



Information Research – Vol. 31 No. 2 (2026)

Verification of AI-generated content

Ines Hocenski, Tomislav Jakopec, Josipa Selthofer

DOI: <https://doi.org/10.47989/ir31261629>

Abstract

Introduction. This study investigates the effectiveness of Artificial Intelligence (AI)-generated content detection tools in distinguishing between human-written and AI-produced texts in English and Croatian. As AI models are evolving rapidly, identifying generated content is becoming increasingly challenging, particularly in multilingual contexts.

Method. The research analysed twenty-four texts (both human and AI-generated) using four detection tools: Winston AI, Originality.ai, ZeroGPT and Smodin. Three expert authors produced human-written and AI-generated texts across both languages. Qualitative assessment by experts focused on analytical depth and factual accuracy, while detection and readability results were obtained quantitatively from each tool.

Analysis. Comparative analysis was conducted to evaluate accuracy, consistency, and language sensitivity of the tools. Readability metrics were also examined to identify stylistic differences between human and AI writing.

Results. The findings show high detection accuracy for English texts and inconsistent or failed detection for Croatian texts. Originality.ai demonstrated the highest overall reliability. Winston AI and Originality.ai provided readability scores, which indicated that human-written texts were more complex and suited to highly educated readers, while AI-generated texts were simpler in structure.

Conclusion. The study highlights the linguistic limitations of existing detection tools and the need for multilingual calibration. Despite its pilot scope, the research provides valuable insight into tool accuracy, readability patterns and cross-linguistic challenges in AI detection.

Introduction

The rapid proliferation of generative Artificial Intelligence (AI) has triggered significant debate across academia, education, and publishing, particularly regarding the reliability of automated systems designed to distinguish human-written from machine-generated texts.

AI represents one of the most significant technological achievements of humankind and the modern world. Among the first scientific papers that mention the term AI was the one written in 1943 by McCulloch and Pitts, which explores the *mathematical model* of a neural network. Recent papers (Bini, 2018, p. 3) define AI as ‘the computer’s ability to mimic human intelligence, including learning, decision making and problem solving.’

AI has evolved from the early theoretical models of neural computation in the mid-twentieth century to contemporary large-scale language models capable of generating coherent and contextually relevant text. While foundational work established AI as a field concerned with simulating aspects of human reasoning and decision-making, recent advances in deep learning and transformer architectures have significantly expanded its generative capabilities. These developments have fundamentally altered the landscape of academic writing, raising new challenges related to authorship verification and automated content detection.

Today, AI tools are a standard part of everyday human life, helping people communicate, conduct business, and participate in a wide range of activities. It is nearly impossible to avoid AI in most aspects of scientific work.

Based on the previously presented facts, the study aims to compare human-written texts with those generated using various AI tools across two different languages, English and Croatian. The purpose of this study is to verify the effectiveness of AI-generated content detection tools when applied to multilingual contexts.

Theoretical framework

2.1. AI-generated content and language models

The term AI encompasses a wide range of technologies combined to solve problems, conduct analysis, and interact in a manner that mimics human intelligence. AI is often referred to in literature as cognitive technology (Chan–Olmsted, 2019). To produce more content, authors frequently utilise AI tools (Russell, 2022). Text generation tools can be used to check grammar, structure, or organise a paper according to human input. According to Awan, text generators produce human-like texts that mimic the patterns and styles of a person. This includes natural language processing, content creation, coding, and user assistance (Awan, 2023).

Language is a fundamental means of communication with others, as well as with various machines and devices. The need for generalised models arises from the necessity for machines that can perform complex linguistic tasks, such as translation, summarisation, information retrieval, and communication (Naveed et al., 2023).

Natural language processing (NLP) is a form of AI that is linked to linguistics (Šuman, 2021). It is a model that provides machine understanding of the human language, as well as understanding of human meaning and communicative intent. NLP focuses on the research on how machines process and understand natural human languages and is used for translating data from natural language to machine language (Chan–Olmsted, 2019).

Historical advances in NLP have evolved from statistical to neural language modelling and from pre-trained language models (PLMs) to large language models (LLMs). A PLM is a type of model in NLP that is trained on large amounts of text data before being adapted for specific tasks such as translation, text classification, summarisation, or question answering. To train a variety of models

capable of generating natural language, these models require large amounts of data, which is what large language models enable (Naveed et al., 2024).

LLMs differ in architecture, training objectives, and scale. Transformer-based models such as generative pre-trained transformers (GPT) represent a specific subclass optimised for generative linguistic tasks, whereas the broader LLM category also includes encoder-based and hybrid architectures designed for classification or retrieval.

Although LLMs are a broader term, the term GPT has become more widely known. The relationship between them is that all GPTs are LLMs, while LLMs do not have to be GPTs. Although AI-based content generation brings numerous benefits to humans, it is challenging to determine whether content was written by a human or AI, which leads to problems, particularly in schools, higher education, and the scientific community. According to the available literature, there are two approaches to determining whether the text is AI-generated: Manual identification and automated identification using software tools. The rapid progress of language models has raised concerns about the misuse of text generation systems. With the development of GLTR (Generalised Language Tool for Recognition), it is easier for humans to detect whether the text is AI-generated or genuine. In a study conducted on human participants, GLTR was shown to improve the detection rate of inauthentic texts from 54% to 72% (Gehrmann, 2019). The literature also mentions the Raidar method (Generative AI Detection via Rewriting), which detects AI-generated content by rewriting text and calculating the edit distance of the resulting text. Raidar method shows significant results in detecting existing AI content detection models such as academic and commercial. This method works exclusively based on words and is compatible with LLMs that cannot be accessed within the system (Mao et al., 2024).

2.2. Approaches to AI-generated content detection

In manual identification methods, sentence patterns are utilised because texts written by humans tend to contain a broader range of phrases and sentences with diverse writing styles. In contrast, AI-generated texts often exhibit unique punctuation patterns (Amalia, 2023). Consequently, the searched content that requires more recent information may be incomplete or less detailed (Amalia, 2023). AI-generated content also presents different challenges. Two characteristics commonly discussed in the literature as indicators of AI-generated text are perplexity and burstiness (Ward, 2023). Perplexity represents a statistical measure used in natural language processing (NLP) to assess how predictable a sequence of words is for a language model. Texts with lower perplexity tend to contain more predictable word patterns and statistically probable constructions, which are often associated with machine-generated content. Burstiness, in contrast, refers to the degree of variation in sentence length, syntactic complexity, and structural patterns within a text. Human writing typically exhibits higher burstiness, reflecting natural fluctuations in linguistic style, whereas AI-generated texts often display more uniform sentence structure and a more consistent distribution of linguistic features.

In contrast, burstiness refers to the irregular distribution of content or concepts within the text (Ward, 2024). McLean lists several options for identifying AI-generated content, such as repetitive text, unusual vocabulary, and predictable patterns of sentence length and structure (McLean, 2024). AI-generated content detection tools, such as Copyleaks, ZeroGPT, and Winston AI, are available today. Legal frameworks related to the use of AI are still in their early stages of development. In the United States, Representative Ritchie Torres has proposed the AI Disclosure Act, which requires the labelling of AI-generated content, including photos, text, audio, and video (Smrekar, 2023). In Europe, the use of AI is to be regulated by the first legislative framework, the Artificial Intelligence Act (European Parliament, 2024).

2.3. AI-generated content detection tools

As mentioned earlier, the increasing availability and ease of use of AI programs have led to their growing adoption, both for personal and business purposes. AI-generated content detection tools are designed to distinguish between human-written and AI-generated text. These tools differ from plagiarism detection systems, which identify overlap with existing texts regardless of authorship. Similar research and discussions have been published in several papers. One of the scientific papers is the work of the authors Anderson, Belavy, Perla, et al. The authors analysed the ability of AI to generate academic papers in the field of sports medicine and exercise medicine. They generated two essays using ChatGPT and analysed their shortcomings. The key problems revealed by the research were incorrect references, lack of originality and critical thinking. The AI-generated content detection tool, GPTZero, did not detect plagiarism. The authors also state that paraphrasing the AI-generated text with additional AI tools led to an increase in the percentage of real text, as detected by the plagiarism detection tool (Anderson et al., 2023). A similar issue is addressed by the author Cao, from whose work a section related to universities and the challenges they are facing was highlighted in this paper. She states that traditional plagiarism checks often fail to detect AI-generated texts. As a result, universities are introducing specialised systems, such as GPTZero and Turnitin, which function by combining perplexity measures, stylometric analysis, and partial comparison with reference databases. However, regardless of the above, unresolved issues related to privacy and the risk of false positives remain (Cao, 2025). In a study published in the *Patterns* journal in 2023, researchers analysed ninety-one essays written by native speakers of English for the TOEFL (Test of English as a Foreign Language), using seven AI detectors. For purposes of comparison, the identical detectors were used to analyse ninety-nine essays by American students, in which the detectors correctly identified more than 90% of the essays as being written by humans. Empirical evidence suggests that detection systems may systematically misclassify texts written by non-native speakers (Liang et al., 2023). In their conclusion, they agree with the authors as mentioned above, who emphasise the need for inclusive conversations involving all stakeholders to define acceptable uses of the Generative Pre-trained Transformer (GPT) model in different contexts, especially in academic and professional settings (Liang et al, 2023). Generative AI (GenAI), such as ChatGPT, enables users in educational contexts to generate content that appears original, thereby potentially violating educational standards (Hua, 2024). In the paper presenting an extensive-scale comparison of human-written versus ChatGPT-generated essays, the authors compared essays generated by ChatGPT with those written by humans. They analysed language skills, vocabulary complexity, and structural sophistication. The research results showed that GPT-3 and GPT-4 achieve better results than humans, leading to the conclusion that the trend of using AI is growing, especially when applied to written tasks (Herbold et al., 2023). GPTZero is designed to estimate the probability that a text was AI-generated, rather than to detect plagiarism in the traditional sense. In the context of the previously mentioned research and the issues related to English as a second language, the authors Alexander et al. agree, believing that currently available tools, such as GPTZero, yield inconsistent results, especially for non-native English speakers (Alexander, 2023). Although numerous studies and results have managed to detect generated texts, there are still limitations in their ability to identify AI-generated texts. This is the case because AI models are becoming increasingly sophisticated. The most common problems that content detection tools encounter are those related to more complex or edited AI-generated texts. To avoid recognising the generated text, current AI systems replace the original words with random synonyms (Bellini et al., 2024). This does not mean that AI systems actively resist recognition, but rather that paraphrasing mechanisms within generation tools unintentionally reduce detectability.

An analysis of the available and recent literature reveals that the topic of utilising AI for writing essays, professional papers, and other similar tasks is highly popular, with numerous global studies underway. Although the results are not uniform, the authors agree on one thing: It is

necessary to work on privacy and authorship protection policies when using AI and to improve tools for detecting AI-generated content. In addition to authorship protection, privacy concerns arise because AI detection tools often require uploading sensitive documents, which may conflict with confidentiality obligations.

Methodology and research

3.1. Aim and purpose of the methodology

The research aims to analyse and compare texts created by AI tools with those written by humans across two languages, English and Croatian. Croatian was intentionally selected as a representative low-resource language in contrast with English as a high-resource language. This design enables a controlled comparison of detector performance across linguistic resource conditions. Authors writing in languages with smaller speaker populations such as Croatian often use English-language grammar correction tools, such as Grammarly, which tend to unify text consistency. This can result in more formalised text that human text recognition tools can characterise as machine-written text. The purpose of this study is to identify initial patterns and challenges in the effectiveness of currently available AI content detection tools in a multilingual context. The study does not aim to provide generalisable conclusions but rather to generate insights that can inform future large-scale research. Based on this study, it is expected that the generated content detection tools will not be able to identify the content generated by content generation tools. The scientific contribution of this work is reflected in the verified accuracy of the generated content detection tools. Furthermore, the work aims to raise awareness of the beneficial and harmful aspects of using AI tools.

3.2. Research methodology and process

At the beginning of the research process, research questions were set:

Q1: Do texts written by humans demonstrate greater depth of analysis and factual accuracy compared to AI-generated texts, as assessed qualitatively by expert reviewers?

Q2: Can AI text detection tools identify the difference between human-written and AI-generated texts?

To provide answers, the research was conducted in three phases. In the first phase, three scientists from different scientific fields wrote a professional text of at least 1,800 characters from their respective fields of expertise. In the second phase, the same scientists from the first phase composed a professional text in the field of knowledge of the other two scientists, using different content generation tools: ChatGPT, Gemini, and Perplexity. In the third phase of the research, using AI-generated content detection tools, eighteen of these texts were analysed. In total, twenty-four texts were produced: Three human-written in Croatian + three human-written in English, and six AI-generated per tool across both languages). From these, eighteen were selected for detailed analysis due to tool limitations.

Although twenty-four texts were produced, only eighteen were analysed because some detection tools could not process Croatian texts. Thus, the analysed sample comprised all twelve English texts and six Croatian texts, depending on tool compatibility. Some texts were excluded because certain detection tools did not support Croatian input or imposed character limitations. This introduces a potential selection bias, which is acknowledged as a methodological limitation.

This study builds on a previous pilot exercise in which shorter texts, of approximately 600–800 characters, were tested with detection tools. The earlier pilot suggested limited detection accuracy. The present study expands the text length to at least 1,800 characters per sample to test whether longer texts yield different detection outcomes. Texts were produced in both English and Croatian, which is a morphologically complex South Slavic language with limited

representation in AI training corpora). This bilingual design allows testing whether tool accuracy varies by language type and resource availability.

3.3. Research results

Written texts were checked against four systems designed to detect AI-generated content. The combined results of all analysed tools are available in the appendix.

3.3.1. Winston AI

Winston AI is a robust AI-generated content detector trained on vast amounts of data generated by the most widely used AI text generation tools, as well as human-generated content. According to the available documentation, it can detect if a given text was generated by an AI writing tool.

Winston AI did not provide scores for any of the samples written in the Croatian language. All human-written and AI-generated texts, ChatGPT, Gemini, and Perplexity, in Croatian are marked as N/A (Not Available) in the dataset. For the English (EN) language samples authored by humans, Winston AI demonstrated a consistent scoring pattern of 100% human. Winston AI assigned very low human scores to all AI-generated content in English, indicating a high detection rate for non-human origin. For ChatGPT (EN), all three Samples (01, 02, and 03) were assigned a human score of 1%. Gemini (EN) assigned a human score of 1% for Samples 01 and 02, while Sample 03 received a human score of 2%. Perplexity (EN) assigned a human score of 0% to Sample 01, while Samples 02 and 03 were assigned a human score of 1%.

3.3.2. Originality.ai

Originality.ai represents itself as an AI-generated text detector that is the most accurate at detecting AI-generated content, as demonstrated by both internal testing and third-party testing. It serves mainly for the SEO community, such as digital marketers, content marketers, copy editors, website publishers, and writers (Originality.ai, 2025).

Based on the research results presented in the appendix table related to the human-written texts, the tool assigned high human scores to the samples authored by humans in both languages, though there was some variation. In Croatian (HR), Originality.ai identified Human 01 and Human 03 as 100% human. Human 02 received a slightly lower but still high score of 92%. In English (EN) is like the Croatian samples, Human 01 and Human 03 were identified as 100% human. However, Human 02 received a lower human score of 74%. In AI-generated texts, Originality.ai consistently assigned very low human scores to AI-generated texts, indicating a high detection rate for AI content. In all three Samples (01, 02, and 03), in both English and Croatian, ChatGPT was assigned a human score of 0%. In all Samples (01, 02, and 03), in both English and Croatian, Gemini was assigned a human score of 0%. In English, in all three Samples (01, 02, and 03), Perplexity was assigned a human score of 0%, while in Croatian, Samples 01 and 03 were assigned 0%, while Sample 02 received a human score of 7%.

3.3.3. ZeroGPT, an AI-generated content discovery tool

ZeroGPT is a free text analysis tool that allows users to analyse the text in real-time. Its functionality is fully available to all users without any charges or restrictions on the text length (ZeroGPT, 2025). The tools were selected based on accessibility, public availability, language support, and transparency of scoring outputs. Systems requiring institutional licences, like Turnitin, were excluded due to restricted access.

In human-written texts, the accuracy of ZeroGPT in identifying human-written content varied significantly depending on the language of the sample. For English (EN), the tool demonstrated high accuracy. It identified Human 02 and Human 03 as 100% human, while Human 01 received a human score of 98.61%. In the case of the Croatian (HR) language, the results were less consistent. While Human 01 HR received a relatively high human score of 83.06%, Human 03 HR dropped to 59.83%, and Human 02 HR was assigned only a 30.73% human score, implying the

tool flagged nearly 70% of the text as AI-generated. For AI-generated texts, ChatGPT, Gemini, Perplexity, ZeroGPT consistently assigned low human scores to AI-generated content across both languages, indicating a high detection rate for artificial origin. For ChatGPT in English, human scores ranged from 0% to 2.27%. In Croatian, scores were slightly higher but remained low, ranging from 0% to 7.23%. For Gemini, the tool assigned human scores between 0% and 4.80% for English samples. For Croatian samples, the human scores ranged from 3.92% to 4.81%. For Perplexity, it showed the highest variance in ZeroGPT. While the Croatian samples received very low human scores (0.68% to 3.69%), the English Samples (EN 02 and EN 03) received human scores of 11.85% and 13.53%, respectively.

3.3.4. Smodin AI

Smodin's AI Writer is a personalised writing tool that utilises AI to produce high-quality content instantly. It generates text in over 100 languages, cites sources in your preferred format, researches facts, and provides real-time feedback (Smodin, 2025).

Based on the provided data, Smodin assigned high human scores to samples authored by humans in both languages. For Croatian (HR), the tool assigned a perfect 100% human score to all three human-written Croatian Samples (Human 01, 02, and 03). For English (EN), the tool assigned a human score of 100% to Human 02 and Human 03, while Human 01 received a slightly lower score of 89%. In the case of AI-generated texts, the tool's ability to identify AI-generated content, assigning a low human score, varied between languages and specific AI models. For ChatGPT in English, all samples received a 0% human score, while in Croatian, Samples 01 and 03 received 0%, but Sample 02 was assigned a significantly higher human score of 45%. For Gemini in English, all Samples received a 0% human score, while in Croatian, the human scores were higher and more varied: 43% for sample 01, 19% for Sample 02, and 0% for Sample 03. The English-language samples generated by Perplexity were classified as human-written with varying scores across the samples. Sample 01 received a human-likelihood score of 0%, while Sample 02 and Sample 03 received 19% and 30%, respectively. In Croatian, Samples 01 and 02 received human scores of 25% and 31%, while Sample 03 received 0%.

For the purposes of statistical analysis, a simple model was used to standardise the accuracy percentages of tools detecting whether the text was written by a human or an AI. Table 1. Detection accuracy presents the detection accuracy rates based on the article source. Due to the small sample size, more advanced statistical analyses were not conducted.

Articles	Detection accuracy			
	Winston AI	Originality.ai	ZeroGPT	Smodin AI
Human 01 HR	100%	100%	99%	89%
Human 01 EN		100%	83%	100%
Human 02 HR	100%	74%	100%	100%
Human 02 EN		92%	31%	100%
Human 03 HR	100%	100%	100%	100%
Human 03 EN		100%	40%	100%
ChatGPT 01 HR	99%	100%	98%	100%
ChatGPT 01 EN		100%	94%	100%
ChatGPT 02 HR	99%	100%	100%	100%
ChatGPT 02 EN		100%	100%	55%
ChatGPT 03 HR	99%	100%	98%	100%
ChatGPT 03 EN		100%	93%	100%
Gemini 01 HR	99%	100%	98%	100%
Gemini 01 EN		100%	96%	57%
Gemini 02 HR	99%	100%	100%	100%
Gemini 02 EN		100%	95%	81%
Gemini 03 HR	98%	100%	95%	100%
Gemini 03 EN		100%	95%	100%
Perplexity 01 HR	100%	100%	98%	100%
Perplexity 01 EN		100%	96%	75%
Perplexity 02 HR	99%	100%	88%	81%
Perplexity 02 EN		93%	99%	69%
Perplexity 03 HR	99%	100%	86%	70%
Perplexity 03 EN		100%	97%	100%
Average	99%	98%	91%	91%

Table 1. Detection accuracy.

Since Winston AI was unable to process texts in Croatian, its accuracy rates apply exclusively to the English language. Consequently, the data regarding detection performance by language are presented in Tables 2 and 3.

Articles	Detection accuracy English			
	Winston AI	Originality.ai	ZeroGPT	Smodin AI
Human	100%	91%	100%	96%
ChatGPT	99%	100%	99%	100%
Gemini	99%	100%	98%	100%
Perplexity	99%	100%	91%	84%
Average	99%	98%	97%	95%

Table 2. Detection accuracy English.

Articles	Detection accuracy Croatian		
	Originality.ai	ZeroGPT	Smodin AI
Human	97%	51%	100%
ChatGPT	100%	96%	85%
Gemini	100%	96%	79%
Perplexity	98%	98%	81%
Average	99%	85%	86%

Table 3. Detection accuracy Croatian.

Although the aggregate data indicates high (>90%) and very high (>97%) overall accuracy rates for the tools, a breakdown by the two observed languages shows that English performs better, averaging 97%, compared to Croatian, averaging 90%. These results remain remarkably high, demonstrating the practical applicability of AI-generated text detection systems. Due to the exploratory nature of the study and limited sample size, only descriptive statistics were calculated. Future research should incorporate inter-rate agreement measures such as Cohen's κ .

Discussion

Considering the first research question (Q1), expert reviewers found that AI-generated texts generally lacked in-depth analytical structure and occasionally contained factual inaccuracies. In contrast, human-written texts demonstrated greater depth and contextual accuracy. Regarding the second research question (Q2), the comparison of four detection tools showed high accuracy for English texts, but mixed or inconsistent performance for Croatian texts. When comparing the performance of the detection tools, the research indicates that texts written by humans in English were accurately recognised as such by most tools. Winston AI correctly determined that the human-written texts in English were 100% human. Originality.ai also identified human-written English texts as 'Likely original' with high confidence, showing 100% confidence for two samples and 74% confidence for another. ZeroGPT detected one English human text as 'Human-written' with a very low AI percentage. In this study, a human score below 5% is interpreted as very low, indicating a high probability of AI authorship. Smodin AI also showed a very low probability of AI content in all English human texts, reporting 0% or 11% AI content. For texts written by humans in Croatian, the detection results showed more variability. Winston AI failed to analyse texts written in Croatian altogether. Originality.ai identified all Croatian human texts as 'Likely original' with 100% confidence. Smodin AI reported 0% AI for all three Croatian human texts, supporting its general claim of accurately identifying human-written texts. ZeroGPT, however, displayed inconsistent results for human-written Croatian texts: One sample was detected as 'Mixed signals, with some parts generated by AI/GPT' with a significant AI percentage, another as 'Likely human-written, may include parts generated by AI/GPT' with a moderate AI percentage, and a third as 'Most of your text is AI/GPT-generated' with a high AI

percentage. The paper notes that Smodin AI and ZeroGPT identified these texts as not human-written 'With varying success,' although Smodin's AI human Croatian results were consistently low in AI percentage in the provided scans. Texts generated by AI tools in English through ChatGPT, Gemini, and Perplexity, were consistently detected with a high percentage of AI by all four tools. Winston AI assigned them a very low human score, indicating a high probability that an AI text generation tool was used. Originality.ai identified these as 'Likely AI' with 100% confidence, recognising AI-generated texts with the highest accuracy among the tested tools.

ZeroGPT consistently labelled them as 'AI/GPT generated' with high AI percentages. Smodin AI also reported 100% AI probability for English texts generated by these tools. For texts generated by AI tools in Croatian, through ChatGPT, Gemini, and Perplexity, Winston AI failed to provide results. Originality.ai consistently detected them as 'Likely AI' with 100% confidence. Smodin AI and ZeroGPT determined that these texts were not written by humans 'with varying success.' ZeroGPT showed high AI percentages for these texts, ranging from over 90% to over 99%. Smodin AI reported AI percentages for Croatian AI texts ranging from a moderate 55% to a very high 100%. The reduced accuracy in Croatian can be attributed to several factors, including the scarcity of training corpora in low-resource languages and the high morphological complexity of Croatian compared to English.

Regarding readability, the tools that provided assessments were Winston AI and Originality AI. These tools assessed that texts written by humans in English were intended for people with a higher level of knowledge or education, assigning scores that indicated difficulty. For instance, Winston AI reported scores ranging from twenty-nine to thirty-two out of 100, indicating a 'College graduate level' or 'Professional level.' At the same time, Originality.ai showed Flesch-Kincaid Reading Ease scores around 33–35 (where 45–60 is ideal) and Gunning Fog Indices around seventeen (where 11–13 is ideal). In contrast, Originality.ai found that the style of AI-generated texts in English was significantly more straightforward and understandable to nearly everyone. However, the specific scores from Originality.ai reports for AI English texts (F-K Ease as low as 1.1 and Gunning Fog Index consistently 19) and Winston AI reports (scores ranging from 0 to 16 with levels like 'Professional level') seem contradictory to this general statement about understandability, suggesting potential limitations or differing interpretations of readability metrics. This finding is important because even expert-written professional texts were classified as low-readability, which highlights limitations in how readability metrics interpret scientific writing. Crucially, Winston AI and Originality.ai failed to assess texts written in Croatian for readability. Their scan reports for Croatian texts, both human and AI-generated, show readability results of zero, confirming this failure. The remaining two tools, ZeroGPT and Smodin AI, failed to assess readability for any of the examined texts. Neither the ZeroGPT nor the Smodin AI scan reports contain any readability results.

Research results provided answers to the research questions. Comparing the tools for detecting generated content, the results indicate that texts written by humans in English were accurately recognised as such. Texts written in Croatian were recognised by three research tools, except for the Winston.ai tool.

Texts generated by AI tools were recognised with the highest accuracy by the Originality.ai tool. The Winston.ai tool also failed to provide results for texts generated in Croatian, whereas the Smodin AI and ZeroGPT tools, with varying success, determined that humans did not write these.

Regarding readability, the Winston.ai and Originality.ai tools assessed that texts written by humans in English were intended for people with a higher level of knowledge or education. In contrast, those generated by AI tools were designed for the general population. The tools failed to assess texts written in Croatian. As for the remaining two tools, ZeroGPT and Smodin AI, they were unable to assess readability at all for any of the mentioned texts.

Texts generated by AI tools were generally lacking in-depth analytical structure compared to those written by humans.

The non-trivial rate of misclassification observed across the tools suggests that current AI detectors should not be interpreted as definitive classification systems but rather as probabilistic indicators. The systematic misclassification of Croatian texts likely reflects the limited representation of morphologically rich languages in training corpora used for detector development. These findings align with previous studies demonstrating detector bias against non-native or linguistically diverse writing styles.

The two-language design was intentionally constructed as a controlled high-resource vs low-resource comparison rather than as a multilingual survey.

This study has several limitations. First, the sample size is small and does not allow for statistical generalisation. Second, the AI models used to generate texts are rapidly evolving, which may affect the reproducibility of results. Third, not all detection tools support Croatian, which limited the scope of analysis. These constraints reinforce the framing of the study as a pilot, intended to inform future large-scale research. Since some tools could not process Croatian texts, the dataset used for analysis was partially filtered by tool compatibility, which may have influenced comparative results.

Conclusion

The conclusion summarises the main findings of this study and situates them within the broader context of AI-generated content detection research. It reflects on the study's contributions, methodological limitations, and implications for future multilingual and cross-disciplinary research. The following section synthesises the results in relation to the research questions and outlines recommendations for further exploration of AI detection tools in diverse linguistic environments.

The first question, regarding detail and accuracy, was addressed through expert assessment. The second question was explored through tool comparison. Unlike an earlier study that suggested weak detection, this research showed notable improvements in the accuracy of several tools, particularly Originality.ai. While detection was more reliable for English texts, Croatian samples revealed limitations, especially with Winston AI, which was unable to process them. These findings highlight that the effectiveness of AI detection tools varies across languages and that multilingual testing is essential.

The results suggest improvements relative to earlier studies in AI-content detection, but also reveal ongoing challenges, particularly in multilingual contexts. The difference between detecting human-written and AI-generated text is particularly pronounced in languages spoken by small populations, such as Croatian. When using AI tools to translate into English and correct grammatical errors, the resulting text may be classified as AI-generated. The said fact leads to the conclusion that the consequences of using AI tools to verify text cannot be reliably predicted. The scientific contribution of this study lies in providing preliminary empirical indications of tool accuracy across two languages, thereby contributing to the discussion on the strengths and weaknesses of current detection systems. The study raises awareness of both the benefits and drawbacks of using AI tools, underlining the importance of developing robust detection methods, improving privacy and authorship protection, and fostering critical reflection on the use of AI in education, research, and professional communication. The findings should be interpreted with caution and cannot be generalised because of the limited number of languages. Nevertheless, they highlight key methodological and linguistic challenges that more comprehensive studies will need to address in the future. The findings should be interpreted as exploratory rather than conclusive, highlighting the need for large-scale multilingual validation studies.

Acknowledgements

For data analysis of scanning reports, the AI tool NotebookLM was used.

About the authors

Ines Hocenski is a Senior Teaching and Research Assistant at the Department of Information Sciences, Faculty of Humanities and Social Sciences in Osijek, Croatia. She received her Ph.D. from the University of Zadar. Her research interests include publishing, crisis management, crises in publishing, copyright, and marketing in publishing. She can be contacted at ihocenski@ffos.hr

Tomislav Jakopec is an Associate Professor at the Department of Information Sciences, Faculty of Humanities and Social Sciences, University of Osijek. He holds a Doctor of Science degree in the field of Information sciences from the University of Zadar and a Master of Informatics from the University of Zagreb. His research interests include information technology, web design, databases, and the design of information systems. He can be contacted at tjakopec@ffos.hr

Josipa Selthofer is an Associate Professor in the Department of Information Sciences at the Faculty of Humanities and Social Sciences in Osijek, Croatia. She earned an M.Sc. in Graphic Design from the Faculty of Graphic Arts at the University of Zagreb. She obtained a PhD in Information Science from the University of Zadar, Croatia. She has been the Head of the Double-major Graduate Study in Publishing since October 2021 and has been teaching in both the Double-major Graduate Study in Publishing and the Information Technology graduate program. He writes books and scientific papers in the field of visual communications and graphic design, and presents at international conferences.

References

- Akshay, L. C. (2018). McCulloch – Pitts Neuron – Mankind's first mathematical model of a biological neuron. TDS Archive. <https://medium.com/data-science/mcculloch-pitts-model-5fdf65ac5dd1> (Archived at <https://web.archive.org/web/20251031115553/https://medium.com/data-science/mcculloch-pitts-model-5fdf65ac5dd1>)
- Anderson, N., Belavy, D. L., Perle, S. M., Hendricks, S., Hespanhol, L., Verhagen, E., & Memon, A. R. (2023). AI did not write this manuscript, or did it? Can we trick the AI text detector into generated texts? The potential future of ChatGPT and AI in sports & exercise medicine manuscript generation. *BMJ Open Sport & Exercise Medicine*, 9(1), e001568. <https://doi.org/10.1136/bmjsem-2023-001568>
- Awan, A. A. (2023, May 24). What is text generation? DataCamp. <https://www.datacamp.com/blog/what-is-text-generation> (Archived at <https://web.archive.org/web/20251104161758/https://www.datacamp.com/blog/what-is-text-generation>)
- Amalia, A. (2023). How to identify AI-generated text: With or without software. <https://www.contentgrip.com/how-to-spot-ai-generated-text/> (Archived at <https://web.archive.org/web/20251104161045/https://www.contentgrip.com/how-to-spot-ai-generated-text/>)

Bini, S. A. (2018). Artificial Intelligence, machine learning, deep learning, and cognitive computing: What do these terms mean and how will they impact health care? *AAHKS Symposium*, 33 (8), 2358-2361. <https://doi.org/10.1016/j.arth.2018.02.067>

Bellini, V., Semeraro, F., Montomoli, J., Cascella, M., & Bignami, E. (2024). Between human and AI: Assessing the reliability of AI text detection tools. *Current Medical Research and Opinion*, 40(3), 353-358. <https://doi.org/10.1080/03007995.2024.2310086>

Cao, L. (2025). A practical synthesis of detecting AI-generated textual, visual, and audio content. *arXiv*. <https://doi.org/10.48550/arXiv.2504.02898>

Chan-Olmsted, S. M. (2019). A review of Artificial Intelligence adoptions in the media industry. *International Journal on Media Management*, 21(3-4), 1-23. <http://dx.doi.org/10.1080/14241277.2019.1695619>

European Parliament. (2024, March 13). *European Parliament legislative resolution of 13 March 2024 on the proposal for a regulation on artificial intelligence (Artificial Intelligence Act) (P9_TA(2024)0138)*. https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html (Archived at https://web.archive.org/web/20260322073421/https://www.europarl.europa.eu/doceo/document/TA-9-2024-0138_EN.html)

Gehrmann, S., Strobelt, H., & Rush, A. (2019). GLTR: Statistical detection and visualization of generated text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 111-116). Association for Computational Linguistics. <https://doi.org/10.18653/v1/p19-3019>

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, 13(1). <https://doi.org/10.1038/s41598-023-45644-9>

Hua, H., & Yao, C.-J. (2024). Investigating generative AI models and detection techniques: Impacts of tokenization and dataset size on identification of AI-generated text. *Frontiers in Artificial Intelligence*, 7. <https://doi.org/10.3389/frai.2024.1469197>

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *Patterns*, 4(7), 100779. <https://doi.org/10.1016/j.patter.2023.100779>

Mao, K., Deng, C., Chen, H., Mo, F., Liu, Z., Tetsuya Sakai, & Dou, Z. (2024). ChatRetriever: Adapting large language models for generalized and robust conversational dense retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 1227-1240). Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.71>

McLean, D. (2025, February 10). How to detect AI writing in 2025 (Expert tips). *Elegant Themes*. <https://www.elegantthemes.com/blog/business/how-to-detect-ai-writing> (Archived at <https://web.archive.org/web/20251104095622/https://www.elegantthemes.com/blog/business/how-to-detect-ai-writing>)

Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2025). A comprehensive overview of large language models. *ACM Transactions on Intelligent Systems and Technology*, 16(5), 1–72. <https://doi.org/10.1145/3744746>

Originality.ai. (2026). Home page. <https://originality.ai/> (Archived at <https://web.archive.org/web/20251104100248/https://originality.ai/>)

Russell, S. J., & Norvig, P. (2003). *Artificial Intelligence: A modern approach* (2nd edition). Pearson.

Russell, S. (2022). *Kao čovjek: Umjetna inteligencija - napredak ili prijetnja?* Planetopija, Zagreb.

Solomonoff, R. J. (1985). The time scale of artificial intelligence: Reflections on social effects. *Human Systems Management*, 5, 149–152. <https://doi.org/10.3233/HSM-1985-5207>

Smodin. (2026). Home page. <https://smodin.io/> (Archived at <https://web.archive.org/web/20260323091418/https://smodin.io/>)

Smrekar, M. (2024). Raider, lektor koji prepoznaje AI generirane tekstove. <https://www.bug.hr/umjetna-inteligencija/raidar-lektor-koji-prepoznaje-ai-generirane-tekstove-39480> (Archived at <https://web.archive.org/web/20251104123135/https://www.bug.hr/umjetna-inteligencija/raidar-lektor-koji-prepoznaje-ai-generirane-tekstove-39480>)

Šuman, S. (2021). Pregled metoda obrade prirodnih jezika i strojnog prevođenja. *Zbornik Veleučilišta u Rijeci*, 9(1), 371–384. <https://doi.org/10.31784/zvr.9.1.23>

ZeroGPT. (2026). Home page. <https://www.zerogpt.com/> (Archived at <https://web.archive.org/web/20251104135037/https://www.zerogpt.com/>)

Žejn, A. (2020). Stilometrijska analiza slovenske pripovjedne književnosti do prve izvirne slovenske priče. *Umjetnost riječi: Časopis za znanost o književnosti, izvedbenoj umjetnosti i filmu*, 64(3–4), 263–283. https://doi.org/10.22210/ur.2020.064.3_4/04

Ward, J. (2023). How to detect AI content: Proven methods for distinguishing human from machine writing. https://www.linkedin.com/posts/johnathanward_how-to-detect-ai-content-proven-methods-activity-7155346352015642624-2E-x (Archived at https://web.archive.org/web/20251104143308/https://www.linkedin.com/posts/johnathanward_how-to-detect-ai-content-proven-methods-activity-7155346352015642624-2E-x)

Winston AI. (2025). *Interpreting our AI detection scores*. <https://gowinston.ai/interpreting-our-ai-detection-scores/> (Archived at <https://web.archive.org/web/20251104144024/https://gowinston.ai/interpreting-our-ai-detection-scores/>)

Copyright

Authors contributing to *Information Research* agree to publish their articles under a [Creative Commons CC BY-NC 4.0 license](#), which gives third parties the right to copy and redistribute the material in any medium or format. It also gives third parties the right to remix, transform and build upon the material for any purpose, except commercial, on the condition that clear acknowledgment is given to the author(s) of the work, that a link to the license is provided and

that it is made clear if changes have been made to the work. This must be done in a reasonable manner, and must not imply that the licensor endorses the use of the work by third parties. The author(s) retain copyright to the work. You can also read more at: <https://publicera.kb.se/ir/openaccess>

Appendix

Articles	Human score			
	WinstonAI	Originality.ai	ZeroGPT	Smodin
Human 01 HR	N/A	100%	83,06%	100%
Human 01 EN	100%	100%	98,61%	89%
Human 02 HR	N/A	92%	30,73%	100%
Human 02 EN	100%	74%	100%	100%
Human 03 HR	N/A	100%	59,83%	100%
Human 03 EN	100%	100%	100%	100%
ChatGPT 01 HR	N/A	0%	5,80%	0%
ChatGPT 01 EN	1%	0%	2,27%	0%
ChatGPT 02 HR	N/A	0%	0%	45%
ChatGPT 02 EN	1%	0%	0%	0%
ChatGPT 03 HR	N/A	0%	7,23%	0%
ChatGPT 03 EN	1%	0%	2,06%	0%
Gemini 01 HR	N/A	0%	3,92%	43%
Gemini 01 EN	1%	0%	2,06%	0%
Gemini 02 HR	N/A	0%	4,81%	19%
Gemini 02 EN	1%	0%	0%	0%
Gemini 03 HR	N/A	0%	4,70%	0%
Gemini 03 EN	2%	0%	4,80%	0%
Perplexity 01 HR	N/A	0%	3,69%	25%
Perplexity 01 EN	0%	0%	1,95%	0%
Perplexity 02 HR	N/A	7%	0,68%	31%
Perplexity 02 EN	1%	0%	11,85%	19%
Perplexity 03 HR	N/A	0%	2,93%	0%
Perplexity 03 EN	1%	0%	13,53%	30%

Table 4. The combined results of all analysed tools.

