

Article

Distinguishing Human Journalists from Artificial Storytellers Through Stylistic Fingerprints

Van Hieu Tran , Yakub Sebastian *, Asif Karim * and Sami Azam *

Faculty of Science and Technology, Charles Darwin University, Casuarina, NT 0810, Australia; vanhieu.tran@cdu.edu.au

* Correspondence: yakub.sebastian@cdu.edu.au (Y.S.); asif.karim@cdu.edu.au (A.K.); sami.azam@cdu.edu.au (S.A.)

Abstract: **Background:** Artificial intelligence poses a critical challenge to the authenticity of journalistic documents. **Objectives:** This research proposes a method to automatically identify AI-generated news articles based on various stylistic features. **Methods/Approach:** We used machine learning algorithms and trained five classifiers to distinguish journalistic news articles from their AI-generated counterparts based on various lexical, syntactic, and readability features. BERTopic was used to extract salient keywords from these articles, which were then used to prompt Google’s Gemini to generate new artificial articles on the same topic. **Results:** The Random Forest classifier performed the best on the task (accuracy = 98.3%, precision = 0.984, recall = 0.983, and F1-score = 0.983). Random Forest feature importance, Analysis of Variance (ANOVA), Mutual Information, and Recursive Feature Elimination revealed the top five important features: sentence length range, paragraph length coefficient of variation, verb ratio, sentence complex tags, and paragraph length range. **Conclusions:** This research introduces an innovative approach to prompt engineering using the BERTopic modelling technique and identifies key stylistic features to distinguish AI-generated content from human-generated content. Therefore, it contributes to the ongoing efforts to combat disinformation, enhancing the credibility of content in various industries, such as academic research, education, and journalism.



Citation: Tran, V.H.; Sebastian, Y.; Karim, A.; Azam, S. Distinguishing Human Journalists from Artificial Storytellers Through Stylistic Fingerprints. *Computers* **2024**, *13*, 328. <https://doi.org/10.3390/computers13120328>

Academic Editor: Emanuele Frontoni

Received: 2 November 2024

Revised: 28 November 2024

Accepted: 3 December 2024

Published: 5 December 2024



Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Background

The phenomenal advances in Artificial Intelligence (AI) in the past few years mean it can now answer questions, write essays, conduct academic research, and generate code [1]. Recently, there has been a growing concern that the proliferation of AI-generated news articles may harm individuals and institutions [2]. Generative AI technologies can create fictitious sources, threatening the credibility of many news publishers and journalists [3–5]. The trustworthiness of news citations has also been a major concern in the media and research [6–8]. AI chatbots have been used to generate fictional patient reviews or reinforce existing biases in information selection [9]. In journalistic contexts, the meteoric rise in generative AI and Large Language Models (LLMs) technologies has made it much easier to produce inauthentic news content [10,11]. In the medical and scientific writing context, Frosolini et al. [12] emphasised the importance of references as they are essential for readers to access valid and reliable sources and recommendations.

Before the rise in LLMs, fake news detection research primarily focused on analysing the differences between factual and fake news, where both were written by human agents. Research approaches have been proposed to detect fake news using knowledge-based (fact-checking), information retrieval-based, context-based, and style-based methods. The

knowledge-based technique involves using semantic web/LOD (Linked Open Data), which checks if claims are questionable, and whether articles can be inferred from or similar to the existing knowledge/facts [13–15]. The information retrieval-based methods check if suspicious texts contain factual statements from reliable websites [16–18]. The context-based techniques leverage social network analysis by analysing patterns and trends of rumours/misinformation on social media platforms [19–24]. Finally, the style-based techniques apply deception analysis [25–27] and text categorisation methods to uncover the differences in linguistic and writing style between fake news and factual news [28–30].

There are other novel approaches for fake news detection. Potthast et al. [31] proposed using the unmasking technique, a meta-learning method initially designed for authorship verification. To detect fake claims in the news, Popat et al. [32] looked at a set of lexical features: assertive verbs, active verbs, hedges, implicative, report verbs, discourse markers, subjectivity, and bias. Feature vectors were constructed for each article to compute the normalised frequency of all these linguistic features. For the same problem, Perez-Rosas et al. [33] applied source reliability and credibility classification using distant supervision to assess the credibility of a claim, whereas Agarwal et al. [34] used supervised machine learning algorithms to learn bags of words, N-grams, and Term Frequency-Inverse Document Frequency (TF-IDF) features to analyse news statements. Rubin et al. [30] proposed a method that considers satirical cues in news articles. They found that shallow syntax features (parts of speech and punctuation), absurdity (named entities analysis) feature, and negative affect features (capturing negative emotions) are significant in identifying fake news.

1.2. Related Work

Large Language Models (LLMs) enable the rapid creation and proliferation of artificial news content. Unlike fake news, artificial news contents do not necessarily carry disinformation or non-factual information, making them more challenging to detect using the aforementioned fake news detection techniques. Kumarage et al. [35] focused on stylometric features, such as phraseology, punctuation, and linguistic diversity, to detect the differences between human- and AI-authored tweets. Their experiments proved that stylometric features significantly improved the performance of an AI-generated text detector. However, linguistic diversity features (readability and lexical richness) were found to be the least important features, as these features require longer text sequences for better accuracy.

Herbold et al. [36] compared students' essays with ChatGPT-generated essays by assessing their lexical diversity, syntactic complexity, and semantic and discourse properties. Ma et al. [37] proposed a method for assessing perspectives like syntax, semantics, and pragmatics to distinguish AI-generated scientific abstracts and scientific abstracts written by human writers. They found that AI-generated abstracts were characterised by inconsistencies and a lack of scientific knowledge. Importantly, their syntactic features proved to be the strongest indicators of AI and human text.

Unlike previous research, this study examines the journalistic style characteristics between original news articles written by humans and their AI-generated counterparts. We collected human content from the journalism domain and AI content utilising the BERTopic modelling technique. Human-written news articles were written by famous and experienced journalists who are native English speakers with diverse, advanced writing skills. This allows us to explore various human writing styles and language structures. We also introduced novel methods to collect AI-generated news. BERTopic modelling was applied to the collected articles to generate latent topics and keywords. These keywords were subsequently used as part of the prompt for Google's Gemini to generate topically similar but artificial news content. This approach enables us to generate AI-generated news with content that closely resembles human-written news content, highlighting the distinctiveness in linguistic characteristics between journalists and AI. Gemini was chosen over ChatGPT because the former outperformed the latter in 30 out of 32 academic

benchmarks [38]. Gemini is proficient in generating text and mimicking human writing style [39].

1.3. Research Contributions

In this paper, we make two research contributions:

1. We proposed an innovative approach to prompt Artificial Intelligence (AI) chatbot generating news contents using the BERTopic modelling technique.
2. We identified salient linguistic features distinguishing AI-generated news from original, human-written articles.

The remainder of this paper is organised as follows. Section 2 explains the research methodology, including data collection, data cleansing, and study variables. Section 3 describes different predictive models used in this study, models evaluation metrics, and features selection techniques. The results of data pre-processing, features analysis, model evaluation, model optimisation, and feature selection are presented in Section 4, followed by a discussion of results in Section 5. Section 6 concludes this paper by outlining our research limitations and possible future work.

2. Methods

2.1. Original News Articles

Misleading news spreads rapidly, as it can easily bypass and be accepted by audiences when it mimics the writing style of famous journalists, claimed to be written by them. Therefore, we collected news articles from five award-winning journalists to analyse their linguistic characteristics and identify key indicators distinguishing their contents from false news generated by AI. They often have their work intensely scrutinised through rigorous editorial processes and fact-checking regimes, which significantly lessen the likelihood of misinformation or biased narratives. We gathered thirty articles from each journalist as benchmarks for understanding the attributes and characteristics of journalist-written news stories. In total, 150 articles were collected, and on average, there were 1664 words per article, 52 words per paragraph, and 23 words per sentence. The choice of using a certain number of articles collected from each journalist is driven by our emphasis on high-quality and trustworthy content. The limited dataset size allows us to carefully prioritise the richness of linguistic characteristics that are offered in professional content produced by the chosen journalists. We chose popular journalists on widely read topics such as politics, sports, military affairs, and technology, to gain diverse texts for analysis:

- **Thomas L. Friedman** (The New York Times)—foreign affairs, technology, and globalisation [40].
- **Konrad Marshall** (The Sydney Morning Herald)—Australian cultures and sports [41].
- **Dave Philipps** (The New York Times)—wars, military and veteran affairs [42].
- **Nicholas Kristof** (The New York Times)—human rights, health and global affairs [43].
- **David Swan** (The Australian)—technology [44].

2.2. AI-Generated News Articles

Subsequently, we prompted Gemini to compose articles that aligned with the content of the articles, which had been previously collected from the selected journalists. A total of 150 Artificial Intelligence (AI)-generated news articles were created by Gemini, matching the content of the 150 original articles collected before. We used topic modelling to ensure the relevance of our Gemini prompts. Several approaches were considered, such as Latent Dirichlet Allocation (LDA) [45], Non-negative Matrix Factorization (NMF) [46], Top2Vec [47], and BERTopic [48]. For this study, we chose BERTopic because of its ability to grasp context and represent words semantically, making it ideal for diverse and multilingual datasets. BERTopic is an approach to topic modelling that utilises embedding models, and a modified version of Term Frequency-Inverse Document Frequency (TF-IDF) called c-TF-IDF [48]. This technique helps generate compact topic clusters that are easy to under-

stand while ensuring that important words are retained in the topic descriptions. BERTopic can be used for various types of topic modelling, including guided, semi-supervised, hierarchical, and dynamic topic modelling [49]. By leveraging BERTopic modelling, this strategy enhances the precision and relevance of the generated content, aligning it closely with the themes and subject matter of the corresponding original articles. Table 1 shows an example of Gemini-generated news and Journalist-written news.

Table 1. Journalist-written news and Gemini-generated news using keywords generated by BERTopic.

Prompt: Assuming you are the famed journalist—Nicholas Kristof from The New York Times who has been awarded the Pulitzer Prize in 1990 and 2006; was the Pulitzer Prize Finalist in 1998. Write a piece of news following his writing style which includes the following keywords: ‘children’, ‘gaza’, ‘killed’, ‘palestinian’, ‘far’, ‘american’, ‘killing’, ‘war’, ‘sahloul’, ‘world’.

| AI-generated news | Real news |
|---|--|
| <p>Title: Gaza’s Hollow Lullaby: Where Children Bear the Scars of America’s Blind Eye</p> <p>Content: “From the war-torn alleys of Gaza, a heart-wrenching lullaby echoes—a haunting melody not of comfort, but of loss, echoing the silent screams of children stolen too soon. Here, amidst the rubble and despair, I, Nicholas Kristof, stand witness to a tragedy far graver than the headlines portray. It’s not just the buildings reduced to dust, nor the lives extinguished in the flames of war. It’s the collective heartbreak etched on the faces of children, the shattered innocence that exposes the hollowness of American inaction.</p> <p>Eleven-year-old Sahloul’s vacant eyes pierce my soul. A shrapnel wound mars his once playful expression, a constant reminder of the Israeli airstrike that ripped through his neighborhood, taking his two younger siblings. His voice, barely a whisper, narrates the horror: the deafening explosions, the blinding smoke, the chilling silence that followed, broken only by his own cries for his lost brother and sister. Sahloul is not alone. Dozens of children share his story, each wound a testament to the far-reaching consequences of a conflict fueled by America’s unwavering support for Israel’s military might. The justifications roll off the tongues of American officials with practiced ease: self-defense, Hamas aggression, the complexities of the Middle East. But these words offer little solace to a grieving mother whose child lies beneath the rubble, labeled “collateral damage” in the sterile language of war. They hold no meaning for a young boy who struggles to sleep, haunted by the nightmares of falling bombs and echoing screams.</p> <p>...”</p> | <p>Title: So Many Child Deaths in Gaza, and for What?</p> <p>Content: “Consider this: The most dangerous place to be a child in the world today is Gaza. That’s the assessment of Catherine Russell, the executive director of UNICEF, who is not a bleeding-heart radical but a former ambassador and veteran lawyer who worked for Presidents Biden and Barack Obama.</p> <p>Already it appears that more than twice as many children have died in Gaza just since the war started on Oct. 7 as in all the conflicts worldwide in 2022, according to United Nations figures.</p> <p>“Almost one out of every 150 Palestinian children in Gaza have been killed in just two months,” noted Dr. Zaher Sahloul, president of MedGlobal, an aid group working there. “That is the equivalent of half a million American children.”</p> <p>Dr. Sahloul warned that many others may “die from infections, waterborne diseases or dehydration,” while others will suffer from lifelong physical disabilities. We can and should despise Hamas, a repressive, misogynist and homophobic force that uses Palestinian civilians as human shields. And we can understand how Israel, traumatized by savage killings and rapes by Hamas, is determined to strike back. But just because Hamas is indifferent to the lives of Palestinian children does not mean that Israel or the United States should be reckless as well. The Biden administration has continued to periodically defend Israel not only when it is attacked, which is right, but even when it causes enormous numbers of Gazan civilian deaths.</p> <p>...”</p> |

2.3. Data Preprocessing

To prepare these articles for analysis, each individual article in its raw text form was processed in Python following the steps below.

2.3.1. Step 1: Contraction Expansion

Contractions are formed by combining two words into one, typically through an apostrophe. Contraction expansion ensures clarity and consistency in the text. Using Python’s NeatText library and a manually constructed list of contractions dictionary, a contraction such as “they’re” was expanded to “they are”, “he’s” to “he is”, and “she’d” to either “she had” or “she would” depending on its context. This expansion process not only aids in maintaining grammatical correctness but also facilitates subsequent natural language processing tasks, allowing for a more comprehensive and accurate analysis of the text. Contractions also improve stopwords removals.

2.3.2. Step 2: Stopwords Removal

A standard stopwords removal procedure was applied to further streamline the text data and make it more amenable to the application of BERTopic for generating key topics (or keywords) from each article. Eliminating stopwords allows BERTopic to generate more relevant and reflective topics and keywords from original articles. We preserved punctuation marks, commas, question marks, and any other special character that signifies the end of a sentence as they hold a crucial function in the next stage of the analysis.

2.3.3. Step 3: Sentence Tokenisation and Special Characters Removal

We used the Natural Language Toolkit (NLTK) package [50,51] to divide the whole text, including several paragraphs, into sentences. As mentioned earlier, the punctuation marks, commas, question marks, and any other special character that signifies the end of a sentence will need to be kept before applying the tokeniser package. This is because the tokeniser relies on the presence of end-of-sentence marks to determine the appropriate places to separate sentences from the text. BERTopic algorithm works more effectively when tokenising a long document containing many sentences into multiple documents where each document contains only one sentence [52].

2.3.4. Step 4: Whitespace, End-of-Sentence Marks, and Uninformative Sentences Removal

We standardised all the text to lowercase; for example: “The dataset” will become “the dataset” [53]. Special characters that mark the end of sentences were finally removed as they do not carry any meaning to the content of the text or will not contribute to the later stage of applying topic modelling and data analysis. Removing stopwords and special characters resulted in multiple extra white spaces between the remaining text (Figure 1). These extra whitespaces were also removed. Finally, documents containing a limited amount of text (sentences with less than three words) were removed as they do not carry significant meaning for BERTopic to generate meaningful topics or keywords representing the content of the original article.

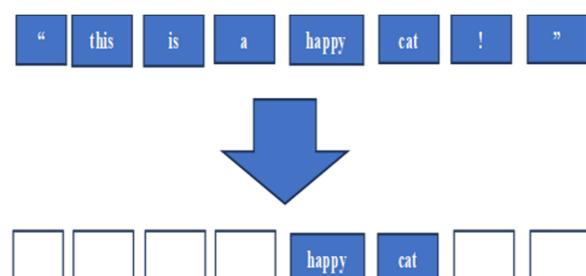


Figure 1. Example of extra white spaces after removing stop words and special characters.

2.4. Features

2.4.1. Response Variable

The response variable contains two labels: Journalist and Gemini. The former indicates that the news was written by a journalist (human) while the latter indicates the news was generated by Artificial Intelligence (AI).

Our predictive features comprise lexical, syntactic and readability features.

2.4.2. Lexical Features

By analysing the Part of Speech (PoS) tags within journalists' news and AI-generated news, we explore differences in the distributions of adverb, adjective, verb, noun and verb, noun-to-verb, adjective-to-noun, adverb-to-verb within the text. PoS tags reveal the grammatical functions of words in sentences, including determiners, pronouns, prepositions, conjunctions, auxiliaries and modals, nouns, verbs, adjectives and adverbs [54]. For instance, Lex et al. [55] applied PoS analysis to classify emotion levels in news-related blogs

and found the number of adverbs to be the most relevant feature. Individuals who use many adverbs tend to write very emotionally. Elsewhere, adjectives and adverbs have been useful for classifying different genres of documents (academic, news, and fiction) and for detecting subjectivity versus objectivity [56]. There are many other studies that reported useful features involving verbs and nouns [57–59]. We also observed differences in the text length between human and AI-generated news (Figure 2). Features, such as the type-token ratio, were included to measure the lexical diversity of text irrespective of their lengths [36,60].

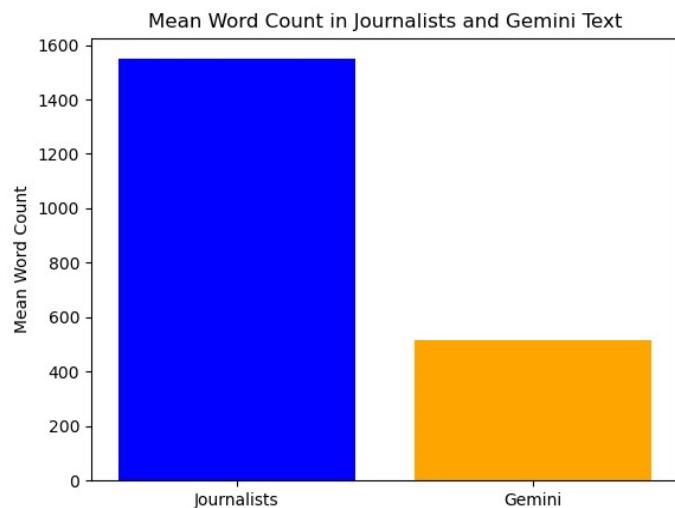


Figure 2. Average word count in journalists and Gemini texts.

We describe the formulas of the lexical-related features below:

$$\text{adverb_ratio} = \frac{\text{Number of Adverbs}}{\text{Total Number of Words}} \times 100 \quad (1)$$

$$\text{adjective_ratio} = \frac{\text{Number of Adjectives}}{\text{Total Number of Words}} \times 100 \quad (2)$$

$$\text{verb_ratio} = \frac{\text{Number of Verbs}}{\text{Total Number of Words}} \times 100 \quad (3)$$

$$\text{noun_and_verb_ratio} = \frac{\text{Total Number of Verbs} + \text{Total Number of Nouns}}{\text{Total Number of Words}} \times 100 \quad (4)$$

$$\text{noun_to_verb_ratio} = \frac{\text{Number of Nouns}}{\text{Number of Verbs}} \times 100 \quad (5)$$

$$\text{adjective_to_noun_ratio} = \frac{\text{Number of Adjectives}}{\text{Number of Nouns}} \times 100 \quad (6)$$

$$\text{adverb_to_verb_ratio} = \frac{\text{Number of Adverbs}}{\text{Number of Verbs}} \times 100 \quad (7)$$

$$\text{TTR} = \frac{\text{Unique words}}{\text{Total number of words}} \times 100 \quad (8)$$

2.4.3. Syntactic Features

We also analysed various aspects related to sentences and paragraphs' structure [36,61,62]. Through these features, we gain valuable insights into the diversity of sentence and paragraph lengths as well as the variability of sentence/paragraph lengths compared to their mean length between journalists' news and Gemini-generated news. We computed various sentence- and paragraph-level syntactic features as follows:

- *Average sentence length:* the mean value of the number of words in each sentence.

- *Sentence length range*: the number of words in the longest sentence—the number of words in the shortest sentence
- *Sentence length coefficient of variation*: sentence lengths divided by the mean value of sentence lengths.
- *Average paragraph length*: the mean value of the number of words in each paragraph in the text.
- *Paragraph length range*: the number of words in the longest paragraph minus the number of words in the shortest paragraph.
- *Paragraph length coefficient of variation*: paragraph lengths divided by the mean value of paragraph lengths.
- *Sentence complex tag and sentence complex depth* were calculated to evaluate the differences between journalists' news and Gemini-generated news in terms of their syntactic complexity [36,63]. The sentence complex depth feature was computed based on the maximum depth of the dependency parse tree, which reflects the number of relationship layers that exist between words in a sentence. In addition, the sentence complex tags feature was generated by calculating the average dependency score of sentences, which indicates how often the selected grammatical relationship shows up between given words. Sentence complex tags and sentence complex depth features engineering tasks were conducted using SpaCy dependency parser in Python.

2.4.4. Readability Features

We also measured whether an audience finds it easy or difficult to read news articles from journalists and AI-generated news articles [64]. We used the Flesch Reading Ease score as a readability measure [65]. The closer the score to 100, the easier the audience can read the text [66]. Additionally, we computed the Coleman–Liau readability formula [67]. In contrast to the Flesch Reading Ease score, which is measured based on total words, sentences and syllables, the Coleman–Liau score was measured based on the alphabetical letters; the equations are given below [66,67].

$$\text{Coleman – Liau score} = 5.88 \times AWL + 29.6 \times \frac{n_{st}}{n_w} - 15.8 \quad (9)$$

$$\begin{aligned} \text{Flesch Reading Ease score} = \\ 206.835 - \left(1.015 \times \frac{\text{total words}}{\text{total sentences}} \right) - \left(84.6 \times \frac{\text{total syllables}}{\text{total words}} \right) \end{aligned} \quad (10)$$

3. Modelling

3.1. Classifiers

Five classifiers were compared: Naïve Bayes, Passive Aggressive, Random Forest, Support Vector Machine, and Gradient Boosting Machine. These models have been explored in recent studies [34,68–70]. Naïve Bayes was used as the performance baseline for the more advanced classifiers. We trained KNN and SVM models with feature standardisation because they are known to be sensitive to feature scaling [71].

3.2. Cross-Validation

We split the dataset into 80% for the training set and 20% for the testing set. K-fold cross-validation ($k = 5$) was applied on the training set, where the average performance calculated for each fold will be the final model performance [72].

3.3. Predictive Model Evaluation

This study aimed to detect journalist-written and Gemini-generated news; therefore, this is a binary classification problem, and how well the predictive model separates the

classes can be evaluated using a confusion matrix (Table 2) [73]. The formulas for calculating accuracy, precision, recall and F1-score are given below.

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \quad (11)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad (12)$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (13)$$

$$\text{F1} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (14)$$

Table 2. Confusion matrix for binary classification.

| | Actual Positive Class | Actual Negative Class |
|--------------------------|-----------------------|-----------------------|
| Predicted Positive Class | True Positive (TP) | False Negative (FN) |
| Predicted Negative Class | False Positive (FP) | True Negative (TN) |

3.4. Feature Selection

Feature selection is one of the most crucial processes in the machine learning field that needs to be conducted to solve common challenges, such as redundant or irrelevant features, which might negatively affect the model's ability to learn data and decrease its performance [74]. In this study, Pearson's correlation heatmap was generated to examine redundant features, whereas the Random Forest feature importance, Recursive Feature Elimination (RFE), Analysis of Variance (ANOVA), and Mutual Information methods were applied to select the best set of features for the predictive model.

4. Results

4.1. Text Pre-Processing

After completing the text-cleaning process, the NeatText library was used to evaluate the text-cleaning process. The Describe function within the NeatText library was applied in this step to examine how clean the text data were, or if there was any noise in the datasets. The term “noise” within the text data can be explained as stop words, punctuation, special characters, and extra whitespace that we tried to remove from previous steps. Table 3 shows that when applying the Describe method to Marshall’s article. The output includes the following parameters:

- Length: indicates the total length of the text including the total of characters and spaces.
- Vowels: indicates the number of vowels that appear within the text. The vowels include the letters: “a”, “e”, “i”, “o”, “u”, and “y”.
- Consonants: indicates the number of consonants that appear within the text. The Consonants include all the letters except letters: “a”, “e”, “i”, “o”, “u”, and “y”.
- Stop words indicate the number of stop words within the text.
- Punctuation: indicates the number of punctuations within the text.
- Special characters: indicates the number of symbols and characters that are not letters or numbers.
- Tokens (whitespace): indicates the number of tokens separated by whitespace in the text.
- Tokens (words): indicates the number of tokens. The tokens here represent the number of words in the text.

Table 3. Noise evaluation for the raw text of the first article collected from Konrad Marshall.

| Marshall (Article One) | |
|------------------------|-------|
| Key | Value |
| Length | 7621 |
| Vowels | 2270 |
| Consonants | 3760 |
| Stopwords | 685 |
| Punctuations | 173 |
| Special_characters | 173 |
| Tokens (Whitespace) | 1357 |
| Tokens (words) | 1352 |

After the cleaning process had been completed, including the removal of whitespace, stop words, special characters, and punctuations as well as the tokenised raw text in sentences, we again, applied the NeatText and Describe method to check the noise in the text data. From the output (Table 4), we can see that text length, number of vowels, and consonants are reduced as a result of the removal of unnecessary words, characters, and tokenisation. Each of the rows (row 1, row 2, row 3, row 4, ..., row 41) indicates that the output of each sentence; there are 41 rows, implying that 41 sentences have been tokenised from the original text. The number of punctuations and special characters equal to 0 implies that we have successfully removed them from the original text. The number of tokens (whitespace) is equal to the number of tokens (words), indicating that there is no extra whitespace. From the output we obtained, we can be certain that there is no noise within the text data, which means that the text datasets have been cleaned and are usable in the topic engineering stage and further data analysis tasks.

Table 4. Noise evaluation for the processed text of the first Konrad Marshall's article.

Row 1 (original): The first time I spoke at length with Damien Hardwick about coaching mortality, he sat at the big oval-shaped desk inside the Tigerland war room, only months after surviving an ill-conceived coup.
Row 1 (processed): time spoke length damien hardwick coaching mortality sat big ovalshaped desk inside tigerland war room months surviving illconceived coup

Row 3 (original): There was precious little to indicate that one day he himself would be the trigger man.

Row 3 (processed): precious little indicate day trigger man

Row 41 (original): Maybe he just wants to go for a walk.

Row 41 (processed): maybe wants walk

| Key | Value | Key | Value | Key | Value |
|---------------------|-------|---------------------|-------|---------------------|-------|
| Length | 137 | Length | 40 | Length | 16 |
| Vowels | 43 | Vowels | 14 | Vowels | 4 |
| Consonants | 76 | Consonants | 21 | Consonants | 10 |
| Stopwords | 0 | Stopwords | 0 | Stopwords | 0 |
| Punctuations | 0 | Punctuations | 0 | Punctuations | 0 |
| Special_char | 0 | Special_char | 0 | Special_char | 0 |
| Tokens (Whitespace) | 19 | Tokens (Whitespace) | 6 | Tokens (Whitespace) | 6 |
| Tokens (words) | 19 | Tokens (words) | 6 | Tokens (words) | 6 |

4.2. Feature Analysis

After pre-processing text data and engineering the selected set of features, the mean value of each feature was generated to compare the journalist-written news and Gemini-generated news in terms of lexical, syntactic, and audience comprehension (readability features) (Figure 3). Journalist articles had the mean value of sentence length coefficient of variation (56.95), sentence length range (57.55), average sentence length (22.83), paragraph length coefficient of variation (43.76), paragraph length range (91.24), all of which

were slightly higher than Gemini-generated articles (41.07, 31.45, 19.29, 27.05, and 46.00, respectively). Therefore, these features are significant to the predictive model performance. Whereas, there are small differences between Gemini-generated articles and Journalist articles in terms of adverb ratio (0.04–0.05), verb ratio (0.13–0.16), noun-to-verb ratio (2.32–1.72), adjective-to-noun ratio (0.32–0.37), adverb-to-verb ratio (0.28–0.30), type token ratio (0.53–0.45), sentence complexity tags (1.6–2.19), sentence complexity depth (5.97–6.01), Flesch Reading Ease (59.72–58.99), and the Coleman–Liau Index (11.31–10.68), when examining their mean value. Hence, these features might still have impacts on the model performance, but not much. Otherwise, the adjective ratio might have no impact on the model performance since there is no difference between Journalist articles and Gemini-generated articles in terms of the mean value of the adjective ratio (0.10–0.10).

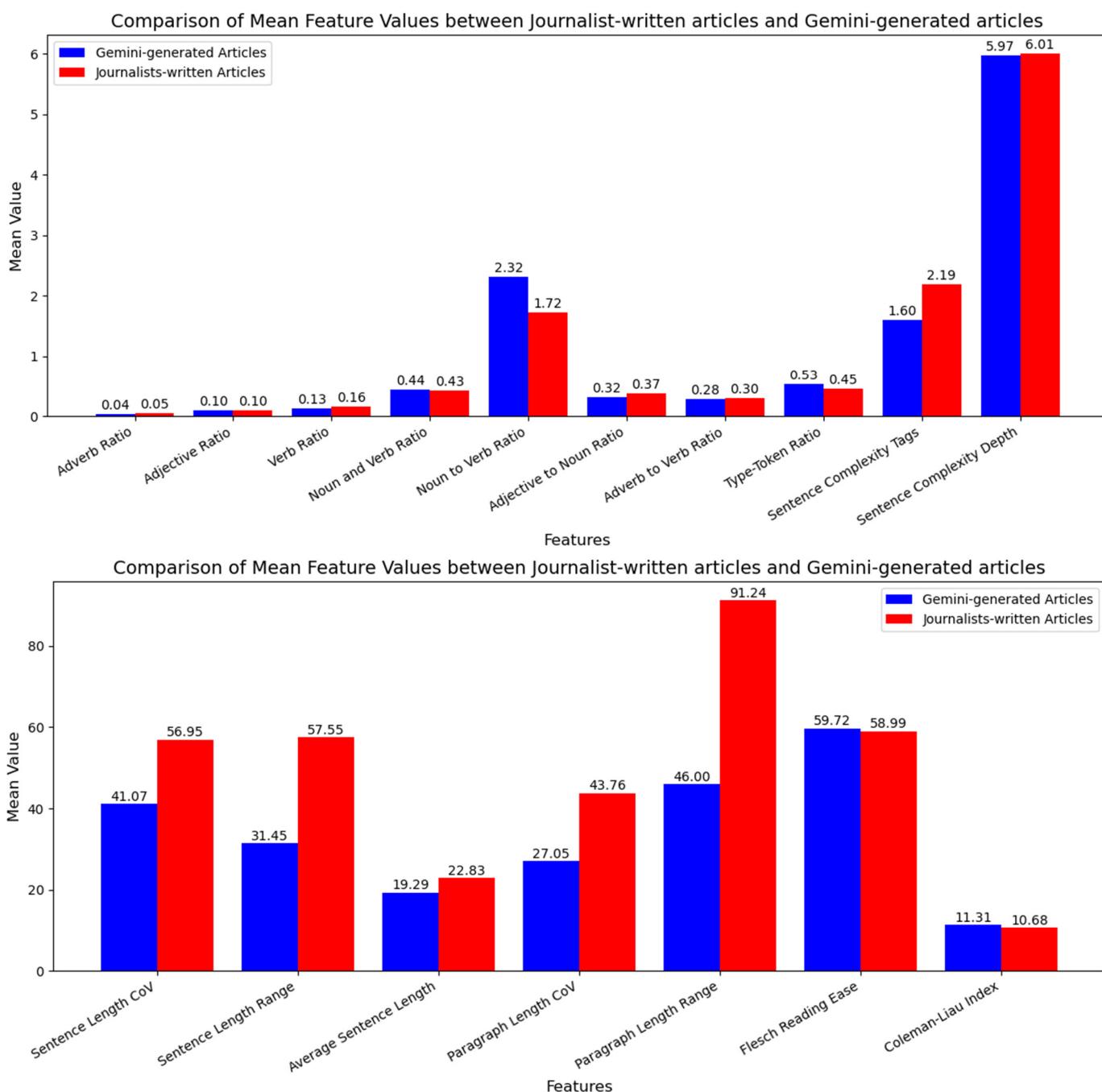


Figure 3. Mean value comparison of all features between Journalist-written articles and Gemini-generated articles.

Overall, the graph can be interpreted as indicating that most of the syntactic features are more important than most of the lexical features and readability features in distinguishing the writing style of Journalist(s) and Gemini.

4.3. Model Evaluation

After building five classification models with all seventeen features, they returned with different performances (Figure 4). The Passive Aggressive classifier has an accuracy score of 0.833, a precision score of 0.872, a recall score of 0.839 and an F1-score of 0.830. The Naïve Bayes classifier has an accuracy score of 0.933, a precision score of 0.936, a recall score of 0.932 and F1-score of 0.933. The Support Vector Machine classifier has an accuracy score of 0.967, a precision score of 0.967, a recall score of 0.967 and an F1-score of 0.967. The Gradient Boosting Machine classifier has an accuracy score of 0.967, a precision score of 0.969, a recall score of 0.967, and an F1-score of 0.967. The Random Forest classifier has an accuracy score of 0.983, a precision score of 0.984, a recall score of 0.983, and an F1-score of 0.983. Therefore, in terms of evaluation metrics (accuracy, precision, recall, and F1), Random Forest has outperformed other classifiers.

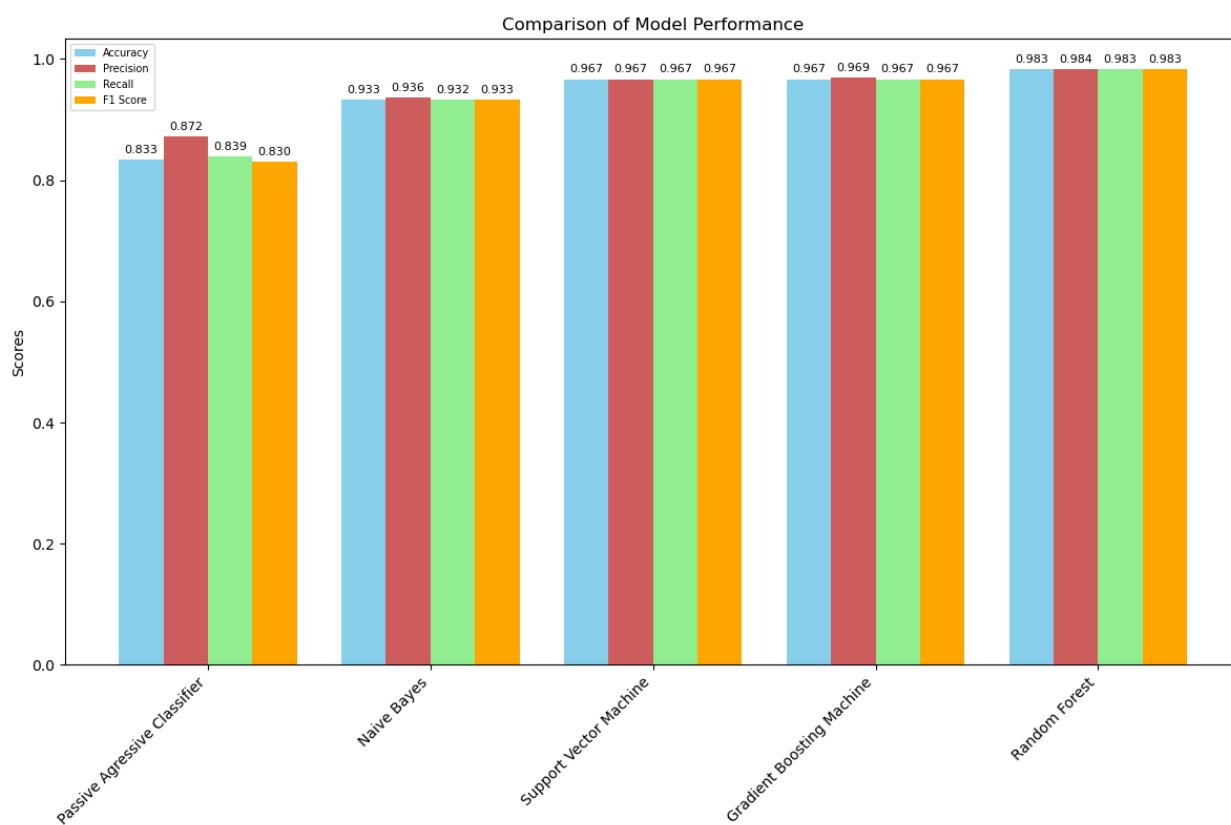


Figure 4. Comparison of models' performances.

As explained above, the Random Forest classifier was selected since it achieved the best performance compared to other classifiers. In the next experiment, model optimisation and feature selection methods were applied to determine the most important features that contributed to the model performance in predicting the originality of news (Journalist/Gemini).

4.4. Model Optimisation

The Random Search method was applied to optimise the Random Forest model. The Random Forest model before applying the optimisation method (with default hyperparameter tuning) had an accuracy score of 0.983, a precision score of 0.984, a recall score of 0.983, and an F1-score of 0.983, whereas, the Random Forest model, after applying optimisation

method (with hyperparameter tuned), had a score of 0.967 as accuracy, precision, recall, and F1-score. Therefore, the Random Forest model with default hyperparameter tuning was selected as it performed better.

4.5. Feature Selection Results

In this experiment, Pearson's correlation score was generated to examine on redundant features, whereas the Random Forest feature importance, Recursive Feature Elimination (RFE), Analysis of Variance (ANOVA) and Mutual Information methods were applied to select the best set of features for the predictive model. Figure 5 shows several redundant or highly correlated features such as verb ratio and noun-to-verb ratio, adjective ratio and adjective-to-noun ratio, adverb-to-verb ratio, and adverb ratio, Coleman–Liau index, and Flesch Reading Ease score. One of the features in these pairs needs to be removed as having redundant features will increase the dimensions of the dataset, increase the training time and increase the chance of overfitting [74].

| Noun to Verb Ratio | 1 | -0.27 | -0.07 | 0.48 | -0.4 | -0.49 | -0.46 | -0.41 | 0.34 | -0.31 | 0.25 | -0.04 | -0.51 | 0.15 | -0.93 | 0.16 | -0.15 |
|------------------------------|--------------------|-------------------------|----------------------|------------------|----------------------|------------------------|---------------------|-----------------------|--------------------|------------------------------|------------------------|-------------------------|--------------|-----------------|------------|---------------------|---------------------------|
| Adjective to Noun Ratio | -0.27 | 1 | 0.13 | -0.21 | 0.29 | 0.26 | 0.15 | 0.27 | 0.26 | 0.31 | 0.093 | 0.3 | 0.18 | 0.86 | 0.094 | -0.49 | -0.32 |
| Adverb to Verb Ratio | -0.07 | 0.13 | 1 | -0.12 | 0.068 | 0.21 | 0.32 | 0.26 | -0.21 | 0.012 | -0.19 | -0.007 | 0.87 | -0.01 | -0.071 | -0.4 | 0.15 |
| Tyep Token Ratio | 0.48 | -0.21 | -0.12 | 1 | -0.45 | -0.64 | -0.4 | -0.57 | 0.21 | -0.32 | 0.039 | -0.23 | -0.3 | 0.11 | -0.4 | 0.36 | 0.043 |
| Paragraph Length CoV | -0.4 | 0.29 | 0.068 | -0.45 | 1 | 0.74 | 0.28 | 0.37 | 0.12 | 0.29 | -0.056 | 0.2 | 0.23 | 0.12 | 0.37 | -0.098 | -0.23 |
| Paragraph Length Range | -0.49 | 0.26 | 0.21 | -0.64 | 0.74 | 1 | 0.51 | 0.58 | -0.2 | 0.25 | -0.17 | 0.12 | 0.4 | -0.018 | 0.43 | -0.26 | 0.041 |
| Sentence Length CoV | -0.46 | 0.15 | 0.32 | -0.4 | 0.28 | 0.51 | 1 | 0.78 | -0.48 | 0.12 | -0.42 | -0.082 | 0.51 | -0.099 | 0.43 | -0.18 | 0.39 |
| Sentence Length Range | -0.41 | 0.27 | 0.26 | -0.57 | 0.37 | 0.58 | 0.78 | 1 | -0.22 | 0.44 | -0.028 | 0.34 | 0.41 | 0.037 | 0.33 | -0.24 | 0.053 |
| Coleman-Liau Index | 0.34 | 0.26 | -0.21 | 0.21 | 0.12 | -0.2 | -0.48 | -0.22 | 1 | 0.24 | 0.44 | 0.35 | -0.36 | 0.51 | -0.33 | 0.19 | -0.88 |
| Sentence Length Complex Tags | -0.31 | 0.31 | 0.012 | -0.32 | 0.29 | 0.25 | 0.12 | 0.44 | 0.24 | 1 | 0.55 | 0.81 | 0.15 | 0.19 | 0.27 | -0.062 | -0.46 |
| Sentence Complex Depth | 0.25 | 0.093 | -0.19 | 0.039 | -0.056 | -0.17 | -0.42 | -0.028 | 0.44 | 0.55 | 1 | 0.76 | -0.31 | 0.19 | -0.28 | -0.033 | -0.55 |
| Average Sentence Length | -0.04 | 0.3 | -0.007 | -0.23 | 0.2 | 0.12 | -0.082 | 0.34 | 0.35 | 0.81 | 0.76 | 1 | -0.023 | 0.27 | -0.038 | -0.13 | -0.54 |
| Adverb Ratio | -0.51 | 0.18 | 0.87 | -0.3 | 0.23 | 0.4 | 0.51 | 0.41 | -0.36 | 0.15 | -0.31 | -0.023 | 1 | -0.12 | 0.42 | -0.3 | 0.23 |
| Adjective Ratio | 0.15 | 0.86 | -0.01 | 0.11 | 0.12 | -0.018 | -0.099 | 0.037 | 0.51 | 0.19 | 0.19 | 0.27 | -0.12 | 1 | -0.24 | -0.15 | -0.45 |
| Verb Ratio | -0.93 | 0.094 | -0.071 | -0.4 | 0.37 | 0.43 | 0.43 | 0.33 | -0.33 | 0.27 | -0.28 | -0.038 | 0.42 | -0.24 | 1 | 0.15 | 0.18 |
| Noun and Verb Ratio | 0.16 | -0.49 | -0.4 | 0.36 | -0.098 | -0.26 | -0.18 | -0.24 | 0.19 | -0.062 | -0.033 | -0.13 | -0.3 | -0.15 | 0.15 | 1 | -0.022 |
| Flesch Reading Ease Score | -0.15 | -0.32 | 0.15 | 0.043 | -0.23 | 0.041 | 0.39 | 0.053 | -0.88 | -0.46 | -0.55 | -0.54 | 0.23 | -0.45 | 0.18 | -0.022 | 1 |
| | Noun to Verb Ratio | Adjective to Noun Ratio | Adverb to Verb Ratio | Tyep Token Ratio | Paragraph Length CoV | Paragraph Length Range | Sentence Length CoV | Sentence Length Range | Coleman-Liau Index | Sentence Length Complex Tags | Sentence Complex Depth | Average Sentence Length | Adverb Ratio | Adjective Ratio | Verb Ratio | Noun and Verb Ratio | Flesch Reading Ease Score |

Figure 5. Pearson's features' correlation score.

Figure 6 compares Random Forest classifier performance with different sets of features selected by Random Forest feature importance, Recursive Feature Elimination (RFE), ANNOVA and Mutual Information method. The Random Forest classifier achieved the highest performance with the selected features from Random Forest feature importance

and Recursive Feature Elimination. However, there is no difference between these two feature selection methods in terms of accuracy, precision, recall, and F1-score.

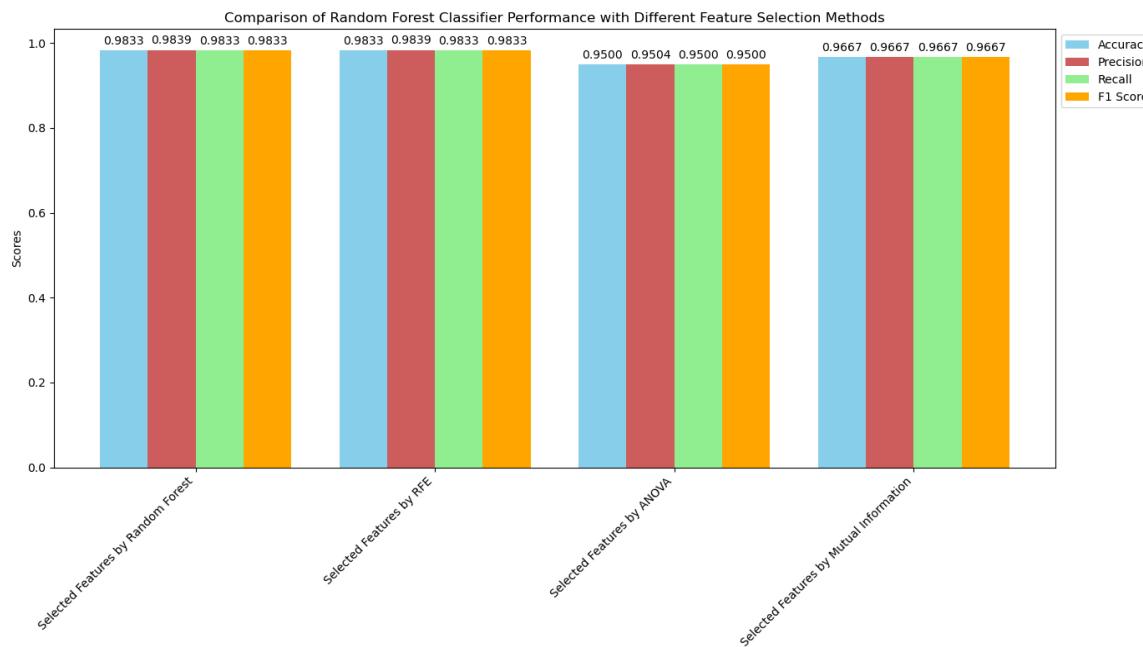


Figure 6. Comparison of Random Forest classifier performance with different feature selection methods.

On the other hand, the Random Forest feature importance method ranked all initial sets of seventeen features as important (Figure 7). However, the set of features that were given by Recursive Feature Elimination was reasonable to obtain, because, with less number of features (five features), the Random Forest model still achieved the same result (Figure 8) compared to seventeen features that were given by Random Forest feature importance method and a smaller number of features will reduce the dimension of the dataset, reduce training time, and avoid overfitting. Overall, sentence length range, paragraph length coefficient of variation, verb ratio, sentence complex tags, and paragraph length range are the most significant features for news originality prediction (journalist/Gemini).

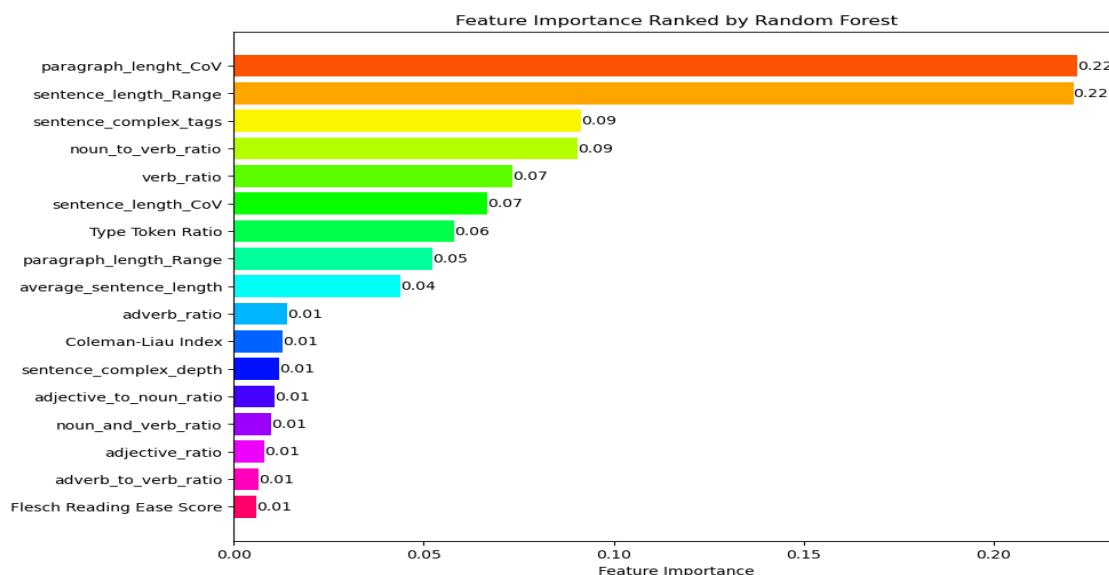


Figure 7. Selected features by the Random Forest feature importance method.

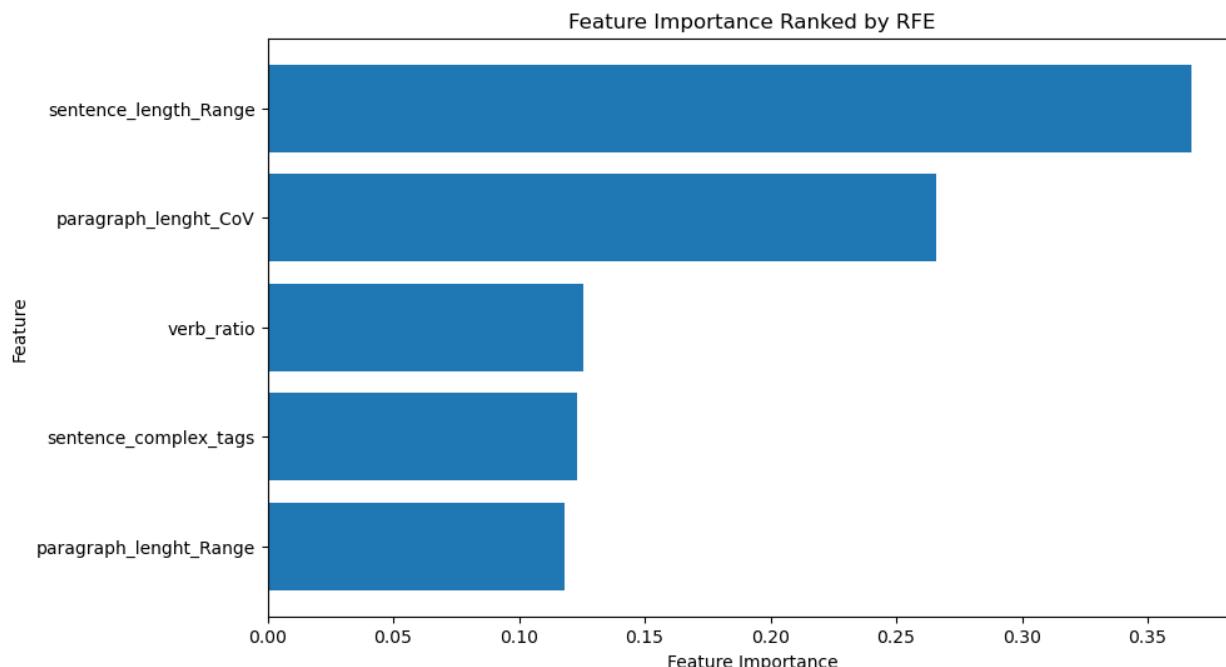


Figure 8. Selected features by the Recursive Feature Elimination method (RFE).

5. Discussions

The research focuses on finding linguistic differences between journalist-written and Gemini-generated news articles, which contribute to detecting AI-generated content. As AI models become increasingly sophisticated, we expect that finding distinguishing features between human-generated and AI-generated text will become increasingly difficult, especially for human readers. Machine learning approaches hold the promise of finding less distinct differences between the texts in ways that humans cannot. Those subtle but significant differences between human- and AI-generated texts may only become more apparent when multiple aspects of stylistic features are considered together: lexical, syntactic and readability features. This is a key proposition of this paper.

Our results suggest that Random Forest performed better than other classifiers with an accuracy score of 0.983, a precision score of 0.984, a recall score of 0.983 and an F1-score of 0.983. This contrasts with the study of Shabani et al. [69], which showed that the Passive Aggressive classifier achieved high performance in human-written real/fake news classification tasks and the Support Vector Machine achieved the best performance with the same task in research [34]. Whereas, with the same objective of authorship verification, Canbay et al. [75] found that the Logistic Regression classifier outperformed the Support Vector Machine classifier.

Other differences that can be considered are the feature extraction methods, as the model's performance could be impacted by the predictors [72]. In their study, linguistic features were extracted using a bag of words as well as Term-Frequency Inverse Document-Frequency (TF-IDF) and Word2Vec, which mainly focus on the frequency of words and the vocabulary of the known words. We extended the feature extraction further by not only learning the distribution and the relationship between words but also learning how sentences and paragraphs are structured, thus measuring the readability scores. Other than that, text length might be another factor that leads to the difference between our model performance. Previous work, such as Agarwal et al. [34], analysed only shorter news statements, unlike our study, which examines a whole news article. Short texts might provide less information compared to the longer texts, hence affecting the model performance [76]. In comparison to research [75], which evaluated if two short texts were authored by the same person. The authors also found that longer texts significantly enhanced the classifier performance.

Our study provides the five most significant linguistic factors to distinguish the journalist writing style from the Gemini writing style. Among all engineered features, the Recursive Feature Elimination method has ranked sentence length range, paragraph length coefficient of variation, verb ratio, sentence complex tags, paragraph length range as the top five important features for Gemini-generated news detection.

Syntactic features play an important role, as four are in the top five important features, and the most significant feature is the sentence length range. The findings suggest that journalists have greater variety and diversity in sentence and graph structure than Gemini. Accordingly, insights gained from analysing this research data show that the journalist's news has mean values of sentence length range, paragraph length coefficient of variation, sentence complex tags and paragraph length range relatively higher than Gemini-generated news. This result was supported by the study of Ma et al. [37], as they stated there is a lack of variability and natural flow in AI-generated text. This could have been the reason why syntactic-related features have been commonly used in the text classification tasks [36,61,62] and the effectiveness of authorship analysis [75]. As such, sentences, paragraph structure, and complexity can be the dominant factors contributing to detect AI-generated content.

Verb ratio is also an important predictor for the Random Forest classifier to distinguish journalist and Gemini news. The average of verbs used in journalists' articles is higher than the average of verbs used in Gemini's articles. This suggests that journalists' writing style focuses more on explaining actions, which makes the article more interesting to engage the audience. The average noun-to-verb ratio (Figure 3) shows that the average noun-to-verb ratio in journalists' news is higher than in Gemini's news. Therefore, Gemini, in contrast, uses more nouns than verbs, focusing more on explaining objects. Wu's study [77] also proved that AI is more objective than humans.

As previously highlighted in the related work section, readability features were not effective predictors because the text collected for analysis was short [36]. In our research, with longer text collected for analysis, readability features were still ranked as insignificant indicators to distinguish human and AI content. This further implies the ability of AI to generate a text that is easy for audiences to follow and understand, which makes it harder to spot AI-generated news just by reading it.

6. Limitations and Future Work

This analysis has demonstrated the effectiveness of BERTopic in extracting valuable insights and generating high-quality content from textual data. However, this field has several avenues for future work and enhancements. Firstly, the contractions should be expanded more effectively to expand the word, including the name, such as "Putin's". The method used in this research was unable to expand this kind of word even when this word was added to my contraction lists. Several methods/libraries, such as contraction maps or pycontractions, have also failed to expand the name contractions.

Another limitation is that the text needs to be input manually into the Python environment to expand the contractions; all the methods/libraries that were used failed to expand the contractions within the text, which were stored in a CSV file.

We chose to maintain basic prompting mechanisms to simulate the average AI-generated news content that most online readers may encounter. By not incorporating Retrieval Augmented Generation (RAG) or fine-tuning support, the current scenario also serves as a useful baseline for our future work in tackling a more complex problem. The stylistic differences between the texts used in this paper may appear to be easily recognisable to some readers because we have presented the journalist- and AI-generated texts side-by-side. In most of the realistic scenarios; however, average readers may find it difficult to pick up those stylistic differences in the absence of such one-to-one comparisons. Content and topic similarities may be enough to fool readers. As such, our method can serve as an early warning system for readers to the detection of AI-generated texts.

In future work, linguistic features that can address emotions within a text should be analysed to broaden the AI-detection approach. One may explore more specific scenarios

where AI-generated news articles include malicious modifications or altered emotional features. Including elements of disinformation can enhance the practicality of the research. Expanding the assessment beyond Google’s Gemini and incorporating RAG or fine-tuning mechanisms are other promising research directions.

7. Conclusions

This research contributes to the ongoing efforts to combat the spread of misinformation and news generated by AI. Leveraging linguistic features and journalistic writing styles is useful for identifying AI-generated news articles. We prompted Google Gemini to generate informative AI-generated news content using keywords generated with BERTopic and compare them with original news articles written by select award-winning journalists. Five supervised machine learning classification algorithms (Naïve Bayes, Random Forest, Passive Aggressive, Support Vector Machine, and Gradient Boosting Machine) were applied to predict articles’ originality (Journalist/Gemini) using lexical-related features, syntactic-related features, and readability-related features which were engineered from the raw text of Journalist and Gemini articles. Random Forest performed the best, with an accuracy score of 98.3%, a precision score of 0.984, a recall score of 0.983, and an F1-score of 0.983. Syntactic features like sentence length range, paragraph length coefficient of variation, sentence complex tags, paragraph length range, and lexical features, like verb ratio, were identified as significant linguistic indicators to detect AI-generated news. By proposing an innovative AI prompting approach and key linguistic indicators to distinguish AI-generated content from human-generated content, this research contributes to future research of evaluating the trustworthiness and authorship of content. This study has practical applications not only in the news and media industry but also across various industries. For instance, the academic research sector can guarantee that research papers are original and free from bias. The education industry can enhance academic integrity and quality assurance in student submissions and learning materials.

Author Contributions: Conceptualization: Y.S.; formal analysis: V.H.T.; investigation: V.H.T.; methodology: V.H.T. and Y.S.; project administration: Y.S., A.K. and S.A.; resources: V.H.T.; supervision: Y.S., A.K. and S.A.; writing—original draft: V.H.T.; Writing—review and editing: V.H.T., Y.S., A.K. and S.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The original data presented in the study are available on FigShare at <https://doi.org/10.6084/m9.figshare.27619839>.

Acknowledgments: We thank the anonymous reviewers for their valuable comments and suggestions.

Conflicts of Interest: The authors declare that they have no conflict of interest.

Abbreviations

| | |
|----------|---|
| Chat-GPT | Chat Generative Pre-Trained Transformer |
| NLP | Natural Language Processing |
| NLTK | Natural Language Toolkit |
| ANNOVA | Analysis of Variance |
| AI | Artificial Intelligence |
| BERTopic | Bidirectional Encoder Representations from Transformers for Topic Modelling |
| TF-IDF | Term Frequency-Inverse Document Frequency |
| ACS | American Chemical Society |
| IT | Information Technology |
| SVM | Support Vector Machines |
| CSV | Comma-Separated Values |
| HDBSCAN | Hierarchical Density-Based Spatial Clustering of Applications with Noise |
| NaN | Not a Number |
| NBA | National Basketball Association |

References

1. Dergaa, I.; Chamari, K.; Zmijewski, P.; Ben Saad, H. From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biol. Sport* **2023**, *40*, 615–622. [CrossRef] [PubMed]
2. Thorp, H.H. CHATGPT is fun, but not an author. *Science* **2023**, *379*, 313. [CrossRef] [PubMed]
3. Moran, C. ChatGPT Is Making up Fake Guardian Articles. Here’s How We’re Responding. The Guardian, 2023. Available online: <https://www.theguardian.com/commentisfree/2023/apr/06/ai-chatgpt-guardian-technology-risks-fake-article> (accessed on 2 August 2023).
4. Robitzski, D. New AI Generates Horrifyingly Plausible Fake News. Futurism, 2019. Available online: <https://futurism.com/ai-generates-fake-news> (accessed on 3 August 2023).
5. Thompson, S.A. AI-Generated Content Discovered on News Sites, Content Farms and Product Reviews. The New York Times, 2023. Available online: <https://www.nytimes.com/2023/05/19/technology/ai-generated-content-discovered-on-news-sites-content-farms-and-product-reviews.html> (accessed on 2 August 2023).
6. Hurst, L. Ai-generated Fake News Websites Driving Spread of Misinformation. Euronews, 2023. Available online: <https://www.euronews.com/next/2023/05/02/rapid-growth-of-news-sites-using-ai-tools-like-chatgpt-is-driving-the-spread-of-misinforma> (accessed on 10 August 2023).
7. Fitch, A. Readers Beware: AI Has Learned to Create Fake News Stories. The Wall Street Journal, 2019. Available online: <https://www.wsj.com/articles/readers-beware-ai-has-learned-to-create-fake-news-stories-11571018640> (accessed on 2 August 2023).
8. Armstrong, M.; Richter, F. Infographic: Fake News Is a Real Problem. Statista Daily Data, 2016. Available online: <https://www.statista.com/chart/6795/fake-news-is-a-real-problem/> (accessed on 2 August 2023).
9. Citron, I.; Creasy, H.; Rose, V.; O’Connor, E.F.; Din, A.H. Fact or fake news: What are Ai Chatbots telling our patients about aesthetic surgery? *J. Plast. Reconstr. Aesthetic Surg.* **2023**, *86*, 280–287. [CrossRef] [PubMed]
10. Gravel, J.; D’Amours-Gravel, M.; Osmanlliu, E. Learning to fake it: Limited responses and fabricated references provided by ChatGPT for medical questions. *Mayo Clin. Proc. Digit. Health* **2023**, *1*, 226–234. [CrossRef]
11. Li, X.; Zhang, Y.; Malthouse, E.C. A preliminary study of CHATGPT on news recommendation: Personalization, provider fairness, fake news. *arXiv* **2023**, arXiv:2306.10702.
12. Frosolini, A.; Gennaro, P.; Cascino, F.; Gabriele, G. In reference to ‘role of chat GPT in public health’, to highlight the AI’s incorrect reference generation. *Ann. Biomed. Eng.* **2023**, *51*, 2120–2122. [CrossRef]
13. Wu, Y.; Agarwal, P.K.; Li, C.; Yang, J.; Yu, C. Toward Computational Fact-checking. *Proc. VLDB Endow.* **2014**, *7*, 589–600. [CrossRef]
14. Ciampaglia, G.L.; Shiralkar, P.; Rocha, L.M.; Bollen, J.; Menczer, F.; Flammini, A. Computational Fact Checking from Knowledge Networks. *PLoS ONE* **2015**, *10*, e0128193. [CrossRef]
15. Shi, B.; Weninger, T. Fact Checking in Heterogeneous Information Networks. In Proceedings of the 25th International Conference Companion on World Wide Web—WWW’16 Companion, Montreal, QC, Canada, 11–15 April 2016; pp. 101–102. [CrossRef]
16. Etzioni, O.; Banko, M.; Soderland, S.; Weld, D.S. Open information extraction from the web. *Commun. ACM* **2008**, *51*, 68–74. [CrossRef]
17. Magdy, A.; Wanas, N. Web-based statistical fact checking of textual documents. In Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, Toronto, ON, Canada, 30 October 2010; pp. 103–110. [CrossRef]
18. Ginsca, A.L.; Popescu, A.; Lupu, M. Credibility in Information Retrieval. *Found. Trends Inf. Retr.* **2015**, *9*, 355–475. [CrossRef]
19. Acemoglu, D.; Ozdaglar, A.; ParandehGheibi, A. Spread of (mis)information in social networks. *Games Econ. Behav.* **2010**, *70*, 194–227. [CrossRef]
20. Budak, C.; Agrawal, D.; El Abbadi, A. Limiting the spread of misinformation in social networks. In Proceedings of the 20th International Conference on World Wide Web—WWW’11, Hyderabad, India, 28 March 2011; pp. 665–674. [CrossRef]
21. Nguyen, N.P.; Yan, G.; Thai, M.T.; Eidenbenz, S. Containment of misinformation spread in online social networks. In Proceedings of the 3rd Annual ACM Web Science Conference on—WebSci’12, Evanston, IL, USA, 22–24 June 2012. [CrossRef]
22. Kwon, S.; Cha, M.; Jung, K.; Chen, W.; Wang, Y. Prominent Features of Rumor Propagation in Online Social Media. In Proceedings of the 2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, 7–10 December 2013; pp. 1103–1108. [CrossRef]
23. Mocanu, D.; Rossi, L.; Zhang, Q.; Karsai, M.; Quattrociocchi, W. Collective attention in the age of (mis)information. *Comput. Hum. Behav.* **2015**, *51*, 1198–1204. [CrossRef]
24. Tambuscio, M.; Ruffo, G.; Flammini, A.; Menczer, F. Fact-checking Effect on Viral Hoaxes. In Proceedings of the 24th International Conference on World Wide Web—WWW ’15 Companion, Florence, Italy, 18–22 May 2015. [CrossRef]
25. Wei, Z.; Chen, J.; Gao, W.; Li, B.; Zhou, L.; He, Y.; Wong, K.F. An Empirical Study on Uncertainty Identification in Social Media Context. *World Sci. Ebooks* **2017**, 79–88. [CrossRef]
26. Chen, Y.; Conroy, N.K.; Rubin, V.L. News in an online world: The need for an ‘automatic crap detector’. *Proc. Assoc. Inf. Sci. Technol.* **2015**, *52*, 1–4. [CrossRef]
27. Rubin, V.; Conroy, N.; Chen, Y. Towards News Verification: Deception Detection Methods for News Discourse Towards News Verification: Deception Detection Methods for News Discourse. 2015. Available online: https://sites.socsci.uci.edu/~lpearl/colareadinggroup/readings/RubinEtAl2015_DeceptionDetectionNews.pdf (accessed on 12 November 2023).
28. Badaskar, S.; Agarwal, S.; Arora, S. Identifying Real or Fake Articles: Towards better Language Modeling. In Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 7–12 January 2008; Volume II.

29. Afroz, S.; Brennan, M.; Greenstadt, R. Detecting Hoaxes, Frauds, and Deception in Writing Style Online. In Proceedings of the 2012 IEEE Symposium on Security and Privacy, San Francisco, CA, USA, 20–23 May 2012. Available online: <https://ieeexplore.ieee.org/document/6234430> (accessed on 1 November 2024).
30. Rubin, V.; Conroy, N.; Chen, Y.; Cornwell, S. Fake News or Truth? Using Satirical Cues to Detect Potentially Misleading News. In Proceedings of the Second Workshop on Computational Approaches to Deception Detection, San Diego, CA, USA, 17 June 2016. [CrossRef]
31. Potthast, M.; Kiesel, J.; Reinartz, K.; Bevendorff, J.; Stein, B. A Stylometric Inquiry into Hyperpartisan and Fake News. In Proceedings of the Association for Computational Linguistics, Vancouver, BC, Canada, 30 July–4 August 2017.
32. Popat, K.; Mukherjee, S.; Strötgen, J.; Weikum, G. Credibility assessment of textual claims on the web. In Proceedings of the 25th ACM International Conference on Information and Knowledge Management, Indianapolis, IN, USA, 24–28 October 2016. [CrossRef]
33. Perez-Rosas, V.; Kleinberg, B.; Lefevre, A.; Mihalcea, R. Automatic detection of fake news. In Proceedings of the 27th International Conference on Computational Linguistics, Santa Fe, NM, USA, 20–26 August 2018.
34. Agarwal, V.; Sultana, H.P.; Malhotra, S.; Sarkar, A. Analysis of classifiers for fake news detection. *Procedia Comput. Sci.* **2019**, *165*, 377–383. [CrossRef]
35. Kumarage, T.; Garland, J.; Bhattacharjee, A.; Trapeznikov, K.; Ruston, S.; Liu, H. Stylometric Detection of AI-Generated Text in Twitter Timelines. *arXiv* **2023**, arXiv:2303.03697.
36. Herbold, S.; Hautli-Janisz, A.; Heuer, U.; Kikteva, Z.; Trautsch, A. A large-scale comparison of human-written versus ChatGPT-generated essays. *Sci. Rep.* **2023**, *13*, 18617. [CrossRef]
37. Ma, Y.; Liu, J.; Yi, F.; Cheng, Q.; Huang, Y.; Lu, W.; Liu, X. AI vs. Human-Differentiation Analysis of Scientific Content Generation. *arXiv* **2024**, arXiv:2301.10416.
38. Google. Introducing Gemini: Our Largest and Most Capable AI Model. Google. 2023. Available online: <https://blog.google/technology/ai/google-gemini-ai/#performance> (accessed on 5 January 2024).
39. Yana, G.Y. Does Gemini Generate Humanized Writing? Read or Die! 2024. Available online: <https://medium.com/read-or-die/does-gemini-generate-humanized-writing-746ce5e78d1c> (accessed on 1 June 2024).
40. Thomas, L. Friedman Official Biography. 2024. Available online: <https://www.thomaslfriedman.com/official-bio/> (accessed on 11 January 2024).
41. 2022 Gordon Award Winner Konrad Marshall. Melbourne Press Club. Available online: <https://www.melbournepressclub.com/article/2022-gordon-award-winner-konrad-marshall> (accessed on 14 January 2024).
42. Dave Philipps. The New York Times. Available online: <https://www.nytimes.com/by/dave-philipps> (accessed on 16 January 2024).
43. Nicholas, D. Kristof of The New York Times. The Pulitzer Prizes. Available online: <https://www.pulitzer.org/winners/nicholas-d-kristof> (accessed on 19 January 2024).
44. David Swan | LinkedIn. (n.d.). Available online: <https://au.linkedin.com/in/davidswanoz> (accessed on 20 January 2024).
45. Maier, D.; Waldherr, A.; Miltner, P.; Wiedemann, G.; Niekler, A.; Keinert, A.; Pfetsch, B.; Heyer, G.; Reber, U.; Häussler, T.; et al. Applying LDA topic modeling in communication research: Toward a valid and reliable methodology. *Commun. Methods Meas.* **2018**, *12*, 93–118. [CrossRef]
46. Egger, R.; Yu, J. A topic modeling comparison between LDA, NMF, Top2Vec; Bertopic to demystify twitter posts. *Front. Sociol.* **2022**, *7*, 886498. [CrossRef]
47. Cao, J.; Fang, A.C. Investigating variations in adjective use across different text categories. *Adv. Comput. Linguist. J. Res. Comput. Sci.* **2009**, *41*, 207–216.
48. Distante, E. BERTopic: Topic Modeling As You Have Never Seen It Before. Medium, 2022. Available online: <https://medium.com/@data-replay-it-datedtech/bertopic-topic-modeling-as-you-have-never-seen-it-before-abb48bbab2b2> (accessed on 20 December 2023).
49. BERTopic, Spacy Universe (No Date) BERTTopic. Available online: <https://spacy.io/universe/project/bertopic> (accessed on 20 December 2023).
50. Natural Language Toolkit. NLTK. Available online: <https://www.nltk.org/> (accessed on 1 December 2023).
51. NLTK tokenize package. NLTK. Available online: <https://www.nltk.org/api/nltk.tokenize.html> (accessed on 1 December 2023).
52. Albanese, N.C. Topic Modeling with LSA, PLSA, LDA, NMF, Bertopic, top2vec: A Comparison. Medium, 2022. Available online: <https://towardsdatascience.com/topic-modeling-with-lsa-plsa-lda-nmf-bertopic-top2vec-a-comparison-5e6ce4b1e4a5#5763> (accessed on 10 December 2023).
53. Deepanshi. Text Preprocessing in NLP with Python Codes. Analytics Vidhya. Available online: <https://www.analyticsvidhya.com/blog/2021/06/text-preprocessing-in-nlp-with-python-codes/> (accessed on 10 December 2023).
54. Nugues, P.M. Words, parts of speech; morphology. In *Springer eBooks*; Springer: Berlin/Heidelberg, Germany, 2006; pp. 113–145. [CrossRef]
55. Lex, E.; Granitzer, M.; Muhr, M.; Juffinger, A. Stylometric features for emotion level classification in news related blogs. *Adapt. Pers. Fusion Heterog. Inf.* **2010**, 132–133. [CrossRef]

56. Rittman, R.; Wacholder, N. Adjectives and adverbs as indicators of affective language for automatic genre detection. In Proceedings of the AISB 2008 Convention Communication, Interaction and Social Intelligence, Aberdeen, UK, 1–4 April 2008; Volume 1, p. 65.
57. Sokolova, M.; Lapalme, G. Verbs as the Most Affective Words. 2008. Available online: <http://rali.iro.umontreal.ca/rali/sites/default/files/publis/VerbsAffect2.pdf> (accessed on 10 December 2023).
58. Marinellie, S.A.; Johnson, C.J. Nouns and Verbs: A comparison of Definitional style. *J. Psycholinguist. Res.* **2004**, *33*, 217–235. [CrossRef]
59. Polinsky, M.; Magyar, L. Headedness and the Lexicon: The Case of Verb-to-Noun Ratios. *Languages* **2020**, *5*, 9. [CrossRef]
60. Koizumi, R.; In’Nami, Y. Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System* **2012**, *40*, 554–564. [CrossRef]
61. Mendhakar, A.; Darshan, H.S. Parts-of-Speech (PoS) Analysis and Classification of Various Text Genres. *Corpus-based Stud. Across Humanit.* **2023**, *1*, 99–131. [CrossRef]
62. Rittman, R.; Wacholder, N.; Kantor, P.; Ng, K.B.; Strzalkowski, T.; Sun, Y. Adjectives as indicators of subjectivity in documents. *Proc. Am. Soc. Inf. Sci. Technol.* **2004**, *41*, 349–359. [CrossRef]
63. Weiss, Z.; Riemenschneider, A.; Schröter, P.; Meurers, D. Computationally Modeling the Impact of Task-Appropriate Language Complexity and Accuracy on Human Grading of German Essays. In Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications, Florence, Italy, 2 August 2019. [CrossRef]
64. Karlgren, J.; Cutting, D. Recognizing text genres with simple metrics using discriminant analysis. In Proceedings of the 15th Conference on Computational Linguistics, Kyoto, Japan, 5–9 August 1994. [CrossRef]
65. McLellan, T.; Dobinson, G. Tertiary education of journalists and the readability of Australian newspapers. *Aust. Journal. Rev.* **2004**, *26*, 155. Available online: <http://search.informit.com.au/fullText;dn=200408351;res=IELAPA> (accessed on 6 February 2024).
66. Kelly, N.E.W.; Murray, K.E.; McCarthy, C.; O’shea, D.B. An objective analysis of quality and readability of online information on COVID-19. *Health Technol.* **2021**, *11*, 1093–1099. [CrossRef]
67. Scott, B. The Coleman-Liau Readability Formula. ReadabilityFormulas.com. 27 September 2023. Available online: <https://readabilityformulas.com/the-coleman-liau-readability-formula/> (accessed on 25 December 2023).
68. Rohith, H.P.; Sooda, K.; Rai, K.B.; Srinivas, D.B. A natural language processing system for truth detection and text summarization. In Proceedings of the 2023 7th International Conference on Computing Methodologies and Communication (ICCMC), Erode, India, 23–25 February 2023. [CrossRef]
69. Shabani, V.; Havolli, A.; Maraj, A.; Fetahu, L. Fake news detection using naive Bayes classifier and passive aggressive classifier. In Proceedings of the 2023 12th Mediterranean Conference on Embedded Computing (MECO), Budva, Montenegro, 6–10 June 2023. [CrossRef]
70. Fayaz, M.; Khan, A.; Bilal, M.; Khan, S.U. Machine learning for fake news classification with optimal feature selection. *Soft Comput.* **2022**, *26*, 7763–7771. [CrossRef]
71. Alshaher, H. Studying the Effects of Feature Scaling in Machine Learning. Ph.D. Dissertation, North Carolina Agricultural and Technical State University, Greensboro, NC, USA, 2021.
72. Cai, J.; Luo, J.; Wang, S.; Yang, S. Feature selection in machine learning: A new perspective. *Neurocomputing* **2018**, *300*, 70–79. [CrossRef]
73. Hossin, M.; Sulaiman, M.N. A review on Evaluation Metrics for Data Classification Evaluations. *Int. J. Data Min. Knowl. Manag. Process* **2015**, *5*, 1–11. [CrossRef]
74. Cross-Validation: Evaluating Estimator Performance—Scikit-Learn 0.21.3 Documentation. Scikit-Learn.org. 2009. Available online: https://scikit-learn.org/stable/modules/cross_validation.html (accessed on 12 February 2024).
75. Canbay, P.; A Sezer, E.; Sever, H. Binary background model with geometric mean for author-independent authorship verification. *J. Inf. Sci.* **2021**, *49*, 448–464. [CrossRef]
76. Jia, O.; Huang, H.; Ren, J.; Xie, L.; Xiao, Y. Contrastive learning with text augmentation for text classification. *Appl. Intell.* **2023**, *53*, 19522–19531. [CrossRef]
77. Wu, Y. Is Automated Journalistic Writing Less Biased? An Experimental Test of Auto-Written and Human-Written News Stories. *Journal. Pract.* **2019**, *14*, 1008–1028. [CrossRef]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.