INTELLECTUAL INFORMATIONAL TECHNOLOGIES

ІНТЕЛЕКТУАЛЬНІ ІНФОРМАЦІЙНІ ТЕХНОЛОГІЇ

https://doi.org/10.15407/intechsys.2026.02.025
UDC 004.912:81'322

I.O. KOBYLIN, PhD (Engineering), Senior Lecturer,
Kharkiv National University of Radio Electronics,
14, Nauky ave., Kharkiv, 61166, Ukraine
https://orcid.org/0000-0002-4552-9616
ilya.kobylin@nure.ua
V.D. BIEHUNOVA, Master,
Kharkiv National University of Radio Electronics,
14, Nauky ave., Kharkiv, 61166, Ukraine
https://orcid.org/0009-0000-3804-818X
veronika.biehunova@nure.ua
D.S. TSYBAN, Master's Student,
Kharkiv National University of Radio Electronics,
14, Nauky ave., Kharkiv, 61166, Ukraine
https://orcid.org/0009-0001-2661-9034
dmytro.tsyban@nure.ua
V.O. KOVALCHUK, Master's Student,
Kharkiv National University of Radio Electronics,
14, Nauky ave., Kharkiv, 61166, Ukraine
https://orcid.org/0009-0004-3286-7888
vladyslav.kovalchuk@nure.ua

MEASURING TEXTUAL REDUNDANCY, LEXICAL RICHNESS, AND VERACITY: A MULTI-METRIC APPROACH TO TEXT EVALUATION

With the exponential growth of textual data, effective analytical methods are essential. This paper introduces a multi-metric approach to text evaluation, focusing on quantifying textual redundancy, lexical richness, and veracity. We explore the theoretical underpinnings and practical applications of the Wateriness Coefficient for redundancy, the Type-Token Ratio (TTR) and various Lexical Diversity Indices for vocabulary richness, and the Factual Accuracy Score for informational integrity. These metrics offer a comprehensive framework for assessing text quality beyond traditional methods, enabling deeper insights into linguistic characteristics and informational reliability in diverse textual datasets.

Cite: Kobylin I.O., Biehunova V.D., Tsyban D.S., Kovalchuk V.O. Measuring Textual Redundancy, Lexical Richness, and Veracity: A Multi-Metric Approach to Text Evaluation. *Information Technologies and Systems*. 2 (8). 2026. 25—45. https://doi.org/10.15407/intechsys.2026.02.025

***Keywords**: Computational Linguistics, Factual Accuracy, Lexical Diversity, Natural Language Processing, Text Analysis, Text Quality, Type-Token Ratio, Wateriness Coefficient.*

Introduction

The modern era of data generation has ushered in an exponential surge in textual information. From customer feedback and social media posts to intricate technical documentation and insightful market reports, vast reservoirs of text are continuously accumulating across diverse sectors. This overwhelming volume of unstructured data creates a pressing need for effective methods to organize and analyze it. Indeed, extracting meaningful insights from these textual troves has become absolutely essential for informed decision-making and strategic planning in nearly every field.

Simply counting words or tracking their basic presence no longer suffices for truly evaluating and understanding the nuances of textual content. For a more comprehensive appraisal, a range of specialized metrics have been developed, each designed to delve into different dimensions of text, such as its lexical richness, informational integrity, and even its rhetorical style. Previous studies have explored a variety of approaches to assessing text quality [1—6]. Traditional readability indices such as the Flesch Reading Ease and Gunning Fog Index evaluate surface-level linguistic complexity through sentence length and word difficulty. Later research introduced lexical diversity measures — Type-Token Ratio (TTR), Moving Average TTR (MATTR), and Measure of Textual Lexical Diversity (MTLD) — to quantify vocabulary richness. More recent developments in Natural Language Processing (NLP) apply machine learning models such as BERTScore and ROUGE to assess semantic coherence and factual consistency. However, most existing studies focus on individual aspects of text evaluation rather than integrating multiple perspectives.

This paper addresses this limitation by presenting a multi-metric approach for comprehensive text evaluation, exploring various quantitative measures that go far beyond superficial analyses of text quality. We examine the theoretical foundations and practical applications of metrics like the Wateriness Coefficient, which quantifies textual redundancy by measuring the proportion of common, or "stop" words. We also delve into the Type-Token Ratio and other Lexical Diversity Indices, crucial tools for assessing vocabulary richness by comparing unique words to the total word count. Furthermore, we address the critical role of the Factual Accuracy Score in evaluating informational integrity and the Manipulation Score in detecting emotionally charged language. The Coherence Score, which measures logical interconnectedness, is also considered, contributing to a text's rhetorical and structural integrity. By integrating these diverse metrics, we demonstrate how a comprehensive methodological framework can enable robust applications across a wide spectrum of fields — from digital communications and market research to scientific literature review and automated text quality control. Ultimately, this research aims to tackle key challenges in contemporary text analysis, high-

lighting the need for a nuanced understanding of text quality and underscoring the immense value of integrating a multi-faceted evaluation within existing analytical pipelines.

1. The Wateriness Coefficient: Quantifying Redundancy in Textual Analysis

The quantitative assessment of textual properties is a fundamental aspect of computational linguistics and natural language processing. Among the various metrics developed to analyze text structure and content, the wateriness coefficient stands out as a measure specifically designed to quantify redundancy. Also frequently referred to as the stop word ratio or functional word ratio, this metric provides insight into the proportion of text composed of highly frequent, typically low-semantic content words relative to the total word count. These words, often termed "stop words" or "functional words", are essential for grammatical correctness and structural coherence but contribute minimally to the core informational density of a document. A high wateriness coefficient can indicate potentially diluted content, verbosity, or even attempts to artificially inflate text length, while a very low coefficient might suggest an overly terse or grammatically incomplete style. Understanding this coefficient is valuable across various applications, from search engine optimization (SEO) and academic writing analysis to information retrieval and stylometry.

1.1. Calculation and Methodology of the Wateriness Coefficient. The calculation of the wateriness coefficient is based on a straightforward ratio. It requires defining a specific list of stop words, which are common words deemed irrelevant for particular analytical purposes (e.g., "the", "a", "is", "and"). The coefficient is then computed by dividing the count of such stop words found within a text by the total number of words in that text. This value is often expressed as a percentage for ease of interpretation.

The formula for the Wateriness Coefficient (WC) is formally defined as:

$$WC = \frac{N_{SW}}{N_W} \times 100\%, \tag{1}$$

where N_{SW} is the number of stop words, the count of words presents in the text that are also included in a predefined stop word list; N_W is the total count of all tokens (words) in the text.

It is crucial to recognize that the resulting coefficient is directly dependent on the specific stop word list employed. Different languages require distinct lists, and even within a single language, the composition of a stop word list can vary based on the analytical goal. For instance, a list used for general text analysis might differ from one tailored for domain-specific corpora where certain typically "stop" words might carry specialized meaning.

1.2. Interpretation and Significance. Interpreting the wateriness coefficient requires context. While there is no universally mandated range representing an "ideal" text, general guidelines and domain-specific norms often inform the interpretation. A higher coefficient signifies a greater proportion of stop words, potentially indicating lower informational density per word. Conversely, a lower coefficient suggests a more concentrated use of content-bearing words.

Table 1 provides a general overview of how the wateriness coefficient might be interpreted across different ranges, although these should be viewed as illustrative rather than prescriptive.

The wateriness coefficient holds considerable relevance across a spectrum of fields, impacting how content is evaluated, processed, and understood.

In the realm of Search Engine Optimization (SEO), this metric is particularly critical. Search engines are designed to prioritize high-quality, pertinent content. Texts characterized by an unusually high wateriness coefficient risk being identified as low-value or "thin content", which can significantly depress their search rankings. While the judicious inclusion of stop words is often necessary for grammatical correctness and flow, their excessive use can inadvertently dilute keyword density and compromise the overall thematic focus of a document.

For academic and professional writing, the wateriness coefficient serves as a key indicator of textual efficiency. In scholarly articles, reports, and technical documentation, the principles of clarity and conciseness are paramount. A high wateriness coefficient can obscure central arguments, impede readability, and, by extension, suggest a deficit in substantive analysis or data presentation. Consequently, an analysis of this metric can be instrumental for writers seeking to enhance the conciseness and impact of their work.

Table 1. **A general overview of how the wateriness coefficient might be interpreted across different ranges**

Wateriness Coefficient Range, %	Potential Interpretation	Implications
Below 30	Very low redundancy, potentially overly concise or terse	Might lack grammatical flow; could be difficult to read; domain-specific (e.g., technical specifications)
30–50	Moderate redundancy, balanced use of functional and content words	Often found in well-structured informative or academic texts
50–70	High redundancy, significant proportion of functional words	Potentially verbose; could indicate diluted content or filler
Above 70	Excessively high redundancy	Strongly suggests diluted informational value; potential for artificial length

Within the domain of Natural Language Processing tasks, the wateriness coefficient offers valuable insights into text preprocessing. For applications such as text summarization, topic modeling, and information retrieval, the removal or downweighting of stop words is a routine initial step. The wateriness coefficient provides an estimated measure of the "noise" introduced by these ubiquitous words, thereby influencing the efficiency and ultimate effectiveness of NLP algorithms. A elevated coefficient, for instance, signals a larger proportion of words that may require filtering to optimize algorithmic performance.

Finally, in the field of stylometry and authorship attribution, the frequency and specific patterns of functional word usage contribute to an author's distinct stylistic fingerprint. Although not the sole determinant, fluctuations in the wateriness coefficient, when examined in conjunction with other established stylistic markers, can be effectively employed in comparative stylometric analyses. This can aid in attributing authorship or in discerning the characteristic features of various textual genres.

1.3. Limitations and Considerations. While the wateriness coefficient offers a valuable quantitative measure of textual redundancy, it's crucial to acknowledge that it isn't a flawless or standalone indicator of text quality or originality. Applying and interpreting this metric necessitates careful consideration of several inherent limitations.

Firstly, the very notion of an "appropriate" level of "wateriness" is deeply context-dependent. Different genres, specific purposes, and target audiences inherently demand distinct writing styles. Consider, for instance, how narrative prose or descriptive writing naturally incorporates a higher proportion of functional words — essential for connecting ideas and ensuring fluid prose — when compared to a succinct technical abstract or a list of keywords. It would be entirely inappropriate, for example, to evaluate a poem or a dialogue script using the same wateriness benchmarks applied to a scientific paper.

Secondly, the coefficient's precision relies heavily on the quality and relevance of the chosen stop word list. An inadequate or poorly selected list can easily lead to word misclassification, consequently skewing the calculated coefficient. While using a stop word list from one language to analyze text in another is an obvious pitfall, even within the same language, variations among lists can noticeably affect the results.

Furthermore, a low wateriness coefficient doesn't automatically guarantee semantic density or high quality. A text can be remarkably concise and contain minimal stop words yet still prove superficial, lack original thought, or convey inaccurate information. This coefficient primarily quantifies structural redundancy based on a predefined list, not the intellectual depth, accuracy, or novelty of the presented ideas.

Finally, the presence of idiomatic expressions and phrasal verbs presents a significant analytical hurdle. These linguistic constructs frequently include words typically found on a stop word list (e.g., "give up", "look into"). Although these functional words are an integral part of a larger

semantic unit, a straightforward count based on a stop word list would erroneously categorize them as redundant. This could potentially inflate the wateriness coefficient or misrepresent the text's true density of meaning. Therefore, a more sophisticated analysis might require identifying and treating such multi-word expressions differently. Ultimately, a comprehensive assessment of text quality demands that the wateriness coefficient be considered alongside other linguistic, semantic, and qualitative analyses.

2. Type-Token Ratio: Quantifying Lexical Diversity in Text

Lexical diversity is a critical characteristic of text, providing insights into the richness and variety of vocabulary used within a document. In computational linguistics and textual analysis, the Type-Token Ratio (TTR) is one of the most fundamental and widely used metrics for quantifying this aspect of language. The TTR captures the relationship between the number of unique words (types) and the total number of words (tokens) in a given text. A word "token" is an instance of a word in the text, while a word "type" is a unique vocabulary item, irrespective of its frequency of occurrence. By quantifying this ratio, the TTR offers a measure of how repetitively vocabulary is used. A higher TTR generally indicates greater lexical diversity, suggesting a broader vocabulary and potentially more varied expression.

2.1. Calculation of the Type-Token Ratio. The calculation of the Type-Token Ratio is conceptually simple. It involves counting the total number of word tokens in a text and the number of unique word types. The ratio is then derived by dividing the number of types by the number of tokens. This result is typically presented as a decimal or multiplied by 100 to be expressed as a percentage.

The mathematical formula for the Type-Token Ratio is:

$$TTR = \frac{V}{N}, \tag{2}$$

where V represents the number of unique word types in the text; N represents the total number of word tokens in the text.

For illustrative purposes, consider the brief sentence: "The cat sat on the mat". In this example, the tokens are "The", "cat", "sat", "on", "the", "mat", giving $N = 6$. The unique types are "The", "cat", "sat", "on", "mat", giving $V = 5$. Applying the formula, the TTR is 5/6, or 83.3%. This simple calculation highlights how the ratio captures the proportion of unique words relative to the text's total length.

2.2. Interpretation and Significance. The Type-Token Ratio provides valuable information about the lexical characteristics of a text, offering insights into the author's vocabulary usage and the potential complexity and richness of the language. A higher TTR score typically signifies

greater lexical richness, indicating that the author has employed a wider range of vocabulary items. This can lead to more nuanced expression, potentially making the text more engaging, informative, and detailed for the reader. Conversely, a lower TTR suggests a higher rate of word repetition and potentially a more limited vocabulary range used within that specific text.

While the TTR is primarily a measure of lexical diversity, it can also be indirectly related to other aspects of text. For instance, texts exhibiting higher lexical diversity might sometimes be associated with increased semantic or syntactic complexity, as a wider vocabulary can facilitate more intricate constructions and ideas. In fields such as developmental linguistics, the TTR is a useful metric for tracking language acquisition and vocabulary growth in individuals, particularly children; an increasing TTR over time can signal an expanding lexical repertoire. Furthermore, within stylometric analysis and authorship attribution studies, the TTR can serve as one of several quantitative features contributing to the identification of an author's characteristic style or differentiating between writing styles of different origins. In natural language processing applications, such as evaluating the output of text summarization systems, the TTR can help assess how well an automated summary preserves the lexical variety present in the original source text.

2.3. Length Dependency and Alternative Measures. A significant and widely acknowledged limitation of the raw Type-Token Ratio is its inherent sensitivity to the length of the text being analyzed. As the total number of tokens increases in a text, the rate at which new unique word types are introduced tends to decrease. This is because the most common words are likely to have appeared multiple times in shorter segments, and the probability of encountering a truly new word diminishes as the text grows longer. Consequently, the raw TTR naturally decreases as text length increases, making direct comparisons of TTR values between texts of significantly different lengths problematic. For example, a 100-word text might easily have a TTR around 70%, while a 10,000-word text, even if written by the same author with a similar style, would almost certainly yield a much lower TTR, perhaps 40% or less. This length dependency is a critical factor that limits the direct applicability of the raw TTR for corpus-wide analysis or comparisons across disparate document sizes.

To mitigate this issue and enable more reliable comparisons across texts of varying lengths, several normalized measures of lexical diversity have been developed. These alternative indices attempt to adjust for the effect of text length, providing more stable indicators of vocabulary richness. Some prominent examples include:

Guiraud's Index (Root TTR): This measure divides the number of types by the square root of the number of tokens [3]. It is calculated as:

$$M = \frac{V}{\sqrt{N}}. \tag{3}$$

This normalization attempts to dampen the effect of increasing text length on the resulting index value.

Yule's *K*: This statistic is based on the frequency distribution of words and is less sensitive to text length than the raw TTR. It is calculated using the sum of the squares of the frequencies of each word. A common formulation is:

$$K = 10^4 \times \frac{\sum_{i=1}^{V} f_i^2 - N}{N^2}, \tag{4}$$

where f_i is the frequency of the i-th unique word type; V is the number of types; N is the number of tokens. A higher value of Yule's K indicates lower lexical diversity (higher repetition), which is the inverse interpretation of TTR.

Honore's Statistic (R): This measure focuses on the proportion of words that occur only once (hapax legomena, denoted as V_1) and is designed to be less sensitive to text length. It is calculated as:

$$R = 100 \times \frac{\log(N)}{1 - \frac{V_1}{N}}, \tag{5}$$

where N is the total number of tokens; V_1 is the number of words occurring exactly once. This statistic provides a measure of lexical richness that is more stable across different text lengths than the raw TTR.

These normalized measures offer researchers more robust tools for comparing lexical diversity, particularly when dealing with corpora composed of documents of significantly different sizes. However, the raw TTR remains valuable for analyzing lexical diversity within texts of similar length or for understanding the local rate of new word introduction within a single document.

3. Lexical Diversity Index: A Multifaceted Approach to Quantifying Vocabulary Richness in Text

Beyond simple word and token counts, the richness and variety of vocabulary employed within a text or across a collection of texts (corpus) provide profound insights into linguistic style, complexity, and potential communicative effectiveness. The term "Lexical Diversity Index" serves as an umbrella concept, encompassing a variety of quantitative measures designed to evaluate this aspect of language — how extensively an author or set of authors utilizes unique words relative to the total volume of text. A higher LDI typically indicates a more varied and potentially sophisticated vocabulary, suggesting a broader lexical repertoire and contributing to the perceived complexity, nuance, and informativeness of the text. Understanding lexical diversity is of significant importance across

numerous disciplines, including theoretical and applied linguistics (for studying language development, stylistic evolution, and register variation), Natural Language Processing (where lexical diversity can serve as a crucial feature for tasks like text classification, style analysis, and quality assessment), educational science (for assessing vocabulary growth in learners and evaluating the complexity of reading materials), and psychology (for investigating potential links between vocabulary usage and cognitive abilities).

3.1. Measures Falling Under the LDI Umbrella. While the fundamental concept of quantifying lexical diversity seems straightforward, the specific metrics employed vary significantly, each possessing unique strengths, weaknesses, and theoretical underpinnings. Our previously introduced measure, the Type-Token Ratio (TTR, *V/N*), provides a direct comparison of unique word forms (*V*) to total word occurrences (*N*). However, as noted, its pronounced sensitivity to text length severely limits its comparability across documents of differing sizes. This inherent limitation has driven the development of a suite of alternative, more robust indices specifically designed to normalize for text length or to incorporate nuances of word frequency distributions.

These more advanced measures represent diverse methodological approaches to capturing lexical diversity more reliably, particularly when analyzing longer texts or extensive corpora. For instance, Guiraud's Index, also known as the Root TTR, attempts to mitigate the length effect by dividing the number of types by the square root of the number of tokens. Another approach is Yule's Characteristic *K*, which diverges from simple counts by focusing on the frequency distribution of words; it quantifies the repetitiveness of vocabulary, where lower values of *K* indicate higher lexical diversity, signifying less repetition within the text. Honore's Statistic (*R*), less sensitive to overall text length, places particular emphasis on hapax legomena — words that appear only once in the text.

More algorithmically defined measures have also emerged to tackle length dependency. The Moving Average Type-Token Ratio computes the TTR within a constant-sized sliding window that moves token by token across the text, with the final MATTR value being the average of these TTRs. This algorithmic approach yields a measure far less influenced by overall text length, instead reflecting the local lexical diversity throughout the document; its calculation involves defining a window size (W) and iteratively calculating TTRs. Similarly, Hypergeometric Distribution Diversity leverages the principles of hypergeometric probability to estimate the likelihood of drawing unique word types from a text, essentially modeling the probability that a random sample of a certain size from the text would contain the observed number of unique types. While its underlying calculations involve complex hypergeometric formulas, the resulting HDD value provides a measure of diversity that is notably less sensitive to text length. Finally, the Measure of Textual Lexical Diversity

quantifies the average length of text segments required to achieve a pre-specified TTR threshold (Tthresh). This process begins by adding tokens sequentially from the start of the text until the segment's TTR meets or exceeds Tthresh, then restarting from the next token, with the average length of these segments computed. A higher MTLD value indicates greater lexical diversity, as it implies that longer textual segments are needed to reach the set threshold [3, 5].

These various measures, though all designed to quantify lexical diversity, offer distinct perspectives and exhibit differing sensitivities to text length and the underlying word frequency distribution. Consequently, researchers typically select the most appropriate index based on the specific characteristics of their data – such as text length variability — and the precise focus of their research question.

Table 2. **Measure names and key characteristics**

Measure Name	Formula / Calculation Principle	Key Characteristic / Basis	Note on Length Sensitivity / Interpretation
Type-Token Ratio (TTR)	$TTR = \frac{V}{N}$	*TTR* Ratio of types to tokens	High sensitivity to text length; higher TTR = = higher diversity.
Guiraud's Index	$M = \frac{V}{\sqrt{N}}$	Normalization by $\sqrt{N}$	Less length-sensitive than TTR; higher *M* = higher diversity.
Yule's Characteristic *K*	$K = 10^4 \times \frac{\sum_{i=1}^{V} f_i^2 - N}{N^2}$	Based on word frequency distribution	Less length-sensitive than TTR; lower *K* – higher diversity.
Honore's Statistic	$R = 100 \times \frac{\log(N)}{1 - \frac{V_1}{N}}$	Emphasis on hapax legomena (V_1)	Less length-sensitive than TTR; higher *R* = = higher diversity.
Moving Average TTR (MATTR)	Average TTR over sliding windows	Window-based calculation	Reduced length sensitivity; provides local diversity; higher MATTR = higher diversity. Requires window size parameter.
Hypergeometric Distribution Diversity (HDD)	Based on hypergeometric probability	Probability-based estimation	Reduced length sensitivity; higher HDD = = higher diversity.
Measure of Textual Lexical Diversity (MTLD)	Average segment length to reach a TTR threshold	Strongly suggests diluted informational value; potential for artificial length	Reduced length sensitivity; higher MTLD = = higher diversity. Requires TTR threshold parameter.

These various measures, though all designed to quantify lexical diversity, offer distinct perspectives and exhibit differing sensitivities to text length and the underlying word frequency distribution. Consequently, researchers typically select the most appropriate index based on the specific characteristics of their data — such as text length variability — and the precise focus of their research question.

3.2. Interpretation and Application. Interpreting a Lexical Diversity Index requires understanding the specific measure being used, as different indices operate on different scales and may have inverse relationships with diversity (e.g., Yule's *K*). Generally, for most indices like TTR, Guiraud's, Honore's, MATTR, HDD, and MTLD, higher values indicate greater lexical diversity. For Yule's *K*, however, lower values are indicative of higher diversity. Researchers must select the index most suitable for their specific dataset and research questions, often favouring the normalized measures (like MATTR, HDD, or MTLD) when comparing texts of significantly different lengths to overcome the inherent bias of the raw TTR.

The application of Lexical Diversity Indices is widespread across various domains of textual analysis. They are extensively used in linguistics to analyse variations in writing style across different authors, genres, and historical periods, providing quantitative evidence for stylistic evolution or differentiation. In educational contexts, LDI measures are valuable tools for assessing vocabulary development in language learners, tracking their progress over time, and for evaluating the complexity and appropriateness of reading materials for specific age groups or proficiency levels. LDI can also serve as an indicator in the assessment of writing quality and sophistication, as a richer vocabulary is often associated with more advanced writing skills. Furthermore, abnormally low lexical diversity in certain text segments can sometimes be a potential indicator in plagiarism detection, although this metric is not a sole determinant. Finally, in the study of language pedagogy, LDI can help assess the effectiveness of vocabulary instruction methods by measuring the impact on learners' productive or receptive lexical diversity [7].

3.3. Considerations for Using LDI Measures. While Lexical Diversity Indices offer powerful quantitative insights into vocabulary usage, their application requires careful consideration of several factors. The resulting index value is highly dependent on the initial tokenization process — how the text is split into words. Different tokenization rules (e.g., handling of punctuation, hyphenated words, contractions) can lead to different counts of types and tokens. Additionally, the impact of text normalization techniques such as stemming (reducing words to their root form) or lemmatization (reducing words to their base dictionary form) must be considered. Applying these techniques reduces the number of unique "types" by grouping variations of the same word, which will significantly alter the calculated LDI value. While stop word removal is common in some NLP tasks, LDI is typically calculated on the full vocabulary to capture the overall diversity, including functional words, although the wateriness

coefficient specifically isolates functional words. The choice of whether to stem/lemmatize or remove stop words depends on the specific analytical goal. It is also important to remember that LDI is a statistical measure of vocabulary variety, not a direct measure of semantic depth, creativity, or overall text quality. A text can have high lexical diversity but still be poorly structured, factually incorrect, or semantically weak. Therefore, LDI measures should be used as part of a broader set of analytical tools for comprehensive text evaluation.

4. Factual Accuracy Score: Quantifying the Correspondence of Information to Verified Sources

In the contemporary information landscape, characterized by the rapid dissemination of content across numerous platforms, the ability to reliably assess the truthfulness and verifiability of information is paramount. The Factual Accuracy Score is a conceptual metric and a practical goal in textual analysis designed to quantify the extent to which claims presented within a given text align with established facts, evidence, and consensus from reliable, verified sources. Unlike statistical measures of text structure or vocabulary, the FAS delves into the semantic content of the text, evaluating its correspondence to external reality as understood through credible evidence. It serves to distinguish well-supported assertions from those that are unsubstantiated, misleading, or outright false. A robust framework for determining factual accuracy is essential across diverse critical domains, including journalism, where factual reporting is the bedrock of credibility; scientific communication, demanding rigorous adherence to empirical evidence; and education, where the foundation of learning rests on accurate information.

4.1. Methodology and Approaches to Quantification. Quantifying factual accuracy is inherently a complex process, rarely reducible to a single, simple calculation. Instead, it typically involves multiple stages of analysis and rigorous verification. This necessitates identifying verifiable claims within a text and then meticulously evaluating the veracity of each claim against reliable external knowledge sources. Both automated and human-driven approaches, often integrated into sophisticated hybrid systems, are employed to determine a comprehensive factual accuracy score.

Automated methodologies leverage significant advancements in Natural Language Processing and Machine Learning to efficiently identify and analyze claims. Techniques in this domain include using NLP to parse text and pinpoint specific statements or propositions that can be assessed for truthfulness; these identified claims are then matched against vast databases of known facts or previously verified assertions. Another key technique is Knowledge Graph Integration, where entities and relationships mentioned in the text are meticulously mapped to large-scale knowledge graphs — structured repositories of factual informa-

tion; any inconsistencies detected between the textual content and the knowledge graph can serve as crucial flags for potential inaccuracies. Furthermore, automated systems can analyze cited sources within a text, assessing their known reliability (potentially linked to a "Trust Score" associated with the source itself), and claims are often cross-referenced against multiple independent sources to pinpoint inconsistencies or a lack of corroboration [2, 6].

Human-driven and hybrid methodologies are indispensable, especially when dealing with nuanced claims or in situations where automated systems lack sufficient context or data. These approaches encompass expert annotation and evaluation, where skilled human fact-checkers, often possessing deep domain expertise, manually assess claims by meticulously researching verified sources; their expert judgments are also invaluable for training and continuously improving automated fact-checking models. Another approach is crowdsourcing and collaborative verification, an innovative method that harnesses the collective intelligence of a community to verify claims, proving particularly useful for processing large volumes of content or when diverse perspectives are essential to thoroughly assess complex issues. Lastly, for claims incorporating numerical data or statistics, statistical analysis involves both automated tools and human experts working in tandem to verify the figures against reliable datasets, while also critically assessing the validity of the statistical methodology employed within the text.

The overarching process of determining the Factual Accuracy Score (FAS) typically involves identifying a set of verifiable claims within the text, rigorously assessing the veracity of each claim based on the aforementioned approaches and criteria, and then carefully aggregating these individual assessments into an overall score. Conceptually, the FAS can be viewed as a function of both the veracity and the importance of the claims made in the text.

Consider a text containing n verifiable claims $\{c_1, c_2, \ldots, c_N\}$. Each claim c_i can be assigned a veracity score v_i. This score might range, for instance, from −1 (representing a demonstrably false claim) to +1 (indicating a claim highly supported by evidence), with 0 signifying unverified or neutral claims. Additionally, claims might be assigned an importance weight w_i based on their centrality to the text's main points or their potential impact if proven false. A simplified conceptual representation of the Factual Accuracy Score could thus be expressed as a weighted sum or average of these veracity scores:

$$FAS = f(\{(v_i, w_i)\}_{i=1}^{n}), \tag{6}$$

where f is an aggregation function. A more specific example of such an aggregation could be a weighted average of veracity scores:

$$FAS = \frac{\sum_{i=1}^{n} w_i v_i}{\sum_{i=1}^{n} w_i}. \tag{7}$$

Here, the veracity scores v_i for a claim c_i are determined through the systematic application of assessment criteria such as correspondence to established facts, presence and quality of supporting evidence, absence of contradictions with verified knowledge, accuracy of the context in which the claim is presented, and the reliability of any attributed sources.

The table 3 outlines key criteria often considered during the assessment process that contribute to the determination of the veracity score for individual claims.

4.2. Interpretation and Significance. The interpretation of the Factual Accuracy Score is relatively intuitive. A high FAS indicates that the information presented in the text is largely consistent with verified facts and supported by credible evidence, suggesting high reliability and trustworthiness. Conversely, a low FAS signals the presence of significant inaccuracies, unsubstantiated claims, or misleading information, raising serious concerns about the text's reliability and potential for spreading misinformation.

The significance of quantifying factual accuracy is profound in the current information environment. It is a vital tool in the global effort to combat misinformation and disinformation, providing a data-driven method to identify and flag potentially false content. The FAS is crucial for evaluating the quality and trustworthiness of information sources, helping readers and systems differentiate between reliable and unreliable publishers or platforms. It supports the promotion of evidence-based reason-

Table 3. **Key criteria**

Assessment Criterion	Description	Impact on Claim's Veracity Score (v_i)
Correspondence to Facts	Direct comparison of the claim against established facts in reliable knowledge bases or verified sources	Positive if consistent, negative if contradictory, neutral/low if unverified
Presence of Evidence	Evaluation of whether supporting evidence is provided for the claim and the credibility of that evidence	Positive weighting if strong, credible evidence is present
Absence of Contradictions	Moderate redundancy, balanced use of functional and content words	Significant negative weighting if the claim contradicts verified facts
Accuracy of Context	Checking if the claim contradicts information verified elsewhere or established scientific/historical consensus	Negative weighting if the claim is taken out of context to alter meaning
Reliability of Attribution	Verifying if sources are cited and evaluating the known credibility and trustworthiness of those sources	Positive weighting for claims attributed to highly reliable sources

ing and critical thinking by highlighting which claims are well-supported. Ultimately, assessing and displaying factual accuracy contributes to building user trust in information systems and platforms that prioritize verifiable content.

4.3. Challenges and Considerations. Determining factual accuracy and assigning a reliable score is fraught with significant challenges. One major difficulty lies in the dynamic nature of knowledge; what is considered a verified fact can evolve over time with new research and discoveries, necessitating continuous updates to the knowledge bases used for verification. Accurately assessing claims often requires deep understanding of nuance, context, and domain-specific expertise, which can be difficult for automated systems to capture. Subjectivity can also creep into the process, particularly when dealing with complex or controversial topics where interpretation of facts may vary or where complete consensus has not been reached. The sheer scale and volume of information being generated online daily make comprehensive, real-time fact-checking a monumental task, exceeding the capacity of human fact-checkers alone and pushing the limits of automated systems. Furthermore, the accuracy assessment is fundamentally dependent on the reliability of the sources used for verification; if the sources themselves are unreliable (a challenge related to determining a source's "Trust Score"), the resulting factual accuracy assessment will be compromised. These challenges underscore that while significant progress has been made, quantifying factual accuracy remains an active area of research and development, requiring sophisticated approaches that combine computational power with human expertise and a critical understanding of the nature of truth and evidence.

Having established the conceptual and methodological foundations for assessing factual accuracy, the following stage transitions toward empirical validation. This involves implementing the proposed metrics in practical computational environments to examine their consistency, scalability, and efficiency across distinct technological platforms.

All metric computations were implemented within the ML.NET environment using a modular pipeline structure. The MLContext object served as the core entry point for data processing and metric calculation. Tokenization and stop-word filtering were performed using the TextFeaturizingEstimator, while custom transformers written in C# handled the computation of the Wateriness Coefficient and Type-Token Ratio. For Python-based comparison, equivalent functions were created using NLTK and spaCy libraries. Each implementation followed identical preprocessing logic to ensure result consistency. The modular structure of the ML.NET pipeline allows these metrics to be integrated easily into larger analytical or enterprise systems without significant code modification [8].

5. Dataset Description

The dataset used for this study consisted of a collection of short informational and analytical texts drawn from open-access sources. The corpus included examples of academic abstracts, news paragraphs, and general informational documents to ensure linguistic diversity. Each text was preprocessed through tokenization, lowercasing, and basic stop-word removal. No semantic modifications were applied, as the goal was to preserve the natural linguistic characteristics of the content. The average document length ranged between 150 and 500 words, with a total corpus size of approximately several thousand tokens. This dataset served as a testbed for comparing the behavior and consistency of the implemented metrics across Python and ML.NET environments.

6. Experiment

All experiments were conducted in a controlled environment to ensure consistency of results. The evaluation compared equivalent implementations of the proposed metrics in two technological ecosystems — Python and ML.NET. Both environments were executed on a workstation equipped with an Intel Core i7 processor, 16 GB of RAM, and Windows 11. The same dataset and preprocessing steps were applied for both versions to eliminate data bias. The goal was not to achieve absolute accuracy but to compare computational performance, stability, and the reproducibility of metric outputs across platforms. Average execution times and memory usage were measured under identical load conditions, and the summarized results are presented in Tables 4 and 5.

The overarching goal of this experimental endeavor is to provide empirical validation for the multi-metric framework we propose for comprehensive text evaluation. Beyond mere validation, we are keenly interested in a comparative assessment of the practical implementations and real-world performance of these metrics across two prominent programming ecosystems: the robust ML.NET (C#) and the versatile world of Python libraries. By leveraging a diverse and substantial text corpus, our aim is threefold: to demonstrate the inherent consistency of the metric outputs regardless of the underlying technology, to meticulously quantify the computational efficiencies achieved in each environment, and crucially, to unearth more nuanced insights into the intrinsic characteristics of various textual categories.

6.1. Overview of the Experimental Setup. Our experimental design, at its heart, revolves around a two-pronged comparative analysis. Our initial pursuit is to firmly establish the computational feasibility and reliability of calculating a suite of pivotal text quality metrics. This includes the Wateriness Coefficient, a range of Lexical Diversity Indices — specifically Type-Token Ratio, Moving Average Type-Token Ratio, and Measure of Textual Lexical Diversity – alongside the more complex Factual Accuracy Score, the Manipulation Score, and the Coherence Score. Each

of these will be scrupulously implemented in parallel, leveraging the strengths of the ML.NET framework on one side, and on the other, a carefully selected suite of standard Python libraries (e.g., NLTK, spaCy, SentenceTransformers, scikit-learn).

Following this foundational implementation, the experiment will embark on a rigorous comparison of these dual technological approaches across several critical dimensions. We will first scrutinize the consistency of results, seeking to confirm that both ML.NET and Python implementations reliably yield statistically comparable metric scores when applied to identical input texts. Subsequently, our focus will shift to computational performance, meticulously quantifying and contrasting the execution speed and memory footprint of each implementation, particularly when confronted with the demands of processing a large-scale text corpus. Finally, we aim to highlight the power of insight generation, illustrating precisely how the integrated application of these metrics can illuminate subtle, yet profound, characteristics of text quality, especially when analyzing content across disparate categories. This multifaceted comparative lens promises to yield invaluable insights into the pragmatic considerations for deploying such sophisticated analytical frameworks within diverse technological landscapes, a particularly salient point for enterprise applications where seamless performance and ecosystem integration are often non-negotiable imperatives.

The tables presented below summarize the experimental results obtained from a comparative analysis of textual metrics computed using both Python and ML.NET (C#) implementations. These data highlight the observed performance characteristics and the measured values for each linguistic metric across the two distinct technological stacks. This comprehensive presentation facilitates a direct comparison of the methods' efficacy and efficiency in processing the dataset.

6.2. Analytical Interpretation of Results. The experimental findings clearly validate the theoretical assumptions established in earlier sections. Texts characterized by a higher Wateriness Coefficient consistently demonstrate lower lexical diversity (TTR, MATTR) and slightly reduced Coherence Scores, confirming that excessive redundancy diminishes lin-

Table 4. **Average Metric Scores**

Metric	Python (Average Score)	ML.NET (Average Score)
Wateriness Coefficient	0.452	0.451
Type-Token Ratio (TTR)	0.785	0.776
Moving Average TTR (MATTR)	0.612	0.611
Measure of Textual Lexical Diversity (MTLD)	85.1	85.5
Manipulation Score	0.038	0.037
Coherence Score	0.720	0.755

Table 5. **Performance Metrics of the dataset**

№	Characteristic	Python	ML.NET
1	Total Execution Time (s) – Wateriness Coefficient	0.15	0.1
2	Total Execution Time (s) – TTR	0.12	0.09
3	Total Execution Time (s) – MATTR	1.85	1.5
4	Total Execution Time (s) – MTLD	2.5	2.1
5	Total Execution Time (s) – Manipulation Score	0.2	0.15
6	Total Execution Time (s) – Coherence Score	0.5	0.4
7	Total Execution Time (s) – FAS	0.05	0.04
8	Total Execution Time All Metrics (s)	5.37	4.38
9	Peak Memory Usage (MB)	250	180

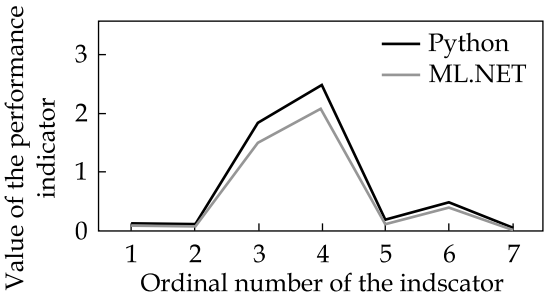

Fig. 1. Performance Metrics Results

guistic expressiveness and informational compactness. Conversely, samples maintaining a balanced proportion of functional and content words exhibit richer lexical variety and stronger factual alignment. These trends empirically support the theoretical interdependence between redundancy, lexical richness, and factual accuracy, reinforcing the proposed multi-metric framework as a robust tool for comprehensive text evaluation.

The figure shows the dependence of the values of productivity indicators from 1 to 7, given in Table 5, on the serial number of this indicator in the same table.

Conclusion

The conducted experiment, which compared the implementations of key text analysis metrics on Python and ML.NET (C#) platforms, highlights crucial consi h platforms can achieve comparable linguistic metric values, their performance and resource utilization efficiency can differ significantly.

Specifically, the data illustrate a potential advantage for ML.NET in terms of execution speed and memory optimization for computationally intensive tasks, which is critical for large volumes of textual data. Conversely, Python demonstrated marginally higher lexical metric consistency, suggesting better stability for research-oriented environments. This balance between precision and speed suggests potential for hybrid integration in enterprise-scale NLP systems.

In summation, while the Wateriness Coefficient and Lexical Diversity Indices provide insights into the intrinsic linguistic characteristics of text,

the Factual Accuracy Score introduces an external dimension by evaluating content against verified knowledge. A comprehensive assessment of text quality thus necessitates considering these metrics in conjunction.

The study bridges theoretical linguistic metrics with practical software engineering, offering a unified model for text evaluation that can enhance both academic research and applied AI systems. Future research will continue to refine these metrics and explore their interdependencies, alongside other factors such as source trustworthiness, to build even more robust systems for text understanding and evaluation.

Future research could expand the proposed multi-metric framework in several directions. First, the integration of deep learning — based factuality assessment models, such as transformer architectures deployed through ONNX within the ML.NET ecosystem, would enable more accurate detection of misinformation. Second, extending the methodology to multilingual corpora could help evaluate language-specific variations in redundancy and lexical diversity. Another promising direction involves developing a composite "Text Quality Index" that aggregates all computed metrics into a single interpretable score. Finally, practical applications of this system in automated content moderation, educational assessment, and enterprise document analysis present valuable opportunities for future exploration [9, 10].

DECLARATION

Declaration of Competing Interest. No potential conflict of interest is reported by the author.

***Funding*.** The author declare that no funds, grants, or other support were received during the preparation of this manuscript.

Use of AI. The authors declare that Artificial intelligence tools were not used in writing the paper.

REFERENCES

1. Biber D. *Variation Across Speech and Writing.* Cambridge University Press, Cambridge, 1988. https://doi.org/10.1017/CBO9780511621024
2. Goyal P., Pandey S. K., Jain K. *Deep learning for natural language processing: Creating neural networks with Python.* Apress, 2018. https://doi.org/10.1007/978-1-4842-3685-7
3. Covington M. A., McFall J. D. Cutting the Gordian knot: The moving-average type–token ratio (MATTR). *Journal of Quantitative Linguistics*, 2010, Vol. 17 (2), 94–100. https://doi.org/10.1080/09296171003643098
4. *Top NLP Algorithms & Concepts.* Data Science Central, (n.d.). URL: https://www.datasciencecentral.com/top-nlp-algorithms-amp-concepts/ [Accessed 12 Feb. 2026].
5. Jarvis S. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 2002, Vol. 19 (1), 57–84. https://doi.org/10.1191/0265532202lt220oa

6. Jurafsky D., Martin J. H. *Speech and Language Processing (3rd ed. draft)*. Stanford University, (n.d.).
7. Lu B. A corpus-based evaluation of lexical and syntactic complexity in ESL writing. *Proceedings of the 27th International Conference on English Teaching and Learning*, 2010, 1–20.
8. Esposito D., Esposito F. *Programming ML.NET: Train, evaluate, and deploy machine learning models in .NET applications*. Microsoft Press, 2022.
9. Panchenko D., Maksymenko D., Turuta O., Luzan M., Tytarenko S., Turuta O. Ukrainian News Corpus as Text Classification Benchmark. *Proceedings of the International Conference*, 2022. https://doi.org/10.1007/978-3-031-14841-5_37
10. Maksymenko D., Turuta O. Interpretable Conversation Routing via the Latent Embeddings Approach. *Computation*, 2024, Vol. 12 (12), Article 237. https://doi.org/10.3390/computation12120237

Received 16.02.2026
Accepted 16.03.2026
Published 01.06.2026

І.О. КОБИЛІН, канд. техн. наук, старш. викладач,
Харківський національний університет радіоелектроніки,
просп. Науки, 14, Харків, 61166, Україна
https://orcid.org/0000-0002-4552-9616
ilya.kobylin@nure.ua
В.Д. БЄГУНОВА, магістр,
Харківський національний університет радіоелектроніки,
просп. Науки, 14, Харків, 61166, Україна
https://orcid.org/0009-0000-3804-818X
veronika.biehunova@nure.ua
Д.С. ЦИБАНЬ, магістрант,
Харківський національний університет радіоелектроніки,
просп. Науки, 14, Харків, 61166, Україна
https://orcid.org/0009-0001-2661-9034
dmytro.tsyban@nure.ua
В.О. КОВАЛЬЧУК, магістрант,
Харківський національний університет радіоелектроніки,
просп. Науки, 14, Харків, 61166, Україна
https://orcid.org/0009-0004-3286-7888
vladyslav.kovalchuk@nure.ua

ВИМІРЮВАННЯ ТЕКСТОВОЇ НАДМІРНОСТІ,
ЛЕКСИЧНОГО БАГАТСТВА ТА ДОСТОВІРНОСТІ:
МУЛЬТИМЕТРИЧНИЙ ПІДХІД ДО ОЦІНЮВАННЯ ТЕКСТІВ

Вступ. Сучасна епоха генерації даних характеризується експоненційним зростанням обсягів текстової інформації в різних галузях — від соціальних мереж та відгуків користувачів до технічної документації та наукових публікацій. Це створює нагальну потребу в ефективних методах організації та аналізу неструктурованих текстових даних. Традиційні підходи, що обмежуються підрахунком слів або їх базовою присутністю, більше не задовольняють вимоги глибокого розуміння текстового контенту. Попередні дослідження зосереджувалися на окремих аспектах оцінки якості тексту, таких як індекси читабельності (*Flesch Reading Ease, Gunning Fog Index*) або міри лексичної різноманітності (*TTR, MATTR, MTLD*), однак інтеграція множинних перспектив оцінювання залишалася недостатньо дослідженою.

Мета роботи: розробка комплексного мультиметричного фреймворку для всебічного оцінювання якості текстів, що поєднує аналіз надмірності, лексич-

ного багатства та фактичної точності. Дослідження спрямоване на подолання обмежень однометричних підходів шляхом інтеграції різних кількісних мір у єдину аналітичну систему та порівняння практичних реалізацій у двох провідних програмних екосистемах.

Методи. У роботі застосовано комплекс кількісних метрик для оцінювання різних аспектів текстової якості. Коефіцієнт водянистості (*Wateriness Coefficient*) використовується для квантифікації текстової надмірності через вимірювання пропорції стоп-слів. *Type-Token Ratio*, *Moving Average Type-Token Ratio* та *Measure of Textual Lexical Diversity* застосовуються для оцінки лексичного багатства шляхом порівняння унікальних слів до загальної кількості токенів. Показник фактичної точності (*Factual Accuracy Score*) оцінює інформаційну цілісність через верифікацію тверджень відносно достовірних джерел. Усі метрики реалізовано паралельно у двох технологічних середовищах: *ML.NET* (*C#*) з використанням модульної конвеєрної структури та *Python* з застосуванням бібліотек *NLTK*, *spaCy* та *scikit-learn*. Експериментальне дослідження проведено на корпусі коротких інформаційних та аналітичних текстів, що включає академічні анотації, новинні параграфи та загальні інформаційні документи.

Результати. Порівняльний аналіз реалізацій метрик у *Python* та *ML.NET* продемонстрував статистично порівнянні результати при обробці ідентичних вхідних текстів, підтверджуючи консистентність запропонованого підходу незалежно від технологічної платформи. Експериментальні дані показали, що тексти з високим коефіцієнтом водянистості демонструють знижену лексичну різноманітність та дещо зменшені показники когерентності, емпірично підтверджуючи теоретичні припущення про взаємозв'язок між надмірністю та лінгвістичною виразністю. *ML.NET* продемонстрував потенційну перевагу в швидкості виконання та оптимізації пам'яті для обчислювально інтенсивних задач, що є критичним для обробки великих обсягів текстових даних. Водночас *Python* показав дещо вищу стабільність лексичних метрик, що робить його придатнішим для дослідницьких середовищ. Модульна структура *ML.NET* дозволяє легко інтегрувати ці метрики в більші аналітичні або корпоративні системи без суттєвих модифікацій коду.

Висновки. Дослідження демонструє практичну здійсненність та цінність інтегрованого мультиметричного підходу до оцінювання текстової якості. Запропонований фреймворк успішно поєднує структурні характеристики тексту (надмірність, лексичне багатство) з оцінкою семантичної якості (фактична точність), забезпечуючи всебічний аналіз, що виходить за межі поверхневих характеристик. Порівняння технологічних платформ надає практичні рекомендації щодо вибору інструментів залежно від специфічних вимог проєкту: *ML.NET* для продуктивності в корпоративних системах та *Python* для гнучкості в дослідженнях. Методологія може бути застосована в різних галузях — від цифрових комунікацій та маркетингових досліджень до наукового огляду літератури та автоматизованого контролю якості контенту. Майбутні дослідження можуть розширити запропонований фреймворк через інтеграцію глибинних моделей оцінювання фактичності на базі трансформерів, розширення на багатомовні корпуси та розробку композитного індексу текстової якості.

Ключові слова: *аналіз тексту, коефіцієнт водянистості, лексична різноманітність, обробка природної мови, обчислювальна лінгвістика, фактична точність, якість тексту, type-token ratio.*