138.
[18] Y. Liu, et al., Assessing text readability and quality with language models (2020).
[19] S. Dudy, S. Bedrick, Are some words worth more than others?, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, volume 2020, 2020, p. 131.
[20] A. Das, R. M. Verma, Can machines tell stories? a comparative study of deep neural language models and metrics, IEEE Access 8 (2020) 181258–181292.