mathematics

MDPI

Article

Distinguishing Human- and AI-Generated Image Descriptions Using CLIP Similarity and Transformer-Based Classification

Daniela Onita *,†, Matei-Vasile Căpîlnaș and Adriana Baciu (Birlutiu) *,†

Department of Computer Science and Engineering, "1 Decembrie 1918" University of Alba Iulia, 5, Gabriel Bethlen, 515900 Alba Iulia, Romania; capilnas.matei@uab.ro

* Correspondence: daniela.onita@uab.ro (D.O.); adriana.birlutiu@uab.ro (A.B.)

† These authors contributed equally to this work.

Abstract

Recent advances in vision-language models such as BLIP-2 have made AI-generated image descriptions increasingly fluent and difficult to distinguish from human-authored texts. This paper investigates whether such differences can still be reliably detected by introducing a novel bilingual dataset of English and Romanian captions. The English subset was derived from the T4SA dataset, while AI-generated captions were produced with BLIP-2 and translated into Romanian using MarianMT; human-written Romanian captions were collected via manual annotation. We analyze the problem from two perspectives: (i) semantic alignment, using CLIP similarity, and (ii) supervised classification with both traditional and transformer-based models. Our results show that BERT achieves over 95% cross-validation accuracy (F1 = 0.95, ROC AUC = 0.99) in distinguishing AI from human texts, while simpler classifiers such as Logistic Regression also reach competitive scores (F1 $\approx$ 0.88). Beyond classification, semantic and linguistic analyses reveal systematic cross-lingual differences: English captions are significantly longer and more verbose, whereas Romanian texts—often more concise—exhibit higher alignment with visual content. Romanian was chosen as a representative low-resource language, where studying such differences provides insights into multilingual AI detection and challenges in vision-language modeling. These findings emphasize the novelty of our contribution: a publicly available bilingual dataset and the first systematic comparison of human vs. AI-generated captions in both high- and low-resource languages.

Keywords: human vs. AI text generation; AI-generated text detection; image–text alignment; transformer-based models; multilingual natural Language processing

MSC: 68T07; 68T50; 68T45

check for updates

Academic Editor: Dongping Zhu

Received: 2 September 2025
Revised: 17 September 2025
Accepted: 19 September 2025
Published: 9 October 2025

Citation: Onita, D.; Căpîlnaș, M.-V.; Baciu, A. Distinguishing Human- and AI-Generated Image Descriptions Using CLIP Similarity and Transformer-Based Classification. *Mathematics* **2025**, *13*, 3228. https://doi.org/10.3390/math13193228

Copyright: © 2025 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

With the rapid advancements in generative Artificial Intelligence (AI), large-scale models are increasingly capable of producing fluent, coherent, and contextually relevant text. One area where these models are widely used is automatic image captioning, where the goal is to generate a descriptive text that matches the visual content of an image. While early models struggled to generate human-like captions, recent architectures now produce descriptions that are often difficult to distinguish from those written by humans.

Understanding the boundaries between human and machine language is crucial in contexts such as academic integrity, authorship verification, and creative content attribution.

This leads to a central question: Can we reliably distinguish between human-written and AI-generated image descriptions? To answer this question, we first constructed a bilingual dataset consisting of image descriptions, human-written and AI-generated, in English and Romanian, and then we investigated how to distinguish between human-written and AI-generated image descriptions.

For each language, English and Romanian, the dataset contains an equal number of human-written and AI-generated captions. Human-authored descriptions were collected through manual annotation, while the AI-generated counterparts were produced using a state-of-the-art vision-language model. The AI-generated descriptions were obtained using the BLIP-2 model (Bootstrapped Language-Image Pretraining 2) [1], and then translated into Romanian using MarianMT [2]. This balanced composition allows for controlled comparative analyses of linguistic patterns, stylistic differences, and the detectability of machine-generated content across both languages. The dataset, together with source code for reproducing our experiments, is publicly available on GitHub (https://github.com/Dani25/Human-vs-AI (accessed on 1 September 2025)). The dataset includes references to the original images from the public T4SA corpus [3] but not the images themselves. Instead, we provide the corresponding image file names, together with the human- and AI-generated descriptions.

We investigated the question of reliably distinguishing between human-written image description and AI-generated text from both a semantic similarity and classification perspective.

1. Answer from a semantic similarity analysis perspective. We start by analyzing image–text alignment using CLIP (Contrastive Language-Image Pretraining) [4], which measures how well a given text semantically corresponds to an image. Additional linguistic features were extracted, such as text length in characters and words, lexical diversity and sentence complexity indicators.
2. Answer from a classification perspective. We then frame the problem as a binary classification task and apply a variety of machine learning models, ranging from traditional classifiers like Logistic Regression and Naive Bayes to more advanced architectures such as XGBoost and fine-tuned BERT.

Our findings show that supervised models—particularly transformer-based ones—can learn to detect subtle stylistic and semantic cues that separate human-authored from AI-generated content. In addition to the classification perspective, the linguistic analyses (e.g., length, structure) and CLIP-based visualizations offer deeper insights into the nature of both human- and AI-generated language. By comparing CLIP scores for human- and AI-generated descriptions (in both English and Romanian), we uncover notable differences in how each type of text aligns with visual content—especially the tendency of human texts to be more emotionally or subjectively anchored.

While most existing research on AI-generated text detection has focused primarily on English, the ability of detection models to generalize across languages remains underexplored. This limitation is critical, given that AI systems are increasingly applied in multilingual environments where language diversity can affect model reliability. Investigating bilingual scenarios therefore provides valuable insights into whether current detection approaches are language-dependent, and whether they can be extended or adapted to handle non-English texts effectively. Our study addresses this gap by examining bilingual analysis, thereby contributing to the broader understanding of cross-linguistic robustness in AI detection models. The inclusion of cross-linguistic analysis in this work represents a novel contribution, offering insights into how language affects both human and AI expression and how this impacts detection accuracy.

Although English has been extensively studied in the context of AI-generated text detection, Romanian remains a comparatively low-resource language in NLP research.

This poses several challenges, including the limited availability of annotated datasets, pretrained language models, and evaluation benchmarks. In our work, Romanian captions were either written by human annotators or obtained via translation, which partially mitigates the scarcity of resources but also introduces translation artifacts. Future directions could leverage multilingual transformers (e.g., mBERT, XLM-R) or data augmentation techniques to improve robustness and cross-lingual transfer, thereby enhancing model performance in truly low-resource settings.

Our contribution is a systematic empirical study built on a newly created bilingual dataset. The novelty of this work lies in (i) the introduction of a dataset that contrasts human- and AI-generated image descriptions in both English and Romanian and (ii) the comparative analysis of semantic alignment and detectability across languages and text sources. By relying on standard yet effective methods, we emphasize reproducibility while placing the focus on the dataset itself and the insights it enables.

This work contributes to the broader field of AI-generated content detection and has implications for natural language processing (NLP), multimodal analysis, and the responsible deployment of generative models.

The remainder of this paper is structured as follows. Section 2 reviews related works. Section 3 details the materials and methods used in our study, including dataset construction, model selection, and evaluation protocols. Section 4 presents the experimental results, highlighting performance metrics across different languages and classifiers. Furthermore, in Section 4, we discuss the implications of our findings, analyze model behaviors, and address potential limitations. Finally, Section 5 concludes the paper by summarizing the key contributions and outlining possible directions for future research.

2. Related Work

The ability to detect AI-generated text has become an increasingly important task as generative language and vision-language models evolve. This section reviews prior work across three relevant directions: (1) AI-generated text detection, (2) multimodal captioning models, and (3) classification approaches for authorship attribution and synthetic content detection.

2.1. Detection of AI-Generated Text

Previous research on AI-generated text detection has focused primarily on English, benchmarking detectors on outputs from GPT-2 [5], GPT-3 [6], and ChatGPT [7]. Although these studies advance detection strategies, they rarely evaluate whether such approaches generalize across languages. Recent work has shown that detectors trained on English often lose accuracy when applied to morphologically richer or lower-resource languages [8–10]. This gap is particularly relevant for our study, which explicitly contrasts English (a high-resource language) with Romanian (a low-resource language).

Early approaches relied on surface-level features such as word length, punctuation, or lexical richness [11]. While effective against older generators, such shallow cues fail on recent LLMs that mimic human writing more closely. Fine-tuned transformers such as BERT [12] and RoBERTa [13] significantly improved robustness [14], but still struggle with adversarial paraphrasing and domain transfer [15]. Thus, a key open challenge is building detectors that are both language-agnostic and resilient to evolving generation strategies. Our work contributes by extending the evaluation of detection into a bilingual, cross-lingual setting.

2.2. Multimodal Captioning and Semantic Alignment

Advances in vision-language models have led to highly fluent image captions. Systems such as OSCAR [16], VinVL [17], and BLIP-2 [1] generate descriptions that are often indistinguishable from human text. However, evaluations often stop at surface fluency, without examining whether such captions align semantically across languages. CLIP [4] provides a way to measure this alignment via joint text–image embeddings, but studies typically evaluate only English captions. By applying CLIP to bilingual captions, we critically test whether translation or linguistic structure affects alignment quality, which has implications for low-resource languages.

2.3. Classification and Authorship Attribution

Distinguishing human from AI-generated content overlaps with authorship attribution. Classical models such as SVMs and Random Forests offer interpretability but limited contextual capacity. Modern approaches—especially fine-tuned transformers—achieve higher accuracy across tasks, including multilingual detection [9]. However, robustness remains an open problem: detectors often degrade under paraphrasing or across domains [10]. Hybrid methods combining CLIP-based semantic alignment with supervised classification [18,19] are promising, but empirical evaluation across languages remains limited. Our contribution bridges these gaps by systematically analyzing both semantic alignment and classification-based detection in a bilingual setting (English vs. Romanian), offering one of the first controlled cross-lingual evaluations of human vs. AI-generated captions.

3. Materials and Methods

3.1. Dataset Construction

We constructed a bilingual dataset consisting of 1313 images and image descriptions, both human-written and AI-generated, in English and Romanian. The images were selected from the T4SA corpus [3]. Four image descriptions were associated with each image, two in English and two in Romanian.

The image descriptions were obtained as follows:

1. Human-generated English descriptions were collected from Twitter posts and came with the original data presented in T4SAcorpus [3] (more details in Section 3.1.1 below).
2. AI-generated English descriptions were obtained by us by applying BLIP-2 (more details in Section 3.1.2 below).
3. Human-generated Romanian descriptions were collected by us by setting up a dataset website and involving students from the University of Alba Iulia (more details in Section 3.1.3 below).
4. AI-generated Romanian descriptions were obtained by translating the AI-generated English descriptions from point 2 above using MarianMT (more details in Section 3.1.4 below).

In summary, we constructed a bilingual dataset, English and Romanian, containing 1313 images and 4×1313 text descriptions. Half of these descriptions were human-written and half AI-generated, half in English and half in Romanian. Figure 1 shows an image and the associated text descriptions that were used in the experimental evaluation.

To support transparency and reproducibility, the dataset described above has been released and is publicly accessible on GitHub (https://github.com/Dani25/Human-vs-AI (accessed on 1 September 2025). The dataset is constructed using images from T4SA; however, due to copyright restrictions, we do not redistribute the images themselves. Instead, our released version includes only the image identifiers along with the corresponding textual annotations.



Figure 1. Sample image along with associated captions. The image is accompanied by four types of textual annotations: (1) a subjective English description from Twitter, (2) a description in Romanian written by students, (3) an English description generated by an AI model, and (4) a machine translation into Romanian of the AI-generated text.

3.1.1. Human-Generated English Descriptions

The data were collected from Twitter over a period of 6 months and, using an LSTM-SVM architecture, the tweets were divided into three sentiment categories: positive, neutral, and negative. In our experimental evaluation, we selected 1313 images and the corresponding 1313 tweets from each of the three sentiment categories.

3.1.2. AI-Generated English Descriptions

The AI-generated image descriptions in our dataset were produced using BLIP-2 (Bootstrapped Language-Image Pretraining v2) [1], a state-of-the-art vision-language model designed for image-to-text generation. BLIP-2 follows a modular two-stage architecture that bridges a vision encoder and a large language model through an intermediate transformer module. The intermediate transformer is a lightweight transformer with learnable query tokens that interact with the visual features through cross-attention, aligning them with language representations.

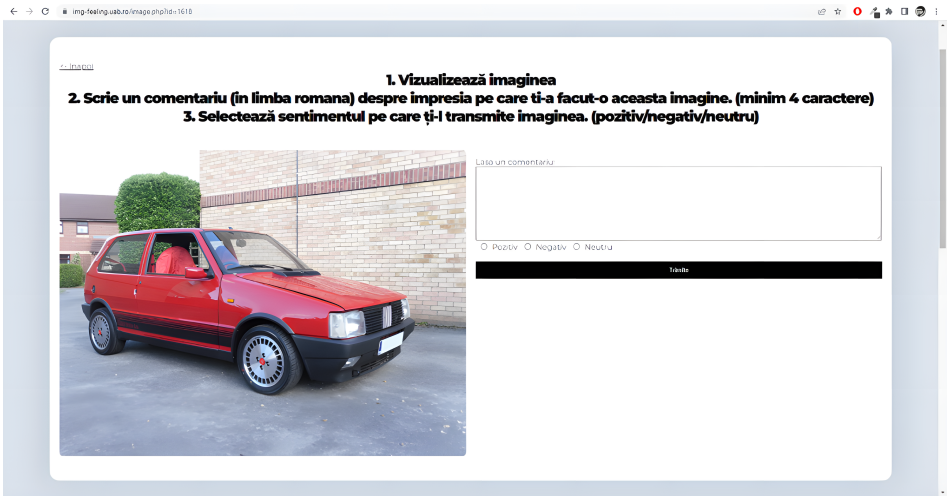
In our experiments, we used the HuggingFace implementation of BLIP-2 to generate image captions in English. Romanian captions were obtained by translating these outputs using a MarianMT translation model.

3.1.3. Human-Generated Romanian Descriptions

To collect human-generated image descriptions, we created a web application (see Figure 2). We invited students from the Faculty of Informatics and Engineering at the University of Alba Iulia to select as many images as they wished and leave a comment describing their impression of each image. Logging in was not required for submitting comments, as anonymous access was deemed sufficient. Furthermore, users were clearly informed about the purpose and implications of the website, including data collection, ensuring that no personal information was recorded.



(**a**) Displayed images.



(**b**) A description may be added once an image is selected.

Figure 2. (**a**) When a user accesses the web page, the web application displays 60 random images from the set of images. Each image has only one description associated with it; thus, once a description is added to an image, that image will not appear for the other users. (**b**) The sequence of steps undertaken by a user to submit an image description in Romanian.

To ensure consistency and avoid excessively short or overly long texts, we imposed a constraint on the length of user-submitted descriptions, limiting them to between 4 and 500 characters. This lower bound filters out empty or trivial inputs (e.g., “ok”, “nice”, or emojis), while the upper bound ensures that the content remains concise and comparable across users.

3.1.4. AI-Generated Romanian Descriptions

To generate Romanian equivalents of English image descriptions produced by the BLIP-2 captioning model, we employed theMarianMT neural machine translation system developed by the Helsinki-NLP group. Specifically, we used the pretrained model Helsinki-NLP/opus-mt-en-ro, which is based on a sequence-to-sequence Transformer architecture optimized for multilingual translation tasks. For implementation, we used the publicly available MarianMT model via the HuggingFace Transformers library [20].

3.2. Dataset Bias and Limitations

While we aimed to construct a balanced bilingual dataset, it is important to acknowledge several potential biases and limitations.

- **Source Bias:** Human-generated English captions come from Twitter posts in the T4SA corpus [3], which may reflect the demographics, topics, and language style of Twitter users rather than the general population. Similarly, Romanian human captions were produced by students, which may introduce age, educational, or cultural bias.
- **AI Generation Bias:** AI-generated English captions were produced using BLIP-2, and Romanian AI captions were obtained via machine translation. Both processes may introduce systematic artifacts, such as repetitive phrasing, translation errors, or style patterns characteristic of the model.
- **Language Resource Limitations:** Romanian is a low-resource language in NLP, limiting the diversity of available models and evaluation benchmarks. This may affect model generalization and the representativeness of our dataset for broader Romanian text.
- **Domain Coverage:** All images originate from T4SA, which focuses on Twitter-related topics. This constrains the diversity of visual content and may limit applicability to other domains or social media platforms.
- **Size Constraints:** With 1313 images and 5252 captions, the dataset is relatively small, which may affect statistical power and limit the training of large models.

These limitations suggest that models trained or evaluated on this dataset may not fully generalize to other languages, platforms, or cultural contexts. Future work could address these biases by collecting data from more diverse sources, using multiple annotator demographics, and leveraging multilingual or cross-domain augmentation techniques.

3.3. Semantic Alignment Analysis

3.3.1. CLIP Similarity Scores

We conducted a semantic alignment study using CLIP (Contrastive Language-Image Pretraining) [4]. For each image-description pair, we computed a CLIP similarity score. The CLIP similarity score is computed as the cosine similarity between the normalized embeddings of an image and a text. It measures the semantic alignment between visual and textual inputs.

$$\text{similarity}(\text{Image}, \text{Text}) = \frac{A \cdot B}{\|A\| \cdot \|B\|} \tag{1}$$

where

- $A \in \mathbb{R}^d$ is the image embedding vector (the image is passed through a visual encoder based on the Vision Transformer architecture (ViT-B/32) [21], which produces a fixed-length embedding vector A),
- $B \in \mathbb{R}^d$ is the text embedding vector (the textual description is processed by a text encoder, which tokenizes and embeds the input text into the same vector space, resulting in vector B),
- $\cdot$ denotes the dot product between vectors,
- $\|\cdot\|$ is the Euclidean (L2) norm.

Both image and text are projected into the same multimodal embedding space of dimension d, allowing for a direct semantic comparison. During pretraining, the CLIP model is optimized using a contrastive loss that encourages aligned image–text pairs to have high similarity scores, while pushing apart non-matching pairs. As a result, the cosine similarity score from Equation (1) reflects how well the semantic content of the image aligns with that of the accompanying text. This formulation enables us to quantitatively evaluate

and compare the relevance of human-written versus AI-generated descriptions. The CLIP scores range from -1 to 1, where higher values indicate a greater similarity between the image and the text.

In our experiments, CLIP similarity is especially helpful since it offers a numerical indicator of the semantic alignment between images and their associated captions in a shared multimodal embedding space. In contrast to metrics that only depend on text, CLIP assesses whether the visual content and the textual description truly match. This enables us to evaluate how accurately human- and artificial intelligence-generated captions convey an image's intended meaning, going beyond superficial linguistic features. Comparing similarity distributions between languages and sources (human vs. AI) allows us to spot systematic alignment or misalignment patterns, providing an alternative viewpoint to classification models and purely linguistic analyses.

3.3.2. Additional Linguistic Analysis

Additional linguistic features were extracted, such as

- Text length expressed in characters and words
- Lexical diversity
  - Type–Token Ratio (TTR) [22]—the proportion of unique words (types) to total words (tokens).
  - Guiraud's Index [23]—a normalized measure of lexical diversity, more robust to text length.
  - Lexical Richness—the proportion of content words (excluding stopwords) to total tokens.
- Sentence complexity indicators: for example, average sentence length, number of words per sentence, or syntactic depth. These provide insight into how elaborate or simplified the text structure is.

3.4. Classification Models

3.4.1. Problem Formulation

We define the task of distinguishing between human- and AI-generated description as a binary classification problem. Given an input text x, the goal is to predict a label $\hat{y} \in \{0, 1\}$, where

- Class 0 (Human-written): Descriptions were written by real users on Twitter (English) or by Romanian-speaking students (Romanian). These texts tend to be subjective, often conveying emotional or personal interpretations of the image.
- Class 1 (AI-generated): Descriptions were generated using the BLIP-2 image captioning model. These descriptions are generally objective and factual. For the Romanian dataset, we translated the English AI-generated texts using the MarianMT machine translation model.

For solving this classification task, we trained and compared the following classifiers: Logistic Regression, Naive Bayes, Linear SVM, XGBoost (eXtreme Gradient Boosting), and BERT (fine-tuned on our dataset).

3.4.2. Preprocessing

For the traditional models (i.e., Logistic Regression, Naive Bayes, Linear SVM, and XGBoost) we applied the following preprocessing steps:

- Lowercasing
- Removal of URLs, mentions, and special characters
- Tokenization

- Vectorization using Term Frequency-Inverse Document Frequency (TF-IDF)

For transformer-based models, such as BERT, we directly fed the raw text into a pretrained tokenizer without additional preprocessing.

3.4.3. Experimental Setup and Evaluation Metrics

We used 5-fold stratified cross-validation the following evaluation metrics:

- Accuracy
- F1 score
- ROC AUC (Area under Receiver Operating Characteristic curve)

3.4.4. Implementation Details

All models were implemented using Python 3.10.16 with libraries such as scikit-learn, xgboost, transformers, and matplotlib/seaborn.

The experiments were carried out on a workstation equipped with a NVIDIA RTX A1000 6 GB (NVIDIA, Santa Clara, CA, USA), Intel i7 GPU (Intel, Santa Clara, CA, USA), and 64 GB RAM (manufacturer unknown), running Python 3.10, scikit-learn 1.3, and HuggingFace Transformers 4.33. The models were evaluated using 5-fold stratified cross-validation to ensure robust performance estimates. We employed the following settings for the hyperparameters:

- For BERT fine-tuning, we used the HuggingFace implementation of bert-base-uncased with the following hyperparameters: maximum sequence length of 128 tokens, batch size of 16, learning rate of 2×10^{-5}, AdamW optimizer, three training epochs, and a linear learning rate scheduler with warm-up.
- For the SVM classifier, we employed a linear kernel with regularization parameter $C = 1.0$.
- For XGBoost, we set the number of estimators to 200, maximum tree depth to 6, learning rate to 0.1, and eval_metric=logloss.

4. Results and Discussions

4.1. Semantic Similarity Analysis

4.1.1. CLIP Similarity Scores

We report findings from our semantic alignment analysis using CLIP similarity scores, which measure how closely an image and its paired description align in a shared multimodal embedding space. For each image–text pair, we employed the CLIPModel.from_pretrained("openai/clip-vit-base-patch32") implementation from Hugging Face to compute raw similarity values via the logits_per_image output—representing unnormalized cosine similarities between encoded image and text embeddings. To ensure interpretability and comparability across samples, the raw scores were normalized using min–max scaling.

Figure 3 presents the normalized CLIP score distributions for image descriptions produced by humans and AI systems in both English and Romanian. The Kernel Density Estimation (KDE) curves reveal a notable performance disparity between languages and sources. AI-generated English descriptions achieve the highest average CLIP scores (mean score: 0.626), with a distribution sharply centered around higher values, outperforming human-written English descriptions (mean score: 0.513). Conversely, in Romanian, AI-generated descriptions (mean: 0.457) slightly outperform human-authored ones (mean: 0.441); the human curve peaks higher and is skewed. In summary, the interpretation from the KDE plots is as follows:

1. AI English (orange) vs. Human English (blue)

- The AI English curve is clearly shifted to the right, peaking around 0.65–0.7, indicating higher CLIP scores. This suggests that AI-generated English descriptions outperform human-generated ones in terms of visual–text alignment, as measured by CLIP.

2. Human Romanian (green) vs. AI Romanian (red)
   - The human Romanian curve peaks around 0.55–0.6, while the AI Romanian curve peaks lower, around 0.45. This implies that human-generated Romanian descriptions are more accurate or relevant than AI-generated Romanian descriptions.
3. Cross-Language and Cross-Source Comparison
   - AI performs best in English, but struggles in Romanian.
   - Humans outperform AI in Romanian, but not in English.

Overall, these results suggest that both the origin (human vs. AI) and the language (English vs. Romanian) influence the semantic alignment with images. The gap between AI and humans is language-dependent, with English favoring AI and Romanian favoring humans. Interestingly, translation may not fully preserve the alignment properties of AI-generated content, despite retaining the core meaning.

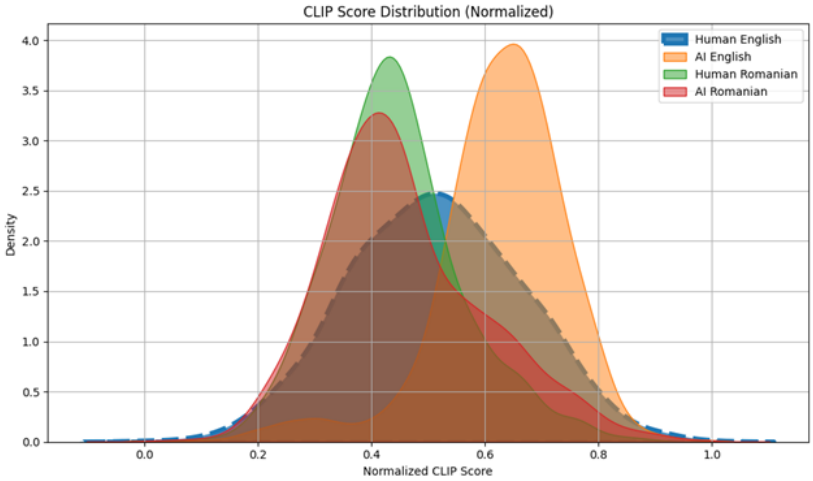


Figure 3. Kernel Density Estimation (KDE) plot illustrating the distribution of normalized CLIP similarity scores for the four categories of image descriptions. AI-generated English descriptions achieve the highest average CLIP score, outperforming human-authored English descriptions. Romanian AI-generated descriptions slightly outperform human-authored ones.

4.1.2. Linguistic Features and Length Distribution

We computed descriptive statistics for description lengths (words and characters) across four text sources. Table 1 presents the summary statistics in both word and character units. The statistics reveal several interesting patterns:

1. Human-written English descriptions are significantly longer than their Romanian counterparts (15.23 vs. 4.31 words on average), a difference likely influenced by the source platforms. While English texts were primarily collected from Twitter, where the character limit is 280, Romanian descriptions were written by students in an academic context with a more generous limit of 500 characters. Interestingly, despite having more space available, Romanian contributors tended to write much shorter and more concise texts. This may reflect differences in task framing, language economy, or writing style expectations in academic settings.

2. AI-generated texts in Romanian and English are comparable in average length (approx. 9.8–9.9 words), which reflects the consistent output structure of the BLIP-2 captioning model and the automatic translation process.
3. Romanian AI texts show the largest variance and maximum length (up to 147 words and 880 characters), possibly due to artifacts from neural machine translation.

Table 1. Descriptive statistics (mean, standard deviation, minimum, maximum) for description lengths (words and characters) for human- and AI-generated texts in Romanian and English. English descriptions are on average longer, particularly for human-authored texts.

Metric	Human EN	AI AN	Human RO	AI RO
Word Mean	15.23	9.92	4.31	9.80
Word Std	5.29	6.35	4.90	7.15
Word Min	3	2	1	1
Word Max	30	100	43	147
Char Mean	115.26	49.01	26.24	52.36
Char Std	28.08	35.25	28.80	48.94
Char Min	35	12	4	11
Char Max	160	545	248	880

Figures 4 and 5 provide complementary views of the word length distributions across the four types of image descriptions (human-written and AI-generated, in both Romanian and English). Figure 4 shows the KDE curves of word counts. These curves highlight the relatively narrow and consistent range of AI-generated texts, suggesting a stylistic uniformity imposed by the captioning model. In contrast, human-written descriptions—especially those in English—exhibit a wider distribution, reflecting greater variability in style and structure. Figure 5 presents a boxplot comparison of word lengths across the same four categories. The plot clearly distinguishes between human- and AI-generated content: human-authored English descriptions are significantly longer and more variable, while both AI-generated categories are more concise and less dispersed. Romanian human-written texts are the shortest on average, despite the input interface allowing up to 500 characters—more than Twitter's 280-character limit, which typically constrains the English human data. This suggests that students voluntarily produced shorter texts, likely influenced by the academic framing of the task.

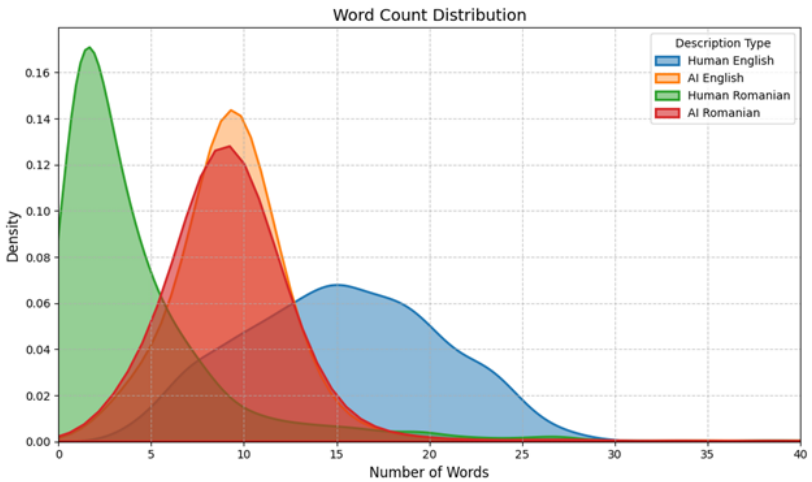


Figure 4. KDE distribution of word counts in Romanian and English texts (human vs. AI-generated). Human-written English texts are significantly longer than Romanian counterparts, while AI-generated captions show more consistency.

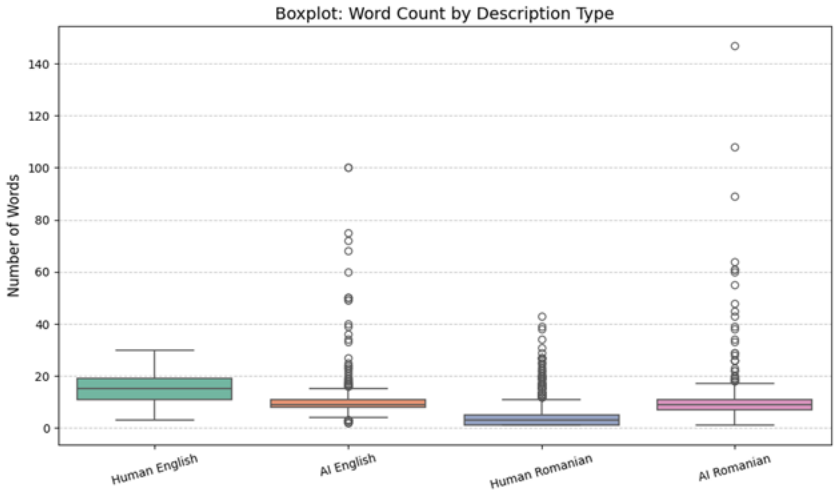


Figure 5. Boxplot comparison of word lengths across Romanian and English descriptions (human- vs. AI-generated). The boxplot clearly separates human vs. AI texts, with humans showing wider variance.

4.1.3. Lexical Diversity Analysis

We further analyzed vocabulary diversity using the following metrics:

- Type–Token Ratio (TTR) [22]—the proportion of unique words (types) to total words (tokens),
- Guiraud's Index [23]—a normalized measure of lexical diversity, more robust to text length,
- Lexical Richness—the proportion of content words (excluding stopwords) to total tokens.

Table 2 summarizes these metrics for all four categories of image descriptions: human- and AI-generated descriptions in both Romanian and English. Human-written Romanian texts exhibit the highest TTR and lexical richness, indicating greater uniqueness and expressiveness, despite being shorter than English ones. English human descriptions score highest on Guiraud's Index, due to longer and denser vocabulary. On the other hand, AI-generated descriptions (especially in English) tend to be more repetitive and structurally constrained, reflected in lower diversity across all metrics.

Table 2. Vocabulary diversity metrics for human- and AI-generated texts in Romanian and English. Human-generated texts show higher lexical diversity; AI-generated descriptions tend to be more repetitive and structurally constrained.

Metric	Human EN	AI EN	Human RO	AI RO
TTR (mean)	0.9784	0.9022	0.9912	0.9464
Guiraud Index (mean)	3.7442	2.7127	1.8275	2.8237
Lexical Richness	0.7875	0.5620	0.8447	0.5938

These results indicate that while human-authored descriptions are more lexically rich and diverse—particularly in Romanian—the AI-generated texts are more uniform and concise, potentially due to model bias toward shorter and more templated outputs. This contrast also highlights the expressive variance between natural human language and algorithmically generated captions, especially when comparing cross-lingual representations.

4.2. Classification Performance

We tested multiple classifiers on the binary classification task (human-generated vs. AI-generated) for both English and Romanian datasets. All models were evaluated using 5-fold stratified cross-validation. The evaluation metrics used were accuracy, F1 score, and ROC AUC (Area under Receiver Operating Characteristic curve). The results of the English dataset are summarized in Tables 3 and 4. Table 3 reports the performance metrics (accuracy, F1 score, and ROC AUC) evaluated on the training set, while Table 4 presents the corresponding metrics obtained by cross-validation, reflecting the generalization performance of the models. For the Romanian dataset, the results are presented in Tables 5 and 6, which separately show the model performances on the training data and on cross-validation, respectively.

Table 3. Classification performance on English dataset—Train scores (mean across folds).

Model	Train Accuracy	Train F1	Train ROC AUC
Logistic Regression	0.9383	0.9354	0.9846
Naive Bayes	0.9539	0.9538	0.9863
Linear SVM	0.9970	0.9969	0.9999
XGBoost	0.9295	0.9274	0.9872
BERT	0.9978	0.9978	0.9999

Table 4. Classification performance on English dataset—Cross-Validation scores (mean ± standard deviations across folds).

Model	CV Accuracy	CV F1	CV ROC AUC
Logistic Regression	0.883 ± 0.009	0.875 ± 0.011	0.944 ± 0.007
Naive Bayes	0.874 ± 0.011	0.877 ± 0.009	0.949 ± 0.008
Linear SVM	0.889 ± 0.008	0.885 ± 0.009	0.946 ± 0.005
XGBoost	0.865 ± 0.012	0.860 ± 0.014	0.940 ± 0.012
BERT	0.953 ± 0.010	0.953 ± 0.010	0.994 ± 0.003

Table 5. Classification performance on Romanian dataset—Train scores (mean across folds).

Model	Train Accuracy	Train F1	Train ROC AUC
Logistic Regression	0.927	0.924	0.988
Naive Bayes	0.963	0.964	0.992
Linear SVM	0.998	0.998	1.000
XGBoost	0.944	0.942	0.975
BERT	0.984	0.984	0.999

Table 6. Classification performance on Romanian dataset—Cross-Validation scores (mean ± standard deviations across folds).

Model	CV Accuracy	CV F1	CV ROC AUC
Logistic Regression	0.8762 ± 0.0129	0.8681 ± 0.0149	0.9495 ± 0.0055
Naive Bayes	0.8663 ± 0.0151	0.8726 ± 0.0120	0.9452 ± 0.0055
Linear SVM	0.8926 ± 0.0062	0.8876 ± 0.0070	0.9571 ± 0.0037
XGBoost	0.8785 ± 0.0142	0.8757 ± 0.0132	0.9375 ± 0.0071
BERT	0.9105 ± 0.0110	0.9133 ± 0.0099	0.9740 ± 0.0037

The classification results demonstrate that it is possible to distinguish between human- and AI-generated descriptions with high accuracy. On the English dataset, BERT achieved a cross-validation accuracy of over 95%, with strong F1 and ROC AUC scores, indicating both robustness and generalization. Traditional models such as Logistic Regression and

SVM also performed competitively, confirming the presence of detectable patterns even in shallow representations. In contrast, XGBoost, while still effective, showed greater variance across folds and lower validation performance, suggesting sensitivity to specific feature distributions and potential overfitting.

When comparing English and Romanian results, we observed similar trends, with BERT maintaining its lead. However, the absolute performance was slightly lower in Romanian, which may be attributed to translation artifacts introduced by MarianMT or to more variation in the student-generated texts.

4.3. Limitations and Error Analysis

Our experiments reveal a clear language-dependent pattern in semantic alignment and classification performance. AI-generated captions outperform human descriptions in English (mean CLIP score: 0.626 vs. 0.513), while in Romanian, humans achieve higher alignment (mean CLIP score: 0.441 vs. 0.457). This discrepancy can be attributed to two main factors: (i) AI English captions are generated directly by BLIP-2, benefiting from model training in English and high-quality English multimodal embeddings; (ii) Romanian AI captions are produced via machine translation using MarianMT, which may introduce artifacts and fail to fully preserve semantic nuances, compounded by Romanian being a low-resource language in NLP. Human authors, conversely, produce contextually richer and more expressive Romanian descriptions. Regarding classification performance, training accuracies for SVM and BERT approach 99%, suggesting a potential for overfitting on the relatively small dataset. However, cross-validation results (e.g., 95% accuracy for English BERT) demonstrate robust generalization. These observations highlight the challenge of multilingual AI-generated text detection: models trained predominantly in English may not transfer seamlessly to low-resource languages, and translation introduces additional complexity that can affect semantic alignment and classifier performance. Overall, these findings underline the importance of (i) carefully considering language and dataset origin in AI detection tasks, (ii) combining semantic similarity metrics (e.g., CLIP) with textual features, and (iii) developing multilingual and multimodal detection approaches that can handle the varying quality and style of AI-generated content across languages.

We performed a qualitative error analysis by reviewing misclassified examples and analyzing patterns. Despite high performance, some misclassifications occurred, especially for texts that were either very short or overly factual. In many cases, human descriptions that lacked emotional language or used generic vocabulary were incorrectly classified as AI-generated. Conversely, BLIP-2 occasionally produced captions with surprising fluency or subjectivity, blurring the line between the two classes. These cases highlight the evolving challenge of AI text detection, especially as generative models continue to improve in fluency and creativity.

Figure 6 is an example which demonstrates how the length and richness of a human-written English description can significantly influence classification outcomes. When the description is detailed and includes descriptive elements (“I can see a beautiful flower…pink-purple petals”), most models correctly identify it as human-generated, with only Naive Bayes misclassifying it. However, when the same visual content is reduced to a short, minimalistic phrase (“A purple flower”), the classification accuracy drops sharply: Logistic Regression, Naive Bayes, Linear SVM, and even BERT label it as AI-generated, leaving only XGBoost with the correct classification. This pattern suggests that several models rely heavily on lexical diversity, complexity, or the presence of stylistic cues—features that are largely absent in very short, factual statements. Interestingly, most models, except XGBoost, still correctly identify the AI-generated description, which is more elaborate and contains nuanced visual details. This indicates that while high fluency in AI-generated text

can mimic human style, subtle statistical and structural cues remain detectable for most classifiers. Overall, the example highlights a vulnerability of current detection systems: short, content-sparse human-written descriptions are particularly prone to misclassification, especially by models sensitive to surface-level linguistic features.



Human Description (detailed)
"I can see a beautiful flower. Some leaves of a plant can be seen, but the focus is on a flower with pink-purple petals."

Predictions:
Logistic Regression: Human
Naive Bayes: AI
Linear SVM: Human
XGBoost: Human
BERT: Human

Human Description (short)
"A purple flower"

Predictions:
Logistic Regression: AI
Naive Bayes: AI
Linear SVM: AI
XGBoost: Human
BERT: AI

AI Description
"The image shows a delicate, pale purple flower with a white center and prominent yellow stamens. It is held aloft by a thin, gracefully curved stem, rising above a large, glossy green leaf that occupies the lower portion of the frame."

Predictions:
Logistic Regression: AI
Naive Bayes: AI
Linear SVM: AI
XGBoost: Human
BERT: AI

Figure 6. Example of model predictions for human- and AI-generated English descriptions. Misclassifications are highlighted in red.

Figure 7 illustrates how model predictions vary depending on the nature and length of the input descriptions in Romanian. When the human-written description is detailed and includes expressive or descriptive language, all models correctly identify it as human-generated. However, when the same content is reduced to a short and factual phrase ("Floare cu petale mov." (en. "Flower with purple petals")), several models—including Logistic Regression, Naive Bayes, and Linear SVM—misclassify it as AI-generated. This suggests that such classifiers rely, at least in part, on lexical richness, complexity, or stylistic cues to distinguish between human- and machine-generated texts. Interestingly, both XGBoost and BERT correctly label the shorter version, possibly due to their higher capacity to capture contextual subtleties or structural patterns beyond surface-level features. In contrast, the AI-generated description—although more fluent and expressive than expected—is consistently recognized as AI by all models, indicating that, despite the high quality of generation, certain detectable patterns still differentiate it from human writing. Overall, this example highlights a limitation in current classifiers: human-written descriptions that are brief or neutral in tone may be prone to misclassification, especially by models more sensitive to surface linguistic cues.



Human Description (detailed)
"Este o imagine frumoasă și reușită, în care este pusă în evidență o floare. Floarea are petalele roz-mov în exterior, iar în interior sunt albe, staminele fiind albe."

Predictions:
Logistic Regression: Human
Naive Bayes: Human
Linear SVM: Human
XGBoost: Human
BERT: Human

Human Description (short)
"Floare cu petale mov."

Predictions:
Logistic Regression: AI
Naive Bayes: AI
Linear SVM: AI
XGBoost: Human
BERT: Human

AI Description
"O floare mov pal cu stamine galbene se înalță delicat deasupra frunzelor verzi lucioase, evidențiindu-se prin simplitate și eleganță naturală."

Predictions:
Logistic Regression: AI
Naive Bayes: AI
Linear SVM: AI
XGBoost: AI
BERT: AI

Figure 7. Example of model predictions for human- and AI-generated Romanian descriptions. Misclassifications are highlighted in red. The translation in English of the Romanian descriptions that appear in the figure is as follows: Human descriptions (detailed) *"It is a beautiful and well-crafted image that highlights a flower. The flower has pink-purple petals on the outside, while the inside is white, with white stamens."*; Human descriptions (short): *"Flower with purple petals"*; AI Description: *"A pale purple flower with yellow stamens rises delicately above the glossy green leaves, standing out through its natural simplicity and elegance."*

4.4. Broader Implications

This work contributes to the growing need for tools that can reliably detect AI-generated content. Our findings suggest that while transformer-based models offer strong performance, domain-specific nuances (such as language, dataset source, and caption intent) must be carefully considered. Additionally, combining semantic similarity tools (e.g., CLIP) with linguistic features and supervised classifiers may provide robust hybrid solutions for multimodal content verification.

The results of our experiments highlight clear differences between human-written and AI-generated image descriptions, both in terms of semantic alignment and linguistic characteristics. Across all experiments, transformer-based models (notably BERT) consistently outperformed traditional classifiers, suggesting that deeper contextual understanding is essential for detecting subtle stylistic and structural cues.

5. Conclusions

This research explored the increasingly relevant task of distinguishing between human-authored and AI-generated image descriptions across two languages: English and Romanian. The study combined semantic similarity analysis through CLIP, traditional linguistic evaluation, and multiple classification models to comprehensively assess whether and how such distinctions can be reliably captured by computational methods.

Our results clearly demonstrate that, despite the growing fluency of modern generative models such as BLIP-2, there remain detectable and learnable patterns that differentiate

AI-generated text from that produced by humans. These patterns are reflected both in statistical and lexical measures (e.g., Type–Token Ratio, Guiraud Index, lexical richness) and in deeper semantic alignment, with CLIP similarity scores showing mean values of 0.626 for AI English captions versus 0.513 for human English captions, and 0.457 for AI Romanian versus 0.441 for human Romanian captions.

From a classification perspective, the fine-tuned BERT models achieved the highest performance in both English and Romanian datasets, with cross-validation accuracy of 95.3% for English and 91.1% for Romanian, F1 scores of 0.953 and 0.913, respectively, and ROC AUC scores of 0.994 and 0.974. Traditional classifiers such as Logistic Regression and SVM performed well in English (CV accuracy around 88–89%) but slightly lower in Romanian (CV accuracy 87–89%), while XGBoost showed greater variability between training and validation scores, hinting at potential overfitting.

One key insight from the Romanian dataset is that human-written descriptions—produced by students—tend to be shorter on average (4.31 words), yet more expressive and semantically aligned than AI-generated translations (9.80 words), as reflected by lexical richness (0.845 vs. 0.594) and vocabulary diversity measures.

Importantly, our analysis also revealed edge cases where the distinction becomes ambiguous, particularly when human authors adopt a highly factual tone or when AI-generated text demonstrates increased stylistic fluency. These findings highlight both the progress of generative models and the limits of current detection techniques.

The novelty of our study lies in the construction of a balanced bilingual dataset and the empirical evidence it provides. By systematically comparing human- and AI-generated image descriptions in English and Romanian, we highlight both linguistic differences and the effectiveness of standard classifiers in detecting machine-generated content. This contribution supports future work on multilingual AI detection and data-driven analysis.

In conclusion, this work shows that distinguishing between human- and AI-generated image descriptions is feasible using current machine learning techniques, particularly with deep contextual models like BERT. At the same time, the subtlety of certain misclassifications and the growing capability of AI systems point to the critical importance of continuing research in this domain.

In future work, our aim is to explore several directions. First, we plan to investigate more adversarial scenarios where AI is fine-tuned to imitate human writing styles. In addition, we want to address challenges related to low-resource languages and code-switched contexts. Another focus will be on multimodal learning strategies that incorporate visual features directly. Finally, we intend to develop hybrid detection approaches that combine textual classifiers with semantic similarity measures, such as CLIP.

The line of research presented in this paper has broader implications for content verification, educational integrity, and digital trust in a world where human- and AI-generated language are becoming increasingly indistinguishable.

Author Contributions: Conceptualization, D.O. and A.B.; methodology, D.O. and A.B.; software, D.O.; validation, D.O. and A.B.; formal analysis, A.B.; investigation, D.O. and A.B.; resources, M.-V.C.; data curation, M.-V.C.; writing—original draft preparation, D.O. and A.B.; writing—review and editing, D.O. and A.B.; visualization, D.O. and A.B.; supervision, A.B.; project administration, D.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data presented in this study are openly available in https://github.com/Dani25/Human-vs-AI (accessed on 1 September 2025).

Acknowledgments: During the preparation of this manuscript/study, the authors used GPT-5 for the purposes of rephrasing. The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial intelligence
LLM	Large language model
NLP	Natural language processing
BERT	Bidirectional Encoder Representations from Transformers
BLIP	Bootstrapping Language-Image Pretraining
CLIP	Contrastive Language-Image Pretraining

References

1. Li, J.; Li, D.; Savarese, S.; Hoi, S.C. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the 40th International Conference on Machine Learning, Honolulu, HI, USA, 23–29 July 2023; Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J., Eds.; PMLR: Cambridge, MA, USA, 2023; Volume 202, pp. 19730–19742.
2. Junczys-Dowmunt, M. Marian: Fast Neural Machine Translation in C++. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 144–149.
3. Vadicamo, L.; Carrara, F.; Cimino, A.; Cresci, S.; Dell'Orletta, F.; Falchi, F.; Tesconi, M. Cross-Media Learning for Image Sentiment Analysis in the Wild. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops, Venice, Italy, 22–29 October 2017.
4. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 18–24 July 2021; pp. 8748–8763.
5. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *OpenAI Blog* **2019**, *1*, 9.
6. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
7. OpenAI. ChatGPT (Mar 14 Version), Large Language Model. 2023. Available online: https://chat.openai.com/ (accessed on 1 September 2025).
8. Ma, R.; Qian, M.; Fathullah, Y.; Tang, S.; Gales, M.; Knill, K. Cross-Lingual Transfer Learning for Speech Translation. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 33–43. [CrossRef]
9. Macko, D.; Kopal, J.; Moro, R.; Srba, I. MultiSocial: Multilingual Benchmark of Machine-Generated Text Detection of Social-Media Texts. In Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), Albuquerque, NM, USA, 29 April–4 May 2025; pp. 123–134. [CrossRef]
10. Wang, Y.; Mansurov, J.; Ivanov, P.; Su, J.; Shelmanov, A.; Tsvigun, A.; Afzal, O.M.; Mahmoud, T.; Puccetti, G.; Arnold, T.; et al. SemEval-2024 Task 8: Multidomain, Multimodel and Multilingual Machine-Generated Text Detection. In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), Mexico City, Mexico, 10–31 January 2024.
11. Gehrmann, S.; Strobelt, H.; Rush, A.M. GLTR: Statistical detection and visualization of generated text. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Florence, Italy, 28 July–2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 111–116.
12. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2018**, arXiv:1810.04805.
13. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv* **2019**, arXiv:1907.11692.
14. Zellers, R.; Holtzman, A.; Rashkin, H.; Bisk, Y.; Farhadi, A.; Roesner, F.; Choi, Y. Defending Against Neural Fake News. In Proceedings of the Advances in Neural Information Processing Systems 32 (NeurIPS 2019), Vancouver, BC, Canada, 8–14 December 2019.

15. Uchendu, A.; Campoy, D.; Menart, C.; Hildenbrandt, A. Robustness of bayesian neural networks to white-box adversarial attacks. In Proceedings of the 2021 IEEE Fourth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE), Laguna Hills, CA, USA, 1–3 December 2021; IEEE: Piscataway, NJ, USA, 2021; pp. 72–80.
16. Li, X.; Yin, X.; Li, C.; Hu, X.; Zhang, P.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. OSCAR: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; Springer: Berlin/Heidelberg, Germany, 2020; pp. 121–137. [CrossRef]
17. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Resnick, C.; Wang, Y.; Gao, J. VinVL: Making Visual Representations Matter in Vision-Language Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 5579–5588. [CrossRef]
18. Hessel, J.; Holtzman, A.; Forbes, M.; Bras, R.L.; Choi, Y. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Punta Cana, Dominican Republic, 7–11 November 2021.
19. Schlee, M.; Kant, G.; Ehrling, C.; Säfken, B.; Kneib, T. Decoding Synthetic News: An Interpretable Multimodal Framework for the Classification of News Articles in a Novel News Corpus. *Artif. Intell. Rev.* **2025**, *58*, 302. [CrossRef]
20. Tiedemann, J.; Thottingal, P. OPUS-MT—Building open translation services for the World. In Proceedings of the 22nd Annual Conference of the European Association for Machine Translation (EAMT 2020), Lisbon, Portugal, 3–5 November 2020; European Association for Machine Translation: Geneva, Switzerland, 2020; pp. 479–480.
21. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
22. Zakharova, E.Y.; Savina, O.Y. Lexical Diversity Measures' Review and Classification. *Tyumen State Univ. Her. Humanit. Res. Humanit.* **2020**, *6*, 20–34. [CrossRef]
23. Daller, H.; van Hout, R.; Treffers-Daller, J. Guiraud's index of lexical richness and advanced variants applied to bilingual speech. In *UWE Bristol Research Repository*; Introduces Guiraud Advanced, combinând tipuri avansate cu $\sqrt{}$Tokens; UWE Bristol: Bristol, UK, 2011.

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.